Tackling the Problem of Construct Proliferation

Shaffer, Jonathan A.; DeGeest, David; Li, Andrew

Published in:
Organizational Research Methods

DOI:
10.1177/1094428115598239

Article

Organizational Research Methods
2016, Vol. 19(1) 80-110

DOI: 10.1177/1094428115598239
orm.sagepub.com

Tackling the Problem of Construct Proliferation: A Guide to Assessing the Discriminant Validity of Conceptually Related Constructs

Jonathan A. Shaffer[1], David DeGeest[2], and Andrew Li[1]

Abstract

Construct proliferation—the accumulation of ostensibly different but potentially identical constructs representing organizational phenomena—is a salient problem in contemporary research. While a number of construct validation procedures exist, relatively few validation studies conduct comprehensive assessments of the discriminant validity of theoretically distinct constructs. In this article, we outline the key considerations a researcher must take into account when attempting to establish the empirical distinctness of new or existing constructs and provide a step-by-step guide on how to assess the discriminant validity of constructs while accounting for three major sources of measurement error: random error, specific factor error, and transient error. Using a number of popular measures from the leadership literature, we provide an illustrative example of how to conduct a study of discriminant validity. We include several analytic strategies in our study and discuss the similarities and differences between the results they yield. We also discuss several additional issues related to this type of research and make recommendations for conducting discriminant validity analyses.

Keywords

construct validity, discriminant validity, measurement error, leadership

[1]Department of Management, Marketing, and General Business, West Texas A&M University, Canyon, TX, USA
[2]Department of HRM & OB, University of Groningen, Groningen, The Netherlands

Corresponding Author:
Jonathan A. Shaffer, Department of Management, Marketing, and General Business, West Texas A&M University, WTAMU Box 60809, Canyon, TX 79016, USA.
Email: jshaffer@wtamu.edu

> Numquam ponenda est pluralitas sine necessitate [Plurality must never be posited without necessity].
>
> William of Ockham (as cited in Kneale & Kneale, 1962)

The development of theoretical propositions that both predict and explain phenomena is one of the most basic goals of science (Campbell, 1990). As such, the usefulness of a given theory can be judged by the extent to which it identifies interrelationships between constructs of interest and provides rationale for why such relationships exist (Greenberg, Solomon, Pyszczynski, & Steinberg, 1988). The development of parsimonious theories is also a goal of science. The principle of Occam's razor states that prediction and explanation being equal, simple theories are superior to complex ones (Cortina & DeShon, 1998). One of the major threats to parsimony and therefore to the development of useful theories is construct proliferation (Le, Schmidt, Harter, & Lauver, 2010). Construct proliferation occurs when research streams are built around ostensibly new constructs that are theoretically or empirically indistinguishable from existing constructs (Harter & Schmidt, 2008). The most direct consequence of construct proliferation may be the obstruction of the development of parsimonious organizational theories (Schwab, 1980). Scholars have also argued that construct proliferation impedes the creation of cumulative knowledge (Le et al., 2010), prevents collaboration between researchers and practitioners (Rousseau, 2007), and diminishes the influence that a scientific discipline has on other disciplines (Pfeffer, 1993).

The problem of construct proliferation may be especially salient to organizational research. As interest in the study of organizations and their constituents has increased, so too has the number of proposed constructs extant in the literature. On its surface, the abundance of such constructs is not necessarily cause for alarm. In fact, Whetten (1989) suggested that "when authors begin to map out the conceptual landscape of a topic they should err in favor of including too many factors, recognizing that over time their ideas will be refined" (p. 490). Over the past century, the conceptual landscape in the field of organizational research has become vast, and the expansion of this landscape has allowed for considerable advances in our understanding of what makes organizations and the people within them effective. However, as the number of extant constructs in the literature has grown, so have concerns over the extent to which many of these constructs are redundant (e.g., Cole, Walter, Bedeian, & O'Boyle, 2012; DeRue, Nahrgang, Wellman, & Humphrey, 2011; Hershcovis, 2011; Le et al. 2010; Tepper & Henle, 2011).

To establish a new construct or to validate an existing construct, researchers must demonstrate two things. First, they must show that the construct is conceptually distinct from related constructs. Researchers have been careful to make conceptual distinctions between new constructs and existing constructs, but, as argued by Le et al. (2010), "because of the conceptual/theoretical fluency of researchers, this requirement is a weak one and is usually easily met" (p. 113). Second, researchers must demonstrate that the construct is empirically distinct from related constructs. To be considered empirically distinct from one another, constructs should not be perfectly (or near perfectly) correlated. To a surprising extent, the need to make empirical distinctions between related organizational constructs remains unmet (Harter & Schmidt, 2008).

Given existing concerns about construct proliferation in organizational research, the purpose of this article is to provide a general guide to assessing the empirical distinctness of new or existing constructs. We draw attention particularly to the effects that three sources of measurement error—specific factor error, random error, and transient error—can have on the conclusions drawn from such assessments. In the remainder of this article, we first outline what we see as the main contributors to the problem of construct proliferation. We then offer a step-by-step guide for evaluating the empirical uniqueness of new or existing constructs that (a) gives recommendations for conducting a comprehensive and efficient literature review that focuses on identifying constructs to be included in a discriminant validity analysis, (b) provides an overview of data analysis strategies for

assessing the empirical uniqueness of a given construct, (c) describes study design features that are most relevant to such assessments, (d) suggests several possible interpretations of why constructs may appear to be empirically indistinguishable, and (e) offers an illustrative example of a discriminant validity analysis that demonstrates how empirical results and research conclusions can differ depending on the data analysis strategy used and the extent to which measurement error is controlled. Lastly, we make recommendations for conducting studies of discriminant validity analyses.

The Problem of Construct Proliferation

The most straightforward way to demonstrate the empirical distinctness of a construct is to evaluate its discriminant validity. Discriminant validity refers to the extent to which measures of theoretically distinct constructs are unrelated empirically to one another (Campbell & Fiske, 1959). Establishing the discriminant validity of measures of newly developed constructs is a particularly critical step in the process of construct validation (Harter & Schmidt, 2008) because this step can reveal whether a given construct is empirically redundant with existing constructs. Perhaps the most well-known methods for establishing discriminant validity are the multitrait-multimethod approach introduced by Campbell and Fiske (1959) and confirmatory factor analysis (CFA; Bagozzi, Yi, & Phillips, 1991). Although these methods are useful, the accuracy of the conclusions drawn from them may be affected by two common limitations. Specifically, examinations of discriminant validity can be less informative or accurate when researchers fail to compare a given focal construct to a broad collection of relevant extant constructs or when they estimate construct-level relationships without accounting for all relevant sources of measurement error. In the following, we explain these limitations in more detail.

Failing to Include Relevant, Conceptually Similar Constructs

The results yielded from assessments of discriminant validity may be more or less informative to the extent that the set of relevant constructs included in the analysis is comprehensive and appropriate. Campbell and Fiske (1959) note that discriminant validity assessments should show that there is no meaningful relationship between construct measures ''purporting to measure different things'' (p. 84). This suggestion is helpful but can be interpreted somewhat broadly. That is, the definition of what qualifies as ''different things'' may vary from study to study. We argue here that in the process of its development, if a proposed construct is not compared to a range of existing constructs that although *conceptually similar* are assumed to be *empirically dissimilar*, then it is difficult to conclude with any real certainty that the new construct is a unique one.

For example, consider a hypothetical situation in which a team of researchers has developed a new measure of workplace effectiveness that they call ''innovative intelligence.'' The researchers have settled on a theoretical definition of innovative intelligence that is distinct from general mental ability (GMA). While GMA refers to ''the ability to deal with cognitive complexity—in particular, with complex information processing'' (Gottfredson, 1997, pp. 92-93), the researchers believe that innovative intelligence represents a distinct ability to create novel solutions to novel problems. By making a conceptual distinction between the two constructs, the researchers have met the first requirement of establishing a new construct. To meet the second requirement, the researchers must examine the empirical distinctness of the innovative intelligence construct. To do so, they plan a construct validity study in which they will administer the newly developed measure of innovative intelligence and several other scales that are assumed to measure different constructs such as demographic variables and attitudinal variables (e.g., age, gender, job satisfaction, and turnover intentions). These variables share no theoretical relationships with innovative intelligence, and the researchers believe that the lack of any empirical relationship between the study variables will reveal

that the new measure of innovative intelligence shows discriminant validity. We argue that such an approach would result in a relatively uninformative test of discriminant validity. A more rigorous test would include measures of existing constructs that may already capture what the measure of innovative intelligence is theorized to uniquely capture, such as GMA or creativity. In the end, failing to include the most relevant constructs against which to compare a given focal construct may lead to the conclusion that the focal construct is unique when in fact it has only been compared to constructs with which it has no theoretical commonalities.

Failing to Correct Estimates of Construct-Level Relationships for Measurement Error

Constructs are operationalized through measures. Statistical relationships between observed scores on given measures serve as representations of the relationships between constructs. However, observed correlations do not necessarily provide an accurate estimate of such relationships because measures are imperfect representations of constructs. According to classical measurement theory, an observed score on a given measure is the sum of the construct score and measurement error. Three sources of measurement error are salient to virtually all areas of organizational research—random error, specific factor error, and transient error (Schmidt & Hunter, 1999). *Random error* occurs as the result of brief lapses in attention, momentary distractions, or other interruptions to thought processes that occur within a single administration of a given measure (Schmidt, Le, & Ilies, 2003). Variations in observed scores that occur as the result of guessing also fall into this category of measurement error (Thorndike, 1951). Random error is thought to be a relevant source of measurement error in most research. *Specific factor error* is measurement error that can be attributed to the wording of the instructions of a given measure or the wording of specific items in the measure (Thorndike, 1951). They "are produced by respondent-specific interpretation of the wording of questionnaire items" and "correspond to the interaction between respondents and items" (Schmidt et al., 2003, p. 209). Because random and specific factor error tend to average out across items, both of these sources of error can be controlled somewhat by extending the length of a given measure. *Transient error* is error that occurs due to temporal changes in the states of test subjects or testing conditions across separate administrations of a given test. Changes in subjects' health, fatigue levels, testing motivation, or general mood and affective state can introduce variance to test scores that are not related to the constructs of interest. Even conditions such as the ambient temperature of the testing location, noise levels, or lighting levels can affect observed scores on a given measure. When a construct is assumed to be stable over time, transient error should have only minor effects on the reliability of measures of that construct. Recent research, however, has shown that transient error may have meaningful effects on the measurement of relatively stable constructs such as personality and trait affect (Reeve, Heggestad, & George, 2005; Schmidt et al., 2003). Its effects on the measurement of less stable variables, such as work attitudes, may be even larger (Le et al., 2010).

If researchers do not account for measurement error, they end up investigating the observed relationships between construct measures rather than the relationships between the constructs themselves. In almost all cases, measurement error attenuates relationships and creates a downward bias in observed relationships between variables (Hunter & Schmidt, 2004). Because researchers are generally interested in the relationships between underlying constructs and not in the relationships between measures of those constructs, failing to take measurement error into account can lead to erroneous research conclusions. This concern may be especially relevant to evaluations of discriminant validity, the results of which can rest solely on the interpretation of the empirical relationships between constructs. Stated plainly, we argue that failing to take all reasonable sources of measurement error into account when evaluating the discriminant validity of a given construct may lead to the conclusion that it is an empirically distinct construct when it is not (Harter & Schmidt, 2008), thereby resulting in the proliferation of redundant constructs in the literature.

Guidelines for Assessing Discriminant Validity

We next suggest a set of guidelines for researchers seeking to assess the discriminant validity of a given construct or set of constructs. These recommendations can apply not only to research that attempts to establish a new construct but also to the evaluation of an existing body of research.

Step 1: Literature Review

Whether the construct of interest is a newly proposed construct or an existing construct, researchers should identify a wide range of theoretically related constructs for possible inclusion in a discriminant validity analysis. Thus, a critical component of any effort to establish the discriminant validity of a given construct is a literature review of disciplines in which it is reasonable to believe that theoretically similar constructs are contained (Gilliam & Voss, 2013). One threat to a high-quality literature review is a "false negative"—that is, the omission of a useful document that would improve the quality of a literature review (Grayson & Gomersall, 2003; White, 1994). False negatives can have a meaningful effect on studies of discriminant validity because they can result in the omission of a conceptually (or empirically) relevant construct from the discriminant validity analysis. To minimize the possibility of such omissions, reviews should span a broad, heterogeneous group of literature related to the construct(s) of interest (Schucan Bird & Tripney, 2011; Grayson & Gomersall, 2003; Hammerstrøm, Wade, & Jørgensen, 2010; Mehdyzadeh, 2004; Taylor, Wylie, Dempster, & Donnelly, 2007). Of course, the reality is that a perfectly comprehensive literature review may be possible only in theory—in practice, researchers cannot review the entirety of the available literature. For this reason, researchers should strive to be efficient in their reviews. For the purposes of a literature review designed to inform a discriminant validity analysis, the value of an exhaustive search is not in identifying every study containing a construct that is conceptually related to the focal construct. Rather, the goal of the literature review should be twofold: to identify a reasonably broad set of theoretically related constructs for inclusion in the discriminant validity analysis and to avoid omitting a relevant construct from the analysis because it lies outside the knowledge base of the researcher (White, 1994). To this end, previous scholars have recommended a variety of tactics to facilitate literature reviews that are both comprehensive and efficient.

First, in addition to general searches of citation databases, White (1994) recommends "backward" searches of the literature—searches that progress "from a known publication to the earlier items it cites"—and "forward" searches—those that proceed "by looking up a known publication ... and finding the items that later cite it" (p. 45). In the context of a discriminant validity analysis, it is important to note that some constructs may be developed explicitly as an extension or refinement of other constructs. Backward and forward searching may be a particularly useful way to identify the predecessors or descendants of such constructs. Relatedly, researchers can engage in what White calls "footnote chasing" (pp. 46-47), that is, reviewing the citation lists contained in key publications about a particular topic.

Second, researchers can also solicit suggestions from colleagues or subject matter experts (Gilliam & Voss, 2013). White (1994) refers to this as "consultation" and likens it to searching citation databases that "are simply inside people's heads" (p. 47). For example, a scholar who is interested in creating a new construct related to organizational justice can share the conceptual definition of the new construct with justice experts. These experts can provide an informed assessment of the potential overlap between this new construct and other existing constructs in the literature and suggest relevant constructs to be included in the empirical construct validation effort. Developing a personal and professional rapport with subject matter experts can make a meaningful difference in a scholar's ability to identify documents with information about constructs relevant to this type of study. Researchers can also solicit advice from a broader audience through the electronic mailing

lists maintained by professional organizations and interest groups such as the Academy of Management or the Society for Organizational and Industrial Psychology.

Third, scholars should consider studies from outside of the focal research domain because valuable information about whether a given construct is theoretically or empirically unique may lie outside the typical purview of an organizational researcher. In the history of science and scholarship, we find numerous examples of related research streams that advance without awareness of one another (e.g., Swales, 1986). Overlap between research streams can create opportunities for synergies when the streams are united, but this overlap can also contribute to the problem of construct proliferation. For example, organizational scholars, at different points in the past, may have imported constructs from domains such as educational or health psychology. Researchers also should be aware that constructs that are conceptually similar may exist under different names. Meta-analyses and published comprehensive literature reviews typically reference related constructs within a body of literature and can provide references to relevant primary studies. Content analysis can provide useful guidance along these lines (Short, Broberg, Cogliser, & Brigham, 2010).

To this point we have mentioned several tactics for conducting a literature review, but which of these tactics gives researchers the most return on their investments of time and effort? Schucan Bird and Tripney (2011) provide a possible answer to this question. Using a case analysis of a literature on culture and sport engagement, the authors suggest that when efficiency is defined as the number of unique items generated for each hour of literature searching, the four most efficient ways to a conduct literature review were to (a) search general bibliographic databases (8.6 unique items per hour), (b) review reference lists (6.6 unique items per hour), (c) search specialist bibliographic databases (5.4 unique items per hour), and (d) consult subject specialists (3.5 unique items per hour).

Step 2: Data Analysis

As mentioned previously, establishing the discriminant validity of a given construct requires a demonstration that the construct is not empirically identical to (or too highly correlated with) a construct from which it is theorized to be distinct. The analytic approach that a researcher chooses when estimating such relationships is critical to any study of discriminant validity because the accuracy of such estimates and the subsequent research conclusions drawn from them are dependent on the manner in which the data are analyzed. In the following, we discuss several methods that can be used to conduct a discriminant validity analysis and review approaches to accounting for measurement error. We note here that the focus of our review is not to provide detailed explanations of each method. Instead, we provide a basic overview of each method.

Multitrait-multimethod matrices. One method for assessing discriminant validity is to conduct a multitrait-multimethod (MTMM) analysis (Campbell & Fiske, 1959). Under the MTMM framework, different constructs are analogous to different traits. Different raters (e.g., two supervisors who independently rate the job performance of the same focal employee), different scales (e.g., the various available measures of Big Five personality traits), or different data collection time periods (e.g., the same construct scale administered to the same study participant at two different points in time) are all considered distinct methods (Conway, 1998). The construct validity of measures is examined by constructing an MTMM matrix that contains the correlations between the constructs of interest and evaluating ''the magnitudes of the correlations that are similar and dissimilar'' (Hinkin, 1998, p. 116). For large MTMM matrices, the number of comparisons to make is numerous (cf. Althauser & Heberlein, 1970; Marsh & Hocevar, 1983), making MTMM a potentially cumbersome method. While the traditional MTMM method has some beneficial features, it also has shortcomings (Marsh & Hocevar, 1983) and in recent years has fallen out of favor in discriminant validity research.

Confirmatory factor analysis. Discriminant validity can also be examined through the use of CFA. Hinkin (1998) speculated that CFA would supplant MTMM as the preferred method for conducting construct validity analyses, and his prediction has proven largely correct—CFA has become a frequently used tool in studies of convergent and discriminant validity (T. A. Brown, 2006; Hoyle, 2000). In the context of discriminant validity, CFA typically involves the comparison of two measurement models—an unconstrained model and a constrained model. In the unconstrained model, two latent variables that represent two conceptually similar constructs are allowed to freely covary with each other. In the constrained model, the covariance of these two latent variables is set to equal 1.0. The fit statistics of the two models are then compared. Often this comparison is made through the use of a chi-square (χ^2) difference test. If the unconstrained model provides a better fit than does the constrained model and the χ^2 difference between the two models is statistically significant, then an argument can be made that the constrained model demonstrates significantly *worse* fit to the data and therefore the two focal constructs are empirically distinct from one another. In contrast, if there is no significant difference between χ^2 for the two models, a researcher would likely conclude that the constructs are not empirically distinct. When a study includes data collected from a single administration of measures, CFA as described previously controls for random error and specific factor error.

CFA offers several advantages to researchers. First, if a study is designed appropriately, CFA can estimate construct-level relationships that account for random, specific factor, and transient error. To achieve this, study data should be collected such that parallel construct measures are administered to the same sample on two different occasions. With these data in hand, a CFA analysis can be conducted that specifies latent constructs represented by parallel construct measures from the separate test administrations. For example, if construct X is measured by Scale A (at Time 1) and Scale B (at Time 2) and construct Y is measured by Scale C (at Time 1) and Scale D (at Time 2), a latent factor indicated by Scale A and Scale B is specified to represent Construct X, and another latent factor indicated by Scale C and Scale D is specified to Construct Y. The correlation between the factors representing Construct X and Y is the estimate of the construct-level relationship between X and Y (cf. Figure 1 in Le, Schmidt, & Putka, 2009). If parallel construct measures are not available, the same measure for a given construct can be given at both test administrations. The latent construct can then be indicated by half of the measure administered in Time 1 and the other half of the measure administered in Time 2. For example, if construct X is measured by Scale A and Construct Y is measured by Scale C at both administrations, each scale can be divided into half-scales (A_1 and A_2 for construct X and C_1 and C_2 for construct Y), and the data can be analyzed such that half-scales $A_{1(\text{Time 1})}$ and $A_{2(\text{Time 2})}$ represent construct X and half-scales $C_{1(\text{Time 1})}$ and $C_{2(\text{Time 2})}$ represent construct Y. Alternately, $A_{2(\text{Time 1})}$ and $A_{1(\text{Time 2})}$ can be used to represent construct X or $C_{2(\text{Time 1})}$ and $C_{1(\text{Time 2})}$ can represent Construct Y. The half-scales can be used as described previously to conduct a CFA analysis that accounts for random, specific factor, and transient error. For a more detailed explanation of this method, we refer readers to Le et al. (2009, 2010). Second, because CFA programs have the flexibility to allow researchers to incorporate multiple sources of error into an analysis and model these types of errors with flexible constraints, CFA arguably improves on MTMM by allowing researchers to test a variety of hypotheses related to the distinctiveness of constructs in a way that MTMM does not. For example, a researcher can use CFA analyses to model different, specific factor structures that underlie constructs (e.g., in a study that includes three constructs, a researcher can compare a one-factor model with a two-factor model and a three-factor model). Researchers can also use CFA to test whether the relationship between two constructs differs from a specific level of correlation, such as .90 or .80. Finally, CFA can also be used to test for the presence of higher-order factors from which lower-level factors are derived.

In practice, CFA has two important limitations that are relevant to tests of discriminant validity. First, while CFA offers the apparent advantage of yielding hypothesis test statistics, interpretations

of research results that rely too heavily on statistical tests can be misleading. The limitations associated with statistical significance testing are not exclusive to CFA (Nickerson, 2000), but they are important to clarify in the context of a discriminant validity analysis. A typical CFA is conducted such that in the constrained model, the covariance between constructs is constrained to 1.0. This means that the χ^2 difference test can be likened to a binary test that compares models in which the focal constructs are either perfectly related or are not perfectly related. In reality, constructs are unlikely to be perfectly related. Thus, the conditions underlying a typical χ^2 difference test are rarely met. Second, sample size has a strong impact on the χ^2 difference test (Le et al., 2010), making it a potentially unreliable way to assess discriminant validity. Taken together, these limitations may cause a χ^2 difference test to suggest evidence of discriminant validity even when constructs are very highly correlated. For a more tangible example of how these two limitations could affect research results, consider an imaginary discriminant validity analysis conducted on a data set containing construct scores from 1,000 participants. Assume that the actual correlation between two constructs in the population is .95 and the correlation in the sample is also .95. In a sample of 1,000 participants, the confidence interval surrounding the sample correlation could be as narrow as .94 to .96. Because the sample size is large and correlation between the constructs is not 1.0, a χ^2 difference test will likely suggest that the two constructs are empirically distinct when, for all applied purposes, they are identical. Said another way, relying on significance tests to interpret research results may seem efficient and objective, but overreliance on significance testing can lead to research conclusions that are incorrect (Schmidt & Hunter, 1997).

Several alternatives to the χ^2 difference test have been recommended. We mention two of them here. First, simulation data provided by Meade, Johnson, and Braddy (2008) suggest that when the Comparative Fit Index (CFI) difference between two measurement models is .002 or less, the models may be considered equivalent. In the context of discriminant validity analyses, a CFI difference between a constrained model and an unconstrained model of .002 or less may suggest construct redundancy. Although examining the CFI difference between measurement models retains the limitation of holding construct relationships to 1.0 in the constrained model, Meade et al. (2008) showed that it (along with alternative fit indices such as the incremental fit index and root mean square error of approximation) was less affected by study sample size and was more likely to accurately assess discriminant validity. A second alternative is to examine the factor correlations between two constructs. Several authors have suggested that constructs can be considered as lacking discriminant validity when the factor correlation between them reaches a magnitude of .85 or higher (Kenny, 2012; van Mierlo, Vermunt, & Rutte, 2009).

Disattenuation formula. Obtaining estimates of construct-level relationships can also be accomplished using the disattenuation formula (Le et al., 2010), and these estimates can be used to determine whether constructs are empirically distinct. According to classical measurement theory, the observed correlation in the population between constructs x and y can be computed as follows (Equation 1 in Schmidt et al., 2003):

$$\rho_{xy} = \rho_{x_t y_t} \times \sqrt{\rho_{xx}\rho_{yy}}. \quad (1)$$

In this equation, ρ_{xy} represents the correlation between scores on the construct measures, $\rho_{x_t y_t}$ is the true relationship between the constructs themselves, and ρ_{xx} and ρ_{yy} are the reliabilities of the measures of x and y. This equation is known as the attenuation formula because it demonstrates how observed correlations are biased (in almost all cases they are downwardly biased or attenuated) relative to the value of true score correlations as a result of measurement error. This equation can be

rearranged to solve for $\rho_{x_t y_t}$. This rearranged formula is called the disattenuation formula and is shown here (Equation 2 in Schmidt et al., 2003):

$$\rho_{x_t y_t} = \frac{\rho_{xy}}{\sqrt{\rho_{xx}\rho_{yy}}}. \tag{2}$$

The correlations that are reported in research are not based on data from the true population but instead are based on sample data. Considerations for sampling error require correlations that are based on sample data to be denoted as $\hat{\rho}_{x_t y_t}$. This indicates that the correlations reported in research are in fact only estimates of the true score correlations that may be found in the population. Similarly, sample observed correlations are denoted as r_{xy}, and observed reliabilities are denoted as r_{xx} and r_{yy}. Taken together, these considerations require that the disattenuation formula be revised as shown in the following (Equation 3 in Schmidt et al., 2003):

$$\hat{\rho}_{x_t y_t} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}. \tag{3}$$

The disattenuation formula is valuable in the way that it allows a researcher to address multiple forms of measurement error. Random and specific factor error are typically assessed by computing the coefficient alpha for a measure (Cronbach, 1951). This estimate of reliability is also called the coefficient of equivalence (CE). CE tends to be the most frequently reported reliability estimate in organizational research. The extent to which the combination of random and transient error influences reliability can be estimated by correlating scores from the same scale administered at two different times. This estimate is known as the test-retest reliability of the scale or the coefficient of stability (CS; Schmidt et al., 2003).

When using the disattenuation formula, estimates of construct relationships that are corrected for random and specific factor error (but not transient error) can be obtained by inputting the coefficient alphas in the denominator portion of the disattenuation formula such that:

$$\hat{\rho}_{x_t y_t} = \frac{r_{xy}}{\sqrt{CE_{xx} CE_{yy}}}. \tag{4}$$

At this point, it is important to note that because estimates of CE do not account for transient error, they tend to overestimate the reliability of measures. Estimates of CS do not account for specific factor error and therefore also tend to overestimate the reliability of measures. Subsequently, estimates of construct-level relationships obtained by correcting for CE or CS will typically underestimate the empirical relationships between constructs.

The coefficient of equivalence and stability (CES) is the only reliability coefficient that estimates the combined biasing effects of random, specific factor, and transient error. Methods presented by Becker (2000) and implemented by Schmidt et al. (2003) outline procedures for comparing estimates of CES and CE to examine the extent to which transient error affects observed correlations between measures. Schmidt et al. refer to this as transient error variance (TEV), which can be estimated by subtracting CES from CE. In combination with procedures presented by Le et al. (2009), estimates of CES can be used to examine construct relationships underlying theoretical models in the organizational sciences.

Like estimates of CE, CES estimates can also be used in the denominator portion of the disattenuation formula to estimate construct-level relationships. But, unlike estimates of CE, estimates of CES account for all three types of measurement error. When two parallel forms of a measure are completed across test administrations, the correlation between scores of the parallel forms is the estimate of the CES (Le et al., 2009; Schmidt et al., 2003). Because the measures of the constructs will have been administered at two separate times, the researcher, in addition to having the data necessary for computing CES, will also have two sets of observed correlations between the constructs

measures—the observed correlations from Time 1 and those from Time 2. For each relationship between a given pair of measures, the average of the two observed correlations *across time periods* should be input into the numerator portion of the disattenuation formula (Le et al., 2009). Thus, if construct X is measured at Time 1 with scale A and measured at Time 2 with scale B and construct Y is measured at Time 1 with scale C and at Time 2 with scale D, the observed correlation that should be input into the numerator is actually the mean observed correlation across two different time periods, or $\bar{r}_{xy}$, where:

$$\bar{r}_{xy} = \frac{r_{AD} + r_{BC}}{2}. \tag{5}$$

Once $\bar{r}_{xy}$ and CES have been estimated for a pair of scales, researchers can estimate the construct-level correlations between the constructs the scales are intended to measure by using the disattenuation formula. Rewriting the disattenuation formula to reflect the appropriate use of $\bar{r}_{xy}$ and CES yields the following equation:

$$\hat{\rho}_{x_t y_t} = \frac{\bar{r}_{xy}}{\sqrt{CES_x CES_y}}. \tag{6}$$

If only one measure for a given construct is available, a researcher can still obtain estimates of CES. To do so, the researcher can administer the available scale to the same group of participants at two points in time. After the data have been collected, the scale can be split into two parallel half scales. When using the split-half approach, the following formula is used to estimate CES (Equation 16 in Schmidt et al., 2003):

$$CES = \frac{2[Cov(1a_1, 2a_2) + Cov(2a_1, 1a_2)]}{\sqrt{Var(1A)} \times \sqrt{Var(2A)}}. \tag{7}$$

In this equation,

> $Cov(1a_1,\ 2a_2)$ is the covariance between the half-scale a1 administered at Time 1 ($1a_1$) and the half-scale a_2 administered at time 2 ($2a_2$); $Cov(2a_1,\ 1a_2)$ is the covariance between the half-scale $a1$ administered at Time 2 ($2a_1$) and the half-scale $a2$ administered at Time 1($1a_2$); Var ($1A$) is the variance of the Full-Scale A administered at Time 1; and Var ($2A$) is the variance of the Full-Scale A administered at time 2. (Schmidt et al., 2003 p. 212)

When there is an odd number of items in a scale, researchers can use the following adjustment to this formula (Equation 17a in Schmidt et al., 2003):

$$CES = \frac{Cov(1a_1, 2a_2) + Cov(2a_1, 1a_2)}{2p_1p_2[\sqrt{Var(1A)} \times \sqrt{Var(2A)}]}. \tag{8}$$

Equation 8 differs from Equation 7 by the addition of $2p1p2$ to the denominator; " $p1$ is the ratio of number of items in the subscale a1 to that in the full scale; $p2$ is the ratio of number of items in the subscale a_2 to that in the full scale" (Schmidt et al., 2003, p. 212). For these equations Var (1A) and Var (2B) are given in Equations 18a and 18b in Schmidt et al. (2003). For informational purposes we repeat them here:

$$Var(1A) = \frac{Var(1a_1)[p_1 - (p_1 - 1)ce_1]}{{p_1}^2} \tag{9}$$

$$Var(2A) = \frac{Var(2a_2)[p_2 - (p_2 - 1)ce_2]}{{p_2}^2}. \tag{10}$$

Though there is no definitive point at which the correlations between two constructs is high enough to consider the constructs as empirically identical, John and Benet-Martínez (2000) suggest that when the construct-level relationship between two constructs reaches at least .90 in magnitude, it is likely the case that they cannot be empirically distinguished from one another. This suggestion aligns with findings from Le et al. (2010), who reported a corrected correlation between job satisfaction and organizational commitment of .91 and wrote that although job satisfaction and organizational commitment can be theoretically distinguished, "the very high correlation among the constructs . . . suggests that the constructs cannot be *empirically* distinguished in any practical sense in real research data" (p. 122).

The simplicity of the disattenuation formula can make it appealing, but it is not without its own limitations. The reader will recall that the disattenuation formula requires only three pieces of information—an observed correlation in the numerator and two reliability estimates in the denominator. The accuracy of output from the disattenuation formula hinges on the accuracy of its input. Accordingly, the primary threat to using the disattenuation formula is sampling error, which has been shown to have large effects on the results of studies with small samples (i.e., between 50 and 300 subjects as suggested in Hunter & Schmidt, 2004). Sampling error can have meaningful effects on reliability estimates and observed correlations from any single study. Subsequently, to the extent that sampling error biases any of the input values to the disattenuation formula, the output will be biased accordingly. Moreover, correcting correlations actually *increases* the effects of sampling error, as shown by the fact that the confidence intervals surrounding corrected correlations widen in direct proportion to the corrections that are made (Hunter & Schmidt, 2004). Even if study measures are perfectly reliable (a condition that is highly unlikely to be met in reality), sampling error can still bias study results. Schmidt and Hunter (1999) refer to measurement error as *systematic error* and sampling error as *unsystematic error*. Applying the disattenuation formula in individual studies allows researchers to account for systematic error, but it does not account for unsystematic error. One way to control for sampling error is to conduct studies on extremely large (or infinitely large!) samples, but this option is not always realistic. A more practical approach is to combine studies in a meta-analysis, which averages out the effects of sampling error across studies (Hunter & Schmidt, 2004; Schmidt & Hunter, 1999).

Step 3: Study Design

Once researchers have identified the constructs that are appropriate to include in their study and determined how they will analyze the data, they must consider design features of their study. There are several aspects of study design that we address here. Specifically, researchers must carefully consider the length of time between any repeated administrations, which scales to use in their study, and the desired study sample size.

Scale administration times. As described in the previous section, to obtain estimates of construct-level relationships that are corrected for the combination of random, specific factor, and transient error, construct measures need to be administered at two separate points in time. When administering study scales at two points in time, the interval of time between scale administrations must be carefully selected. If the interval is too long, then substantive changes may occur to the focal construct between the first and second administration of the survey measures, which could result in inflated estimates of the effects of transient error on observed scores between measures. An important consideration when choosing the interval between scale administrations is the stability of the construct of interest. Some constructs are more stable than others: General mental ability is considered highly stable over long periods of time, while affective states and job attitudes are thought to be less stable (Judge, Higgins, Thoresen, & Barrick, 1999; Reeve & Bonaccio, 2011; Staw, Bell, & Clausen,

1986). In the end, researchers should schedule scale administrations such that it is unlikely for substantive changes in the focal constructs to have occurred between administrations. The length of time that transpires between scale administrations should be based on theoretical justifications available in the literature (Le et al., 2010). For example, in estimating the effects of transient error on personality trait measures, a period of up to two months between scale administrations may be appropriate because substantive changes in personality traits should not occur over this interval. Thus, discrepancies between personality trait scores during this time period are likely the result of transient error (Watson & Humrichouse, 2004). In contrast, a shorter interval may be necessary when the focal construct is attitudinal in nature. We also recommend that during the second scale administration, researchers ask study participants whether meaningful changes pertaining to the focal constructs have occurred between administrations. Participants who indicate that meaningful changes have occurred should be considered for exclusion from the final study sample because including them in the final sample may result in an overestimate of the effects of transient error on the final study results (and, subsequently, an overestimate of the strength of construct-level relationships). A reviewer also suggested that researchers should be mindful of any changes in the environment in which their study takes place. For example, changes to organizational structure, policies, or personnel may result in substantive changes to focal constructs that could complicate an assessment of discriminant validity.

Study scales. As mentioned previously, researchers interested in obtaining fully corrected estimates of construct-level relationships should administer different measures of the same constructs or parallel forms of the constructs across time periods. When several scales for a given construct are available, it is preferable to use this approach (Schmidt et al., 2003). In some instances, multiple measures of the focal constructs may not exist. For a new construct, it is almost certainly the case that no validated scale exists to measure that construct, and the researcher will first need to develop a scale to measure the construct (see Hinkin, 1998, for a review of how to develop new scales). But even for well-known constructs it is possible that only one validated scale is available in the literature. As explained previously, in such situations, researchers can administer the same scale at two points in time and then create two parallel measures of the focal construct by splitting the full scale, post hoc, into two half scales from which the CES of the scale can be estimated. A relatively straightforward way to create half scales is to divide items across scales based on their content, but the process can be more complex if needed or desired. Becker (2000) offers suggestions for splitting scales into equivalent halves that, in addition to grouping items based on their content, include dividing items across half scales based on their means, standard deviations, factor loadings, and scoring direction (i.e., standard vs. reverse-scored items).

Sample size. Although the common advice to have as large of a sample as possible may seem cliché, in light of several empirical considerations, this advice takes on a greater level of importance in studies of discriminant validity. First, because accounting for random, specific factor, and transient error requires collecting data at two separate points in time, it is probable that attrition between scale administrations will reduce the final sample size of such studies. Dillman, Smyth, and Christian (2009) suggest that attrition in longitudinal or panel studies is nearly inevitable and emphasize that attrition can be especially damaging when the participants who drop out of the study differ in some systematic way from those who remain enrolled in the study. To combat attrition, we recommend that researchers should (a) follow the advice of Ployhart and Vandenberg (2010) to estimate the amount of attribution likely to occur by reviewing previous studies conducted in similar domains to determine the size of the initial sample based on the final sample that is needed and (b) determine whether any attrition that occurs is systematic or random in nature. Second, factor analysis requires large sample sizes, especially when there are many scales and/or items to be included in the analysis.

Thus, the size of the sample should be proportional to the number of scales/constructs included in the study. Generally speaking, the more constructs/scales included, the larger the required sample. Some scholars recommend the ratio of sample size to number of items should be 10 to 1; however, this recommendation can be unnecessarily prescriptive in many cases (MacCallum, Browne, & Cai, 2006). Calculating power using estimates of relationships based on prior evidence is also valid in terms of determining sample size (Cohen, 1992; MacCallum, Browne, & Sugawara, 1996), and there are useful online tools for doing so (e.g., Faul, Erdfelder, Lang, & Buchner, 2007; Lenth, 2009).

Step 4: Substantive Interpretation of Results

If study results suggest a lack of discriminant validity between a pair of constructs, researchers must consider what the results could mean. We offer three potential interpretations of such results.

The constructs are unique, but the scale items contained in the construct measures are too similar. One possibility is that the constructs of interest are distinct, but the scales used to assess the constructs do not differentiate between them (cf. Cole et al., 2012; Hershcovis, 2011). The greater the number of scale items that are alike across construct measures, the higher the empirical relationship between those constructs is likely to be. Sometimes scale items across measures are similar because the constructs explicitly share conceptual overlap (e.g., Avolio & Gardner, 2005; Conger & Kanungo, 1994). At other times, the similarity between scale items may be unintentional. From a measurement standpoint, the implications of this possibility are that researchers may need to reevaluate the construct measures of interest and make an effort to create items and measures that more clearly delineate constructs from one another. Relatedly, it may be the case that respondents—even when researchers believe that scale items across construct measures are sufficiently distinct from one another—do not make the kinds of nuanced judgments that are necessary to produce empirical distinctions between constructs. Harter and Schmidt (2008) argue that researchers may sometimes work with the belief that if they ''can make a logical or conceptual distinction between constructs or measures, then this distinction will exist in the minds of employees or respondents to surveys'' (p. 36). Based on the available evidence, the assumption that respondents make such distinctions seems unlikely. People tend to form quick, general impressions that are based more on intuition than on rational thought, and these intuitive impressions can have a strong influence on later judgments (Ambady & Rosenthal, 1992; Barrick, Swider, & Stewart, 2010).

The constructs share a causal relationship. A second interpretation is that the constructs are unique, but changes in one construct lead to corresponding changes in the other construct. Although this is certainly a possibility, this interpretation can be problematic because it would mean that one of the constructs entirely, or almost entirely, causes the other (Le et al., 2010). To examine this possibility further, the researcher would need to collect longitudinal data or experimental data. Other scholars have provided information on how to test for causal relationships, and we refer the interested reader to those authors (e.g., Harter, Schmidt, Asplund, Killham, & Agrawal, 2010; James, Muliak & Brett, 1982).

The constructs are empirically redundant. Finally, it is possible that highly related constructs are empirically redundant and that any proposed theoretical distinction between them is spurious. From a theoretical perspective, evidence of construct redundancy can have implications for taxonomies and models of organizational phenomena. Such findings may suggest that the established theories and taxonomies outlining the interrelationships and casual connections between the constructs of interest are in need of revision. We realize that such an interpretation might seem foreboding—it could mean that much work is needed to ''clean up'' existing theoretical domains. However, we believe that

evidence of empirical redundancy could help advance the field of organizational research. Pfeffer (1993) argues that scientific fields or disciplines with more advanced levels of paradigm development are characterized by a high level of theoretical and methodological consensus. If paradigm development rests with knowledge integration rather than knowledge division, as Pfeffer suggests, then the determination that two supposedly distinct constructs are empirically redundant may represent a step toward achieving such integration.

Empirical Example

Background

The context of our empirical example is the leadership domain, which we chose because the issue of construct proliferation seems to be especially salient to this area of research. Our review of the leadership literature suggests that numerous scholars have voiced concerns over the extent to which construct proliferation has crept into this area of study. As early as 1977, Pfeffer observed that leadership concepts suffered from conceptual and operational ambiguities. Later, in an extensive review, Fleishman et al. (1991) outlined numerous taxonomies of leadership behavior. Yukl, Gordon, and Taber (2002) remarked that the proliferation of leadership theories was "bewildering," stating:

> Sometimes different terms have been used to refer to the same type of behavior. At other times, the same term has been defined differently by various theorists. What is treated as a general behavior category by one theorist is viewed as two or three distinct categories by another theorist. What is a key concept in one taxonomy is absent from another. Different taxonomies have emerged from different disciplines, and it is difficult to translate from one set of concepts to another. (p. 15)

More recently, DeRue et al. (2011) reported evidence of conceptual and empirical overlap among prevalent leadership constructs, noting that "new leader behavior theories continue to be conceived without explicit comparison to or falsification of existing leader behavior theories" (p. 15).

Our empirical study included a total of 12 leadership constructs: transformational leadership, contingent reward, passive and active management by exception, laissez-faire leadership, charismatic leadership, leader-member exchange, initiating structure, consideration, ethical leadership, abusive supervision, and servant leadership. We also included a measure of leadership effectiveness. We chose to include these leadership constructs for three reasons. First, our review suggested that leadership studies that assess the construct validity of leadership constructs typically fail to establish appropriately the discriminant validity of the constructs. For example, abusive supervision has not yet been compared empirically to a comprehensive set of related constructs. This same pattern can be observed in the development of charismatic leadership (Conger, Kanungo, Menon, & Mathur, 1997), transformational leadership (Tejeda, Scandura, & Pillai, 2001), and servant leadership theories (Page & Wong, 2000). Second, studies that have examined the discriminant validity of new leadership constructs do not always account for measurement error (Liden, Wayne, Zhao, & Henderson, 2008; Rowold & Heinitz, 2007). Third, some studies have considered the effects of measurement error but have corrected only for specific factor and random error (e.g., Avolio, Bass, & Jung, 1999; M. E. Brown, Treviño, & Harrison, 2005; Carless, 1998; Carless, Wearing, & Mann, 2000; Heinitz, Liepmann, & Felfe, 2005; Yukl et al., 2002).

Participants and Procedures

Our study sample consisted of full-time working adults who were recruited through Qualtrics (a marketing research company that maintains a variety of survey response panels) and were paid for their

participation. Planning in advance for study attrition, we oversampled at Time 1 to ensure that we would have enough respondents at Time 2 to conduct our analysis. Specifically, we invited twice as many people to participate as were needed for our final sample. Participants were paid at both collection times. We stopped collecting data at Time 2 once we reached the end of the budget that we had allotted for this project. Four hundred and ten participants completed an initial electronic survey containing 13 leadership scales. Two weeks after the initial survey, 220 of the original respondents completed a second administration of the same questionnaire. We examined the mean age, gender, and Time 1 construct scores for each group of study participants (i.e., those completing the study questionnaire only at Time 1 and those completing the questionnaire at both Time 1 and Time 2) and found no meaningful differences between them. Our final sample size was reduced to 185 based on respondents' answers to two survey items. First, we asked respondents whether their relationship with their supervisor had changed significantly over the past two weeks. Twenty-one respondents indicated that their relationship had changed and were excluded from our final data set. Second, we included an attention check in the survey that asked respondents to select "not at all" for that item. Fourteen participants responded incorrectly and were excluded from our analysis. The sample was 39.5% male and had an average age of 38.21 years ($SD = 10.37$). The participants reported working in a wide range of industries (e.g., banking, construction, education, health care, information technology, manufacturing, retail, social services, and transportation).

Measures

Transformational leadership. We used the short form of the MLQ-5X (Avolio et al., 1999) to measure the four dimensions of transformational leadership (TFL). This measure consists of 36 items, 20 of which assess the four leader behavioral dimensions that underlie TFL: idealized influence, individualized consideration, inspirational motivation, and intellectual stimulation. We used these 20 items to construct a single scale measuring overall TFL. Of the remaining items in the MLQ-5X, 16 of them assess contingent reward (CR), management by exception-active (MBEA), management by exception-passive (MBEP), and laissez-faire (LF) leadership behaviors. Each of the scales for these leadership dimensions consists of 4 items. Each item in the MLQ-5X describes a leadership behavior, and respondents were asked to indicate how frequently their manager or immediate supervisor engaged in each behavior on a scale from 1 (*not at all*) to 5 (*frequently, but not always*).

Charismatic leadership. We measured charismatic leadership with the Conger-Kanungo Scale (Conger, Kanungo, & Menon, 2000). This scale consists of 20 items. Respondents were instructed to indicate on a scale of 1 to 5 how accurately each item described their manager or immediate supervisor, with higher scores representing higher perceptions of charismatic leadership.

Leader-member exchange. Leader-member exchange was measured with the LMX-7 (Graen & Uhl-Bien, 1995). The response anchors for the LMX-7 vary from item to item. For example, participants were asked to respond to the item "How well does your supervisor recognize your potential?" on a scale of 1 (*not at all*) to 5 (*fully*), to the item "How would you characterize your working relationship with your supervisor?" on a scale of 1 (*extremely ineffective*) to 5 (*extremely effective*), and to the item "How well does your supervisor understand your job problems and needs?" on a scale of 1 (*not a bit*) to 5 (*a great deal*).

Initiating structure and consideration. We measured these two leadership constructs using the Leader Behavior Description Questionnaire (Form XII; Stogdill, 1963). Each of the scales for these constructs consists of 10 items. Participants were asked to report how often their leader engaged in the behaviors described in each item on a scale that ranged from 1 (*never*) to 5 (*always*).

Ethical leadership. To assess ethical leadership, we used the scale developed by M. E. Brown et al. (2005). Participants were instructed to describe how well each of the scale's 10 statements described their manager or immediate supervisor on a scale that ranged from 1 (*strongly disagree*) to 5 (*strongly agree*).

Servant leadership. We measured servant leadership with the 14-item scale developed by Ehrhart (2004). Participants were asked to indicate the extent to which each of the items accurately described their leader or supervisor on a scale of 1 (*to a very small extent*) to 5 (*to a very good extent*).

Abusive supervision. To measure abusive supervision, we used Tepper's (2000) 15-item scale. Participants were asked to indicate how often their supervisor engaged in the behaviors described in each of the items on a scale of 1 (*I cannot remember him/her ever using this behavior with me*) to 5 (*He/she uses this behavior very often with me*).

Leader effectiveness. We assessed effectiveness with five items. Three of these items—"Overall, to what extent is the supervisor performing his/her job the way you would like it to be performed?" "To what extent has s/he met your own expectations in his/her leadership roles and responsibilities?" and "If you had your way, to what extent would you change the manner in which s/he is doing the job?"—were taken from Tsui (1984). Two items—"Overall, do you like your supervisor?" and "Overall, are you satisfied with your supervisor?"—were adapted from Judge and Bono (2000). Participants responded to these items on a 7-point scale ranging from 1 (*not at all*) to 7 (*to the fullest extent*). We combined the five items into a single measure of leadership effectiveness.

Data Analysis Strategy

As explained previously, the most appropriate way to examine the relationships between the leadership measures is to do so while accounting for random, specific factor, and transient error. For comparison, we present results from both CFA and the disattenuation formula. To apply the disattenuation formula, we needed to compute estimates of CES for each study measure. To do so, we used the split-half approach outlined by Schmidt et al. (2003). When constructing the half-scales for each measure, we tried to ensure that the halves were similar in content. For scales that contained several dimensions (e.g., transformational leadership, charismatic leadership), we split items from each dimension between the two half scales. If a measure contained reverse-scored items, we split the items between half scales. Using the CES for each measure as input for the denominator portion of the disattenuation formula, we estimated the true score correlations among the leadership constructs using Equation 6, shown here again for the sake of clarity:

$$\hat{\rho}_{x_t y_t} = \frac{\bar{r}_{xy}}{\sqrt{CES_x CES_y}}, \tag{11}$$

where

$$\bar{r}_{xy} = \frac{r_{(Construct\ A^{Time1}, Construct\ B^{Time2})} + r_{(Construct\ A^{Time2}, Construct\ B^{Time1})}}{2}. \tag{12}$$

Equation 11 uses the average of the correlations from each full scale *across test administrations* as input into the disattenuation formula. For the CFA analyses, we examined each possible pairing of study constructs separately.[1] Specifically, for each construct pair we examined an unconstrained model and a model in which the covariance between the construct pair was constrained to 1.0. To control for all the sources of measurement error, each construct was represented by half of the scale items administered in Time 1 and another half of the scale items administered in Time 2.

Results

Table 1 presents the means, standard deviations, estimates of CE, and observed correlations for each of the study measures across both administrations. Table 1 also reports the test-retest reliability (CS) for each measure. As shown in Table 1, CE estimates at Time 1 ranged from .65 (management by exception-active) to .96 (transformational leadership, charismatic leadership, servant leadership, and abusive supervision), with an average of .89 across all measures. The CE estimates at Time 2 ranged from .73 (management by exception-active) to .97 (abusive supervision), with an average of .90. CS estimates ranged from .62 (management by exception-active) to .88 (consideration and leader effectiveness), with an average of .79. The correlations reported in Table 1 ranged from .89 to –.68. The strength of none of the observed correlations from Time 1 or Time 2 was equal to or greater than .90.

Table 2 shows the CES estimates for each of the leadership measures. The CES estimates ranged from .60 (management by exception-active) to .88 (ethical leadership), with an average of .77. We estimated TEV for each leadership measure by subtracting the CES estimate for a given measure from the average of the CE estimates for that measure from Time 1 and Time 2 (Schmidt et al., 2003). The extent to which the CE overestimated the reliability of the leadership measures ranged from 4.77% (consideration) to 28.72% (initiating structure), with an average overestimate of 16.22%. These results suggest that on average, CE overestimated the reliability of the leadership measures.

Correcting based on estimates of CES

For the analyses based on applying estimates of CES in the disattenuation formula (Equation 6), the observed correlations that we used in the numerator portion of the formula are shown in Table 3. Again, these correlations represent correlations across test administrations for each construct pair. Table 4 shows the corrected correlations between the study variables. The upper diagonal shows correlations that were corrected based on the CES estimates. Following the cutoff suggested by John and Benet-Martínez (2000), we considered correlations of .90 or greater in magnitude as suggesting a lack of discriminant validity.

In the left side of Table 5 we present CFA results that account for all three sources of measurement error. Here we report three CFA statistics. First, we present the χ^2 difference ($\Delta\chi^2$) between the unconstrained and the constrained models for each construct pair. Because between these two models there was a difference of only 1 degree of freedom, $\Delta\chi^2$ between the models reached statistical significance when it was 3.841 or greater. A $\Delta\chi^2$ that is statistically significant may be interpreted as indicating an empirical difference between models and the presence of two distinct constructs. Second, we present the CFI difference (ΔCFI) between the models. As recommended by Meade et al. (2008), a ΔCFI difference that is greater than .002 suggests an empirical difference between models and construct distinctiveness. Third, we report the factor correlations between each construct pair. Following Kenny (2012) and van Mierlo and colleagues (2009), we considered factor correlations that reached a magnitude of .85 or greater to suggest a lack of discriminant validity.[2] Alongside the CFA results, we include the disattenuated correlations from Table 4 for comparison purposes.[3] Finally, we present a summary statistic that denotes the number of discriminant validity indices for each construct pair that meet or exceed the cutoff values and therefore suggest a lack of distinctiveness between constructs.

Correcting based on estimates of CE

For illustrative purposes, we also present results from the disattenuation formula and CFA that account only for random and specific factor error (that is, they are corrected based on estimates of CE). For many studies, data are collected at only one time period. Therefore, to offer an illustration that most closely reflects typical research results, we used only the data collected at Time 1 of

Table 1. Descriptive Statistics and Correlations for Study Variables.

	Time 1																Time 2			
	M	SD	α	1	2	3	4	5	6	7	8	9	10	11	12	13	M	SD	α	Test-Retest Reliability
1. TFL	3.43	.79	.96		.89	−.18	−.45	−.52	.84	.85	.64	.83	.85	.86	−.54	.80	3.47	.83	.96	.83
2. CR	3.51	.87	.85	.89		−.12	−.40	−.49	.77	.82	.65	.79	.79	.80	−.52	.78	3.56	.93	.89	.78
3. MBE-A	2.91	.81	.65	.01	.02		.27	.22	−.17	−.25	−.06	−.26	−.18	−.21	.39	−.31	2.81	.87	.73	.62
4. MBE-P	2.44	.94	.76	−.33	−.32	.25		.78	−.38	−.45	−.43	−.57	−.48	−.49	.40	−.50	2.51	.94	.79	.70
5. LF	2.14	.99	.84	−.37	−.40	.29	.78		−.46	−.55	−.56	−.66	−.60	−.56	.52	−.58	2.17	1.02	.87	.72
6. CHAR	3.42	.83	.96	.77	.71	−.08	−.31	−.30		.84	.68	.75	.84	.86	−.55	.81	3.39	.86	.96	.78
7. LMX	3.65	.79	.90	.79	.75	−.15	−.43	−.47	.74		.66	.84	.86	.86	−.66	.86	3.58	.89	.92	.85
8. STR	3.82	.69	.90	.63	.64	.05	−.46	−.45	.62	.66		.63	.74	.72	−.51	.66	3.79	.72	.92	.76
9. CONS	3.60	.80	.92	.75	.71	−.24	−.53	−.52	.73	.81	.68		.85	.85	−.68	.85	3.54	.82	.91	.88
10. ETH	3.83	.79	.94	.80	.78	−.14	−.48	−.53	.79	.81	.73	.87		.88	−.66	.84	3.81	.86	.95	.86
11. SERV	3.44	.97	.96	.77	.73	−.10	−.40	−.39	.81	.79	.66	.83	.83		−.62	.83	3.41	.99	.96	.87
12. ABUS	1.46	.72	.96	−.47	−.47	.31	.37	.49	−.45	−.53	−.47	−.63	−.60	−.50		−.67	1.52	.78	.97	.79
13. EFCT	5.17	1.37	.92	.78	.73	−.24	−.50	−.53	.74	.82	.64	.82	.83	.81	−.60		5.10	1.41	.91	.88

Note: $N = 185$. Time 1 (Time 2) correlations are shown below (above) the diagonal. TFL = transformational leadership; CR = contingent reward; MBE-A = management by exception-active; MBE-P = management by exception-passive; LF = laissez-faire; CHAR = charismatic leadership; LMX = leader-member exchange; STR = initiating structure; CONS = consideration; ETH = ethical leadership; SERV = servant leadership; ABUS = abusive supervision; EFCT = leader effectiveness.

Table 2. Comparisons of Reliability Coefficients and Proportions of Transient Error Variances.

Construct	CE	CES	TEV	% Overestimate
Transformational leadership	.96	.80	.16	20.04
Contingent reward	.87	.73	.13	18.30
Management by exception (active)	.69	.60	.10	16.11
Management by exception (passive)	.78	.67	.10	15.36
Laissez-faire leadership	.86	.70	.16	22.92
Charismatic leadership	.96	.79	.17	22.00
Leader-member exchange	.91	.81	.10	12.15
Initiating structure	.91	.71	.20	28.72
Consideration	.91	.87	.04	4.77
Ethical leadership	.94	.88	.07	7.58
Servant leadership	.96	.87	.09	10.59
Abusive leadership	.96	.76	.20	26.67
Leadership effectiveness	.91	.86	.05	5.63

Note. $N = 185$. CE = average coefficient of equivalence (alpha) across Time 1 and Time 2; CES = coefficient of equivalence and stability; TEV = transient error variance over observed scores variance; % Overestimate = percentage that CES is overestimated by CE.

Table 3. Observed Correlations Across Test Administrations Used for Correcting for CES in the Disattenuation Formula.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. TFL		.75	−.09	−.37	−.42	.75	.75	.64	.72	.77	.77	−.53	.73
2. CR	.77		−.08	−.32	−.42	.70	.75	.64	.69	.74	.72	−.54	.71
3. MBE-A	−.14	−.06		.13	.14	−.14	−.17	.03	−.18	−.08	−.09	.28	−.25
4. MBE-P	−.44	−.40	.23		.69	−.38	−.44	−.42	−.48	−.46	−.44	.33	−.50
5. LF	−.44	−.41	.21	.57		−.40	−.48	−.48	−.52	−.51	−.46	.40	−.54
6. CHAR	.77	.69	−.15	−.34	−.38		.69	.61	.63	.73	.77	−.43	.66
7. LMX	.78	.74	−.16	−.37	−.49	.74		.62	.76	.76	.77	−.55	.77
8. STR	.58	.60	−.01	−.44	−.53	.55	.60		.60	.64	.63	−.43	.57
9. CONS	.82	.74	−.24	−.50	−.60	.72	.79	.59		.77	.80	−.61	.77
10. ETH	.82	.75	−.17	−.43	−.55	.75	.78	.70	.80		.81	−.58	.76
11. SERV	.80	.76	−.19	−.41	−.48	.78	.78	.64	.78	.80		−.56	.72
12. ABUS	−.51	−.46	.29	.41	.52	−.47	−.56	−.47	−.63	−.56	−.55		−.55
13. EFCT	.80	.76	−.28	−.47	−.56	.78	.83	.65	.81	.82	.80	−.65	

Note: $N = 185$. Observed correlations between Construct A at Time 1 and Construct B at Time 2 (Construct A at Time 2 and Construct B at Time 1) are shown in the upper (lower) diagonal. TFL = transformational leadership; CR = contingent reward; MBEA = management by exception-active; MBEP = management by exception-passive; LF = laissez-faire; CHAR = charismatic leadership; LMX = leader-member exchange; STR = initiating structure; CONS = consideration; ETH = ethical leadership; SERV = servant leadership; ABUS = abusive supervision; EFCT = leader effectiveness.

our study for these analyses. The lower diagonal of Table 4 shows correlations that were corrected based on the CE estimates, and the right side of Table 5 includes CFA results for this analysis.

Differences in results across indices and correction methods

Overall, our results suggest that any conclusions about the empirical distinctiveness of the leadership constructs that we examined may be meaningfully different based on whether corrections for

Table 4. Comparison of Correlations Corrected for CE and CES.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. TFL		**.99**	−.17	−.55	−.58	**.96**	**.95**	.82	**.92**	**.95**	**.94**	−.67	**.92**
2. CR	**.99**		−.11	−.51	−.58	**.91**	**.97**	.86	.89	**.93**	**.92**	−.67	**.93**
3. MBE-A	.01	.03		.28	.27	−.21	−.24	.02	−.29	−.17	−.20	.43	−.37
4. MBE-P	−.39	−.40	.36		**.92**	−.50	−.55	−.62	−.64	−.58	−.56	.52	−.64
5. LF	−.41	−.48	.39	**.97**		−.53	−.65	−.72	−.72	−.68	−.60	.64	−.71
6. CHAR	.80	.78	−.10	−.36	−.33		**.90**	.78	.81	.89	**.93**	−.58	.87
7. LMX	.85	.86	−.20	−.52	−.54	.80		.81	**.92**	**.91**	**.92**	−.70	**.96**
8. STR	.68	.74	.06	−.56	−.52	.67	.74		.76	.85	.81	−.61	.78
9. CONS	.80	.81	−.31	−.63	−.59	.78	**.90**	.75		**.90**	**.91**	−.76	**.91**
10. ETH	.84	.88	−.18	−.57	−.60	.84	.89	.80	**.94**		**.92**	−.70	**.91**
11. SERV	.81	.81	−.12	−.47	−.43	.84	.85	.71	.89	.87		−.69	.88
12. ABUS	−.49	−.52	.40	.43	.54	−.47	−.57	−.51	−.67	−.64	−.52		−.74
13. EFCT	.83	.83	−.31	−.60	−.60	.79	**.90**	.71	**.90**	.89	.86	−.64	

Note: $N = 185$. Correlations in the lower diagonal are Time 1 correlations corrected using CE estimates from Time 1. Correlations in the upper diagonal are correlation across both time periods corrected using CES estimates. TFL = transformational leadership; CR = contingent reward; MBEA = management by exception-active; MBEP = management by exception-passive; LF = laissez-faire; CHAR = charismatic leadership; LMX = leader-member exchange; STR = initiating structure; CONS = consideration; ETH = ethical leadership; SERV = servant leadership; ABUS = abusive supervision; EFCT = leader effectiveness. Correlations with a magnitude of equal to or greater than .90 are shown in bold.

measurement error are applied to the correlations and the extent to which such corrections are applied. For example, if we interpret our results based on the observed correlation matrices, such as those presented in Tables 1 and 3, we would likely conclude that each leadership measure assessed a unique, empirically distinct construct because none of the correlations between constructs reached or exceeded a magnitude of .90. If we interpret our results based on the correlation matrices derived from corrections for measurement error as estimated by the CE (the lower diagonal of Table 4), we might conclude that because only 6 of the 78 correlations were equal to or greater than .90 in magnitude, most of the leadership measures assessed empirically distinct constructs. The correlation matrix based on corrections for measurement error as estimated by the CES (the upper diagonal of Tables 4) suggests an entirely different conclusion. Twenty-four of the 78 correlations were equal to or greater than .90. From these correlations we would likely conclude that many of the leadership measures included in our study assessed the same construct. For example, the upper diagonal of Table 4 shows corrected correlations between transformational leadership and contingent reward, charismatic leadership, LMX, consideration, ethical leadership, and servant leadership of .99, .96, .95, .92, .95, and .94, respectively.

The results of our analysis may also differ depending on the summary statistic from which study conclusions are drawn. In our fully corrected results shown on the left side of Table 5, $\Delta\chi^2$ was nonsignificant for only 9 comparisons. This suggests virtually no empirical overlap between any of the leadership constructs save the measure of leadership effectiveness. Thus, solely examining the $\Delta\chi^2$ between models would likely lead to the conclusion that virtually all of the leadership scales assess a unique construct. On the other hand, ΔCFI was .002 or less for 32 of the construct comparisons, suggesting substantial empirical overlap between the leadership constructs. Twenty-nine of the factor correlations reached at least .85 in magnitude, while 24 of the disattenuated correlations reached at least .90 in magnitude. In both cases, the large number of construct correlations exceeding the specified threshold suggests that many of these constructs were not empirically distinct from each other.

Table 5. Comparison of Discriminant Validity Indices.

	Corrected for Random, Specific Factor, and Transient Error (Time 1 and Time 2 Data Combined)					Corrected for Random and Specific Factor Error (Time 1 Data Only)				
	(1)	(2)	(3)	(4)	(5) Indices Suggesting Lack of Discriminant	(1)	(2)	(3)	(4)	(5) Indices Suggesting Lack of Discriminant
Construct Pair	$\Delta\chi^2$	ΔCFI	FC	ρ	Validity	$\Delta\chi^2$	ΔCFI	FC	ρ	Validity
TFL - CR	15.30	**.001**	**.98**	**.99**	3	12.14	**.001**	**.99**	**.99**	3
TFL - MBE-A	150.46	.012	−.42	−.17	0	93.54	.006	−.11	.01	0
TFL - MBE-P	170.20	.013	−.56	−.55	0	191.83	.015	−.49	−.39	0
TFL - LF	201.54	.015	−.58	−.58	0	163.18	.013	−.45	−.41	0
TFL - CHA	20.40	**.001**	**.91**	**.96**	3	12.29	**.001**	.82	.80	1
TFL - LMX	25.17	**.002**	**.92**	**.95**	3	20.45	**.001**	**.86**	.85	2
TFL - STR	27.72	**.002**	.72	.82	1	47.53	**.002**	.73	.68	1
TFL - CON	10.31	**.000**	**.91**	**.92**	3	13.23	**.000**	.83	.80	1
TFL - ETH	23.59	**.001**	**.91**	**.95**	3	16.92	**.001**	.84	.84	1
TFL - SERV	5.50	**.000**	**.91**	**.94**	3	5.68	**.000**	.81	.81	1
TFL - ABUS	354.17	.020	−.59	−.67	0	406.98	.022	−.49	−.49	0
TFL - EFCT	**.01**	**.000**	**.88**	**.92**	4	**.62**	**.000**	.82	.83	2
CR - MBE-A	114.19	.030	−.38	−.11	0	85.57	.023	−.04	.03	0
CR - MBE-P	177.77	.049	−.54	−.51	0	173.39	.048	−.49	−.40	0
CR - LF	188.50	.054	−.56	−.58	0	158.71	.045	−.49	−.48	0
CR - CHA	15.01	**.001**	**.88**	**.91**	3	15.35	**.001**	.81	.78	1
CR - LMX	13.19	**.002**	**.92**	**.97**	3	24.10	.004	**.87**	.86	1
CR - STR	13.54	**.001**	.75	.86	1	47.55	.005	.77	.74	0
CR - CON	**3.31**	**.000**	**.90**	.89	3	18.00	**.002**	.82	.81	1
CR - ETH	17.44	**.001**	**.90**	**.93**	3	19.45	.003	**.88**	.88	1
CR - SERV	**.92**	**.000**	**.89**	**.92**	4	8.33	**.001**	.82	.81	1
CR - ABUS	232.42	.028	−.59	−.67	0	*209.79*	*.024*	*−.51*	*−.52*	0
CR - EFCT	**2.33**	**.000**	**.88**	**.93**	4	**2.28**	**.000**	.83	.83	2
MBE-A - MBE-P	61.16	.020	.38	.28	0	49.73	.016	.42	.36	0
MBE-A - LF	37.80	.013	.39	.27	0	31.18	.010	.42	.39	0
MBE-A - CHA	144.48	.011	−.38	−.21	0	105.09	.008	−.19	−.10	0
MBE-A - LMX	156.38	.028	−.43	−.24	0	134.36	.023	−.30	−.20	0
MBE-A - STR	107.75	.020	−.17	.02	0	120.00	.015	−.08	.06	0
MBE-A - CON	169.00	.023	−.51	−.29	0	130.75	.018	−.35	−.31	0
MBE-A - ETH	157.78	.020	−.41	−.17	0	126.12	.015	−.27	−.18	0
MBE-A - SERV	126.84	.013	−.42	−.20	0	97.49	.010	−.21	−.12	0
MBE-A - ABUS	51.18	.007	.56	.43	0	51.05	.006	.14	.40	0
MBE-A - EFCT	*96.98*	*.020*	*−.50*	*−.37*	0	100.92	.020	−.34	−.31	0
MBE-P - LF	**3.79**	**.001**	**1.00**	**.92**	4	4.18	.004	**.99**	**.97**	2
MBE-P - CHA	177.70	.014	−.57	−.50	0	184.08	.014	−.51	−.36	0
MBE-P - LMX	172.27	.031	−.64	−.55	0	194.79	.035	−.60	−.52	0
MBE-P - STR	200.61	.027	−.61	−.62	0	221.65	.029	−.61	−.56	0
MBE-P - CON	195.08	.027	−.67	−.64	0	178.21	.024	−.66	−.63	0
MBE-P - ETH	177.13	.022	−.63	−.58	0	186.66	.023	−.64	−.57	0
MBE-P - SERV	181.76	.019	−.61	−.56	0	188.85	.019	−.55	−.47	0
MBE-P - ABUS	67.57	.009	.53	.52	0	63.24	.008	.54	.43	0
MBE-P - EFCT	193.61	.042	−.68	−.64	0	164.23	.044	−.65	−.60	0
LF - CHA	201.68	.016	−.56	−.53	0	147.04	.011	−.40	−.33	0
LF - LMX	223.27	.041	−.64	−.65	0	201.21	.037	−.56	−.54	0

(continued)

Table 5. (continued)

	Corrected for Random, Specific Factor, and Transient Error (Time 1 and Time 2 Data Combined)					Corrected for Random and Specific Factor Error (Time 1 Data Only)				
	(1)	(2)	(3)	(4)	(5) Indices Suggesting Lack of Discriminant	(1)	(2)	(3)	(4)	(5) Indices Suggesting Lack of Discriminant
Construct Pair	$\Delta\chi^2$	ΔCFI	FC	ρ	Validity	$\Delta\chi^2$	ΔCFI	FC	ρ	Validity
LF - STR	216.93	.028	−.65	−.72	0	228.21	.030	−.57	−.52	0
LF - CON	219.57	.031	−.68	−.72	0	184.38	.025	−.54	−.59	0
LF - ETH	248.65	.032	−.69	−.68	0	213.57	.027	−.60	−.60	0
LF - SERV	194.46	.020	−.63	−.60	0	148.30	.016	−.46	−.43	0
LF - ABUS	42.71	.006	.58	.64	0	31.85	.004	.56	.54	0
LF - EFCT	179.41	.040	−.67	−.71	0	165.12	.035	−.57	−.60	0
CHA - LMX	43.04	.003	**.91**	**.90**	2	30.27	**.002**	.82	.80	1
CHA - STR	55.23	.003	.76	.78	0	37.01	**.002**	.72	.67	1
CHA - CON	43.92	.003	**.90**	.81	1	27.53	**.002**	**.85**	.78	2
CHA - ETH	20.06	**.001**	**.92**	.89	2	26.66	**.002**	**.86**	.84	2
CHA - SERV	12.92	**.000**	**.93**	**.93**	3	9.23	**.000**	**.86**	.84	2
CHA - ABUS	363.47	.021	−.59	−.58	0	212.20	.011	−.49	−.47	0
CHA - EFCT	4.48	**.000**	**.90**	.87	2	**.44**	**.000**	.84	.79	2
LMX - STR	34.39	.003	.78	.81	0	69.56	.007	.79	.74	0
LMX - CON	10.07	**.001**	**.93**	**.92**	3	28.19	**.002**	**.91**	**.90**	3
LMX - ETH	17.43	**.001**	**.93**	**.91**	3	20.17	**.001**	**.89**	.89	2
LMX - SERV	12.70	**.001**	**.90**	**.92**	3	8.63	**.001**	**.85**	.85	2
LMX - ABUS	417.56	.040	−.67	−.70	0	377.27	.036	−.55	−.57	0
LMX - EFCT	**.97**	**.000**	**.93**	**.96**	4	**.76**	**.000**	**.89**	**.90**	4
STR - CON	21.59	**.001**	.77	.76	1	57.42	.004	.82	.75	0
STR - ETH	34.57	**.002**	.81	.85	1	28.14	.003	.82	.80	0
STR - SERV	9.65	**.000**	.77	.81	1	21.52	**.002**	.74	.71	1
STR - ABUS	401.68	.032	−.55	−.61	0	387.01	.031	−.54	−.51	0
STR - EFCT	4.88	**.000**	.74	.78	1	**2.26**	**.000**	.75	.71	2
CON - ETH	32.48	.003	**.94**	**.90**	2	21.65	**.002**	**.96**	**.94**	3
CON - SERV	12.95	**.000**	**.95**	**.91**	3	**.41**	**.000**	**.91**	.89	3
CON - ABUS	220.48	.018	−.73	.76	0	412.19	.034	−.62	−.67	0
CON - EFCT	**1.98**	**.000**	**.92**	**.91**	4	4.26	**.000**	**.88**	**.90**	3
ETH - SERV	8.13	**.000**	**.94**	**.92**	3	15.41	**.001**	**.87**	.87	2
ETH - ABUS	397.77	.032	−.68	−.70	0	415.48	.031	−.61	−.64	0
ETH - EFCT	**2.03**	**.000**	**.92**	**.91**	4	**2.01**	**.000**	**.86**	.89	3
SERV - ABUS	270.84	.019	−.65	−.69	0	200.27	.013	−.49	−.52	0
SERV - EFCT	**1.34**	**.000**	**.92**	.88	3	**1.02**	**.000**	**.88**	.86	3
ABUS - EFCT	252.46	.026	−.71	−.74	0	509.92	.052	−.60	−.64	0
TOTAL	9	32	29	24		8	30	18	6	

Note: N = 185. (1) Chi-square difference between the unconstrained model and a model in which the covariance between the construct pair was constrained to 1.0; (2) CFI difference between the unconstrained model and a model in which the covariance between the construct pair was constrained to 1.0; (3) factor correlation between construct pairs; (4) disattenuated correlation between construct pairs; (5) number of discriminant validity indices for a given construct pair that suggest a lack of discriminant validity. TFL = transformational leadership; CR = contingent reward; MBE-A = management by exception-active; MBE-P = management by exception-passive; LF = laissez-faire; CHAR = charismatic leadership; LMX = leader-member exchange; STR = initiating structure; CONS = consideration; ETH = ethical leadership; SERV = servant leadership; ABUS = abusive supervision; EFCT = leader effectiveness. TOTAL = the number results from each column that suggest a lack of discriminant validity for a given construct pair. Statistics that suggest a lack of discriminant validity between constructs are shown in bold. Two models (MBE-A - EFCT for the CES corrections and CR - ABUS for the CE corrections) were under-identified. Results for the identified model are shown in italics.

Discussion

The results from our empirical example suggest that the extent to which leadership constructs can be considered empirically distinct may be much less than previously believed once the major sources of measurement error have been taken into account. How might we interpret such findings?

Covariance across scale items

A perusal of the individual scale items that constitute the measures included in our study suggests that some items include similar content even though they belong to leadership measures designed to assess different constructs. For example, an item contained in the MLQ-5X asks respondents to indicate the extent to which their leader "seeks differing perspectives when solving problems." This item was highly related to items from the ethical leadership (M. E. Brown et al., 2005), consideration (Stogdill, 1963), and servant leadership (Ehrhart, 2004) scales that ask respondents to indicate the extent to which their leader " ... listens to what others have to say," " ... puts suggestions made by the group into operation," and makes decisions that " ... are influenced by his/her followers' input," respectively. The mean observed correlation between these items in our sample was .55 (Time 1) and .61 (Time 2). Given the similarity in content across items from measures of theoretically distinct constructs, it should not be surprising that it is difficult to distinguish these constructs from one another empirically. At the very least, it may be the case that the working adults in our sample were unable to distinguish between the items to the extent that is necessary for distinct leadership constructs to emerge in our data.

Some existing leadership constructs are empirically identical

Transformational leadership and leader-member exchange are theoretically distinct constructs. In trying to explain the theoretical link between the constructs, scholars have, for example, tested a model in which LMX mediates the effects of TFL on job performance (Wang, Law, Hackett, Wang, & Chen, 2005). The results from our empirical study might suggest that follower perceptions of LMX do not mediate the effects of TFL but that follower perceptions of LMX and TFL are empirically identical. Furthermore, our results suggest that follower perceptions of TFL and LMX are empirically identical to several other leadership constructs. These results may have far-reaching implications for the leadership literature that go beyond the bounds of this article. In general, our results may suggest that rather than continue to treat these various leadership constructs as unique, leadership scholars may need to reconsider current leadership theories and taxonomies in favor of a greater level of parsimony. At the same time, we stress that our results are meant to be illustrative in nature and should not be used as the sole basis of any substantive conclusions regarding leadership theories. We leave it to researchers in the leadership domain to consider the full impact of our findings.

Additional Issues

Several topics related to assessing discriminant validity deserve additional attention, and we address them here. In our empirical study, we measured each leadership construct by administering the same measure on two separate occasions. The main disadvantage of this approach is that participants' responses on the first administration of the leadership measures could have affected their responses on the second administration (Stanley, 1971). This problem is most relevant to cognitive (i.e., intelligence) and psychomotor constructs—the measurement of which can be influenced by memory or practice effects (Thorndike, 1951). Because responses to the leadership measures used in our study are unlikely to be influenced by memory effects and because parallel forms were not available for

some of the measures used in our study, we believe that the split-half approach was appropriate for this study. However, future research might focus on developing parallel measures of leadership constructs and reassessing the findings presented in our study. In addition, a reviewer noted that critics of the methods reviewed here could argue that true score correlations are different from construct-level correlations. These critics might raise the possibility that true score correlations are meaningfully higher than construct-level correlations. If this were the case, then the methods that we outline in this article (which actually estimate true score correlations) would produce overestimates of the relationships between constructs and could lead to the conclusion that constructs are redundant when they are actually unique. Recent research has examined this issue and found that true scores and construct scores were correlated at .98, suggesting that the methods presented in this article will produce accurate estimations of construct-level relationships (Schmidt, Le, & Oh, 2013).

Finally, one interpretation of our results that we did not empirically examine is the possibility that the leadership constructs we included in our study are indicators of one or more higher-order factors as conceptualized by previous scholars (Bass, 1990; Fleishman, 1953; Halpin & Winer, 1957; Hinkin & Schriesheim, 2008; Yukl et al., 2002). We believe a higher-order analysis of our data to be problematic for two reasons. First, our analyses would be conducted post hoc, driven by empiricism instead of theory. This is certainly not ideal given that the purpose of our article is to guide discriminant validity analyses and not to endorse any one leadership taxonomy over the other. Second, the construct-level correlations between many of the constructs in our study are quite high, and even if our data were to suggest the presence of one or more higher-order factors, this would not change the fact that the indicators are correlated so highly as to be empirically redundant. Said another way, even if we attempted to build a theoretical argument that depicted these constructs as indicators of different higher-order factors, the lack of empirical distinctions between them would suggest that they are redundant constructs regardless of their proposed place in a higher-order model. Thus, the conditions necessary for establishing a new construct—conceptual distinction and empirical distinction—also apply to hierarchical factor models.

General Recommendations

Given that the conclusions drawn from our empirical analysis might differ depending on the analytic strategy we used and the summary statistics we presented, we make several research recommendations. The first recommendation we make is that researchers should correct for all three major sources of measurement error when possible. This recommendation is particularly relevant when the magnitude of factor correlations or disattenuated correlations is used to determine the discriminant validity of constructs. This means that researchers should design studies of discriminant validity such that data are collected in a way that accounts for transient error. Failure to do so will likely result in underestimates of the empirical relationships between a given set of constructs and could lead to the conclusion that the constructs are empirically unique when they are actually redundant. Of course, researchers who follow this recommendation will find that the process of collecting study data is more onerous. As Crutzen (2014) writes, ''This leaves us at a crossroad. We more or less ignore transient error and simply go on or we agree that test-retest analyses should be a part of comprehensive assessment of scale quality. In case of the latter, we have to acknowledge that this brings additional workload'' (p. 73). Keeping in mind that the development of theory rests on accurate estimates of the empirical relationships between constructs, we believe that taking on this workload could be beneficial to organizational research.

We also recommend that in addition to the commonly reported results of χ^2 difference tests, researchers include ΔCFI, factor correlations, and disattenuated correlations when reporting the results of a discriminant validity analysis. There are several benefits to reporting this complement of indices. First, for researchers who wish to make empirical comparisons between measurement

models but are concerned about the limitations associated with the $\Delta\chi^2$ test, the ΔCFI assessment allows for a comparison of models that overcomes these limitations to some extent (Meade et al., 2008). However, neither $\Delta\chi^2$ nor ΔCFI provide information about the size of the relationship between constructs. Thus, in keeping with previous calls (Schmidt, 2010) and the current standards of the American Psychological Association (2001), we recommend that researchers report the effect sizes needed to allow readers to assess the strength of the relationships between individual study constructs. Second, researchers should provide estimates of CES whenever possible. Along these lines, we wish to emphasize that because theory development sometimes rests on the accumulation of data over time, one positive externality that may be realized through the reporting of effect sizes is that such data—especially observed correlations and estimates of CES—can be used as ready input for meta-analytic studies. A third benefit of reporting multiple indices of discriminant validity is that it may encourage researchers to avoid relying on any single statistic when interpreting a discriminant validity analysis and instead to take a more holistic view of their data. If multiple indices suggest the same conclusion, a researcher might be relatively confident about his or her determination that constructs are unique or not. Reporting multiple indices will also allow readers to more easily evaluate studies of discriminant validity and draw their own conclusions from the results.[4]

In our empirical example, it was not uncommon for a given indicator of discriminant validity to suggest a different research conclusion than did the other three. When three of our four indices suggested a lack of discriminant validity between a given construct pair, it was most common for $\Delta\chi^2$ to suggest that the two constructs were distinct when the ΔCFI, the factor correlation, and the disattenuated correlation all suggested the constructs were empirically redundant. Alternately, when only one of the four indices suggested construct redundancy, it was most often ΔCFI. When two of the four indices suggested a lack of discriminant validity, there was a lack of consistency in which two indices were in agreement. Given what might seem to be conflicting results across the four discriminant validity indices, our final recommendation is that even if only one index ($\Delta\chi^2$, ΔCFI, the factor correlation, or the disattenuated correlation) exceeds the associated threshold specified in our study, researchers should question the discriminant validity of the study constructs and should consider conducting additional research to further assess the uniqueness of the constructs. We realize that this approach is a relatively cautious and conservative one, but we believe that such caution is warranted to deter unnecessary complication of organizational theories.

Our recommendations to base research conclusions on multiple discriminant validity indices raise an important question about the legitimacy of the specific cutoff values we used in our study and, more broadly, the cutoff values that researchers can use to guide interpretations of their study results. Indeed, some readers might argue that the χ^2 difference test is superior to the other indices we present because it is the only index that has a definitive cutoff value ($p < .05$) from which study conclusions can be drawn. Such an argument would maintain that even though we based the cutoff values for our alternative indicators of discriminant validity on recommendations from previous research, the values were chosen arbitrarily and lack the objective characteristics of a χ^2 difference test with a cutoff value of $p < .05$. We admit that our results would change dramatically if we had chosen more conservative cutoff values (e.g., a ΔCFI of .001 or less, a factor correlation of magnitude .90 or greater, and a disattenuated correlation of magnitude .95 or greater) or less conservative values (e.g., a ΔCFI of .003 or less, a factor correlation of magnitude .80 or greater, and a disattenuated correlation of magnitude .85 or greater). We also expect there to be at least some disagreement among even the most informed researchers as to what the cutoff values should be for the indices we present. After all, ''reasonable people can disagree about any one cutoff point because it is inherently arbitrary'' (John & Benet-Martínez, 1990, p. 361). But the exact cutoff values are not the most important detail here. Rather, we wish to point out that the χ^2 difference test has the appearance of providing an objective and precise decision point to guide research conclusions, but the typical reliance on $p < .05$ as a cutoff value for interpreting research findings is just as arbitrary

as any other value we present in this article. Although $p < .05$ is the accepted convention in psychological research, Rozeboom (1960) states quite pointedly,

> There is absolutely no reason (at least provided by the method) why the point of statistical "significance" should be set at the 95% level, rather than, say the 94% or 96% level. Nor does the fact that we sometimes select a 99% level of significance, rather than the usual 95% level, mitigate this objection—one is as arbitrary as the other. (p. 423)

Rosnow and Rosenthal (1989) express the same argument a bit more playfully: "Surely, God loves the .06 nearly as much as the .05" (p. 1277). Because no one cutoff value is likely to be more "valid" than any other value, we reiterate our recommendation that scholars should assess discriminant validity using a comprehensive set of statistics that allows for holistic, broadly informed interpretations of their research findings.

Conclusion

Over 80 years ago, Kelley (1927) recognized the problem of construct proliferation, calling it the "jangle fallacy" and describing it as "contaminating to clear thinking" (p. 64). In this article, we describe growing concerns related to construct proliferation in organizational science; outline the impact that random, specific factor, and transient error have on assessments of discriminant validity; and provide a procedural guide for researchers who seek to establish the empirical uniqueness of organizational constructs. Although discriminant validity analyses have long been discussed in the organizational sciences, we believe that scholars may underestimate the extent to which the results of a discriminant validity analysis can differ across analytic methods and the extent to which measurement error can bias research conclusions—especially conclusions related to the empirical distinctness of conceptually related constructs. Our hope is that the use of rigorous construct validation methods will increase the ability of future scholars to develop parsimonious theories that can advance scientific knowledge in the field of organizational research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. We conducted confirmatory factor analysis (CFA) based on item-level data. Across both test administrations, we found a total of three missing items—each from a different participant. We replaced the missing responses with the mean score for each corresponding item.
2. The factor correlation between management by exception-passive and laissez-faire in our study was 1.01 once random, specific factor, and transient error had been taken into account. We reported a correlation of 1.0 for that value. The most probable explanation for this result is sampling error. To be clear, this explanation proposes that the corrected correlation in our study that exceeds 1.0 is "within the sampling distribution of corrected correlations produced by a population with a true-score correlation less than or equal to one" (Charles, 2005, p. 209).
3. Le, Schmidt, and Putka (2009) show that when parallel full scales are available for analysis, the factor correlations derived from a CFA analysis and the coefficient of equivalence and stability (CES)–corrected correlations derived from the disattenuation formula should be nearly identical. However, since parallel scales

were not available for some of the constructs examined in our study, we used the split-half method. When the half-scale method applied in our article is used for computing disattenuated correlations, the factor correlation and the disattenuated correlation between a given construct pair will not necessarily be the same. This is because the CFA-based approach that relies on split-half scales uses only half of the items from both administrations of a scale to estimate the factor correlation whereas the CES-based approach uses all of the items from both administrations of a scale to compute the observed correlations used in the numerator (as shown in Equation 5) and the reliability estimates used in the denominator (as shown in Equation 7) portions of the disattenuation formula. As shown in our results section, the two methods tend to yield similar results and to suggest similar conclusions regarding the distinctiveness of two constructs.

4. A reviewer suggested that another CFA-based index of discriminant validity is the comparison between the average variance extracted (AVE), which denotes the amount of variance captured by a focal construct relative to measurement error, with the square of the factor correlation between two constructs (Fornell & Larcker, 1981). This method proposes that discriminant validity is supported when the AVEs for each of the constructs is greater than the squared factor correlation between the two constructs. Due to space constraints, we do not include these analyses in our study. Future research should consider including this method to assess construct discriminant validity.

References

Althauser, R. P., & Heberlein, T. A. (1970). Validity and the multitrait-multimethod matrix. In E.F. Borgatta & G.W. Bohrnstedt (Eds.), *Sociological methodology* (pp. 151-169). San Francisco: Jossey-Bass.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*, 256-274.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Avolio, B. J., Bass, B. M., & Jung, D. I. (1999). Re-examining the components of transformational and transactional leadership using the Mulitfactor Leadership Questionnaire. *Journal of Occupational and Organizational Psychology*, *72*, 441-462.

Avolio, B. J., & Gardner, W. L. (2005). Authentic leadership development: Getting to the root of positive forms of leadership. *The Leadership Quarterly*, *16*, 315-338.

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly*, *36*, 421-458.

Barrick, M. R., Swider, B. W., & Stewart, G. L. (2010). Initial evaluations in the interview: Relationships with subsequent interviewer evaluations and employment offers. *Journal of Applied Psychology*, *95*, 1163-1172.

Bass, B. M. (1990). From transactional to transformational leadership: Learning to share the vision. *Organizational Dynamics*, *18*, 19-31.

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

Brown, M. E., Treviño, L. K., & Harrison, D. A. (2005). Ethical leadership: A social learning perspective for construct development and testing. *Organizational Behavior and Human Decision Processes*, *97*, 117-134.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Campbell, J. P. (1990). The role of theory in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, pp. 39-73). Palo Alto, CA: Consulting Psychologists Press.

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

Carless, S. A. (1998). Assessing the discriminant validity of transformational leadership behavior as measured by the MLQ. *Journal of Occupational and Organizational Psychology*, *71*, 353-358.

Carless, S. A., Wearing, A. J., & Mann, L. (2000). A short measure of transformational leadership. *Journal of Business and Psychology*, *14*, 389-405.

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods*, *10*, 206-226.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Cole, M. S., Walter, F., Bedeian, A. G., & O'Boyle, E. H. (2012). Job burnout and employee engagement: A meta-analytic examination of construct proliferation. *Journal of Management*, *38*, 1550-1581.

Conger, J. A., & Kanungo, R. N. (1994). Charismatic leadership in organizations: Perceived behavioral attributes and their measurement. *Journal of Organizational Behavior*, *15*, 439-452.

Conger, J. A., Kanungo, R. N., & Menon, S. T. (2000). Charismatic leadership and follower effects. *Journal of Organizational Behavior*, *21*, 747-767.

Conger, J. A., Kanungo, R. N., Menon, S. T., & Mathur, P. (1997). Measuring charisma: Dimensionality and validity of the Conger-Kanungo Scale of Charismatic Leadership. *Canadian Journal of Administrative Sciences*, *14*, 290-301.

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

Cortina, J. M., & DeShon, R. P. (1998). Determining relative importance of predictors with the observational design. *Journal of Applied Psychology*, *83*, 798-804.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Crutzen, R. (2014). Time is a jailer: What do alpha and its alternatives tell us about reliability? *The European Health Psychologist*, *16*, 70-74.

DeRue, D. S., Nahrgang, J. D., Wellman, N., & Humphrey, S. E. (2011). Trait and behavioral theories of leadership: An intergration and meta-analytic test of their relative validity. *Personnel Psychology*, *64*, 7-52.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Ehrhart, M. G. (2004). Leadership and procedural justice climate as antecedents of unit-level organizational citizenship behavior. *Personnel Psychology*, *57*, 61-94.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Fleishman, E. A. (1953). The description of supervisory behavior. *Journal of Applied Psychology*, *37*, 1-6.

Fleishman, E. A., Mumford, M. D., Zaccaro, S. J., Levin, K. Y., Korotkin, A. L., & Hein, M. B. (1991). Taxonomic efforts in the description of leadership behavior: A synthesis and functional interpretation. *Leadership Quarterly*, *2*, 245-287.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*, 39-50.

Gilliam, D. A., & Voss, K. (2013). A proposed procedure for construct definition in marketing. *European Journal of Marketing*, *47*, 5-26.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*, 79-132.

Graen, G. B., & Uhl-Bien, M. (1995). Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain perspective. *Leadership Quarterly*, *6*, 219-247.

Grayson, L., & Gomersall, A. (2003). *A difficult business: Finding the evidence for social science reviews* (Working Paper 19). Swindon, UK: ESRC UK Centre for Evidence Based Policy and Practice.

Greenberg, J., Solomon, S., Pyszczynski, T., & Steinberg, L. (1988). A reaction to Greenwald, Pratkanis, Leippe, and Baumgardner (1986): Under what conditions does research obstruct theory progress? *Psychological Review*, *95*, 566-571.

Halpin, A. W., & Winer, B. J. (1957). A factorial study of the leader behavior descriptions. In R. M. Stogdill & A. E. Coons (Eds.), *Leader behavior: Its description and measurement* (pp. 39-51). Columbus, OH: Ohio State University, Bureau of Business Research.

Hammerstrøm, K., Wade, A., & Jørgensen, A. M. K. (2010). Searching for studies: A guide to information retrieval for Campbell Systematic Reviews. Campbell Systematic Reviews 2010: Supplement 1.

Harter, J. K., & Schmidt, F. L. (2008). Conceptual versus empirical distinctions among constructs: Implications for discriminant validity. *Industrial and Organizational Psychology*, *1*, 36-39.

Harter, J. K., Schmidt, F. L., Asplund, J. W., Killham, E. A., & Agrawal, S. (2010). Causal impact of employee work perceptions on the bottom line of organizations. *Perspectives on Psychological Science*, *5*, 378-389.

Heinitz, K., Liepmann, D., & Felfe, J. (2005). Examining the factor structure of the MLQ: Recommendation for a reduced set of factors. *European Journal of Psychological Assessment*, *21*, 182-190.

Hershcovis, M. S. (2011). "Incivility, social undermining, bullying . . . oh my!": A call to reconcile constructs within workplace aggression research. *Journal of Organizational Behavior*, *32*, 499-519.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*, 104-121.

Hinkin, T. R., & Schriesheim, C. A. (2008). A theoretical and empirical examination of the transactional and non-leadership dimensions of the Multifactor Leadership Questionnaire (MLQ). *Leadership Quarterly*, *19*, 501-513.

Hoyle, R. H. (2000). Confirmatory factor analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 465-497). San Diego, CA: Academic Press.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage Publications.

John, O. P., & Benet-Martínez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 339-369). New York, NY: Cambridge University Press.

Judge, T. A., & Bono, J. E. (2000). Five-factor model of personality and transformational leadership. *Journal of Applied Psychology*, *85*, 751-765.

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*, 621-652.

Kneale, W., & Kneale, M. (1962). *The development of logic*. Oxford, UK: Oxford University Press.

Kelley, T. L. (1927). *Interpretation of educational measurements*. Yonkers, NY: World Book Company

Kenny, D. A. (2012). *Multiple latent variable models: Confirmatory factor analysis*. Retrieved from http://davidakenny.net/cm/mfactor.htm

Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods*, *12*, 165-200.

Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes*, *112*, 112-125.

Lenth, R. V. (2006-2014). Java applets for power and sample size [Computer software]. Retrieved from http://homepage.stat.uiowa.edu/∼rlenth/Power/

Liden, R. C., Wayne, S. J., Zhao, H., & Henderson, D. (2008). Servant leadership: Development of a multidimensional measure and multi-level assessment. *The Leadership Quarterly*, *19*, 161-177.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*, 19-35.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130-149.

Marsh, H. W., & Hocevar, D. (1983). Confirmatory factor analysis of multitrait-multimethod matrices. *Journal of Educational Measurement*, *20*, 231-248.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568-592.

Mehdyzadeh, H. (2004). *Searching for the evidence: An introduction to social science information retrieval*. London: Department for Culture, Media and Sport.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301.

Page, D., & Wong, P. T. P. (2000). A conceptual framework for measuring servant leadership. In S. Adjibolosoo (Ed.), *The human factor in shaping the course of history and development* (pp. 69-110). Boston, MA: University Press of America.

Pfeffer, J. (1977). The ambiguity of leadership. *Academy of Management Review*, *2*, 104-112.

Pfeffer, J. (1993). Barriers to the advance of organizational science: Paradigm development as a dependent variable. *Academy of Management Review*, *18*, 599-620.

Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management*, *36*, 94-120.

Reeve, C. L., & Bonaccio, S. (2011). On the myth and the reality of the temporal validity degradation of general mental ability test scores. *Intelligence*, *39*, 255-272.

Reeve, C. L., Heggestad, E. D., & George, E. (2005). Estimation of transient error in cognitive ability scales. *International Journal of Selection and Assessment*, *13*, 316-320.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276-1284.

Rousseau, D. M. (2007). A sticky, leveraging, and scalable strategy for high-quality connections between organizational practice and science. *Academy of Management Journal*, *50*, 1037-1042.

Rowold, J., & Heinitz, K. (2007). Transformational and charismatic leadership: Assessing the convergent, divergent, and criterion validity of the MLQ and the CKS. *Leadership Quarterly*, *18*, 121-133.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416-428.

Schmidt, F. L. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science*, *5*, 233-242.

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 differences constructs. *Psychological Methods*, *8*, 206-224.

Schmidt, F. L., Le, H., & Oh, I. (2013). Are true scores and construct scores the same? A critical examination of their substitutability and the implications for research results. *International Journal of Selection and Assessment*, *21*, 339-354.

Schucan Bird, K., & Tripney, J. (2011). Systematic literature searching in policy relevant, inter-disciplinary reviews: an example from culture and sport. *Research Synthesis Methods*, *2*, 163-173.

Schwab, D. P. (1980). Construct validity in organizational behavior. In B. M. Staw & L. L. Cummings (Eds.). *Research in organizational behavior* (Vol. 2, pp. 3-43). Greenwich, CT: JAI Press.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA): An illustration using entrepreneurial orientation. *Organizational Research Methods*, *13*, 320-347.

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

Staw, B. M., Bell, N. E., & Clausen, J. A. (1986). The dispositional approach to job attitudes: A lifetime longitudinal test. *Administrative Science Quarterly*, *31*, 56-77.

Stogdill, R. M. (1963). *Manual for the Leader Behavior Description Questionnaire, Form XII*. Columbus, OH: Bureau of Business Research, Ohio State University.

Swales, J. (1986). Citation analysis and discourse analysis. *Applied Linguistics*, *7*, 39-56.

Taylor, B., Wylie, E., Dempster, M., & Donnelly, M. (2007). Systematically retrieving research: A case study evaluating seven databases. *Research on Social Work Practice*, *17*, 697-706.

Tejeda, M. J., Scandura, T. A., & Pillai, R. (2001). The MLQ revisited: Psychometric properties and recommendations. *The Leadership Quarterly*, *12*, 31-52.

Tepper, B. J. (2000). Consequences of abusive supervision. *Academy of Management Journal*, *43*, 178-190.

Tepper, B. J., & Henle, C. A. (2011). A case for recognizing distinctions among constructs that capture interpersonal mistreatment in work organizations. *Journal of Organizational Behavior*, *32*, 487-498.

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

Tsui, A. S. (1984). A role set analysis of managerial reputation. *Organizational Behavior and Human Performance*, *34*, 64-96.

van Mierlo, H., Vermunt, J. K., & Rutte, C. G. (2009). Composing group-level constructs from individual-level survey data. *Organizational Research Methods*, *12*, 368-392.

Wang, H., Law, K. S., Hackett, R. D., Wang, D., & Chen, Z. X. (2005). Leader-member exchange as a mediator of the relationship between transformational leadership and followers' performance and organizational citizenship behavior. *Academy of Management Journal*, *48*, 420-432.

Watson, D., & Humrichouse, J. (2006). Personality development in emerging adulthood: Integrating evidence from self-ratings and spouse ratings. *Journal of Personality and Social Psychology*, *91*, 959-974.

Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review*, *14*, 490-495.

White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 41-55). New York, NY: Russell Sage Foundation.

Yukl, G., Gordon, A., & Taber, T. (2002). A hierarchical taxonomy of leadership behavior: Integrating a half century of behavior research. *Journal of Leadership and Organizational Studies*, *9*, 15-32.

Author Biographies

Jonathan A. Shaffer is the Pickens Professor of Management at West Texas A&M University. He received his PhD in organizational behavior from the University of Iowa. His research has been published in journals such as *Journal of Applied Psychology, Personnel Psychology*, and *Journal of Occupational Health Psychology*.

David DeGeest is an Assistant Professor of HRM/OB at the University of Groningen in the Netherlands. His work has appeared in such journals as *Journal of Applied Psychology, Journal of Management*, and *Journal of Organizational Behavior*. His research interests include collaboration in teams, entrepreneurship, and the role of time in research.

Andrew Li is the Williams Professor of Management at West Texas A&M University. He earned his doctoral degree in management from the University of Arizona. His research focuses on work-family interface in organizations, personality, organizational justice, and team research.