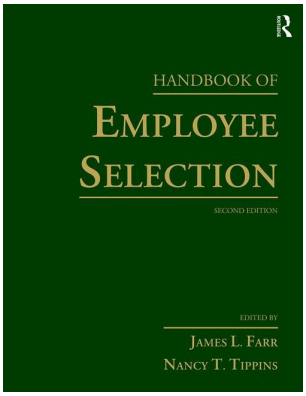

Handbook of Employee Selection

James L. Farr, Nancy T. Tippins, Walter C. Borman, David Chan, Michael D. Coovert, Rick Jacobs, P. Richard Jeanneret, Jerard F. Kehoe, Filip Lievens, S. Morton McPhail, Kevin R. Murphy, Robert E. Ployhart, Elaine D. Pulakos, Douglas H. Reynolds, Ann Marie Ryan, Neal Schmitt, Benjamin Schneider

Situational Specificity, Validity Generalization, and the Future of Psychometric Meta-analysis

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315690193-4
James M. LeBreton, Jeremy L. Schoen, Lawrence R. James
Published online on: 22 Mar 2017

4

Situational Specificity, Validity Generalization, and the Future of Psychometric Meta-Analysis

James M. LeBreton, Jeremy L. Schoen, and Lawrence R. James[1]

Most psychologists would agree that a well-designed employment test should yield evidence of criterion-related validity when tested against a well-measured criterion. If "validity generalization" (VG) were limited to this inference, then there would be no reason for this chapter. Indeed, the authors of this chapter subscribe to this inference, but VG is not limited to this inference. Instead, VG inferences are often extended to suggest that the magnitude of test validities are invariant across situations—that is, situations do not influence the magnitude of criterion-related validity coefficients. This line of thinking is aptly captured in quotes such as the following:

> The evidence from these two studies appears to be the last nail required for the coffin of the situational specificity hypothesis.
>
> (Schmidt, Hunter, Pearlman, & Rothstein-Hirsch, 1985, p. 758)

> The cumulative pattern of findings . . . provides strong support for the hypothesis that there is essentially no situational variance in true validities for classic ability constructs used for selection on similar jobs.
>
> (Schmidt et al., 1993, p. 11)

> these studies found that, on average, all variance across settings (i.e., companies) was accounted for by artifacts. . . . All these pieces of interlocking evidence point in the same direction: toward the conclusion that, for employment tests of cognitive abilities, the situational specificity hypothesis is false.
>
> (Hunter & Schmidt, 2004, pp. 404–405)

> Beginning in 1977, Schmidt and Hunter began publishing empirical evidence discrediting the situational specificity hypothesis. Specifically, they demonstrated that much of the variability in validity coefficients across studies was due to random sampling error.
>
> (McDaniel, Kepes, & Banks, 2011, p. 497)

There is little question that (psychometrically well-developed) tests of knowledge, skills, abilities (i.e., KSAs), and personality traits generally predict (psychometrically well-developed) measures of organizationally relevant criteria. In this sense, the criterion-related validity evidence for these tests can be said to generalize. Whether the validity for a given type of predictor (e.g., critical intellectual skills) against a given class of criterion (e.g., job performance) is generally invariant across situations (i.e., cross-situationally consistent) is another issue. The cross-situational

consistency hypothesis (i.e., VG) has endured a long history of theoretical and empirical debate, the roots of which can be traced, in part, to the person-situation debate (cf. Buss, 1979; Cronbach & Snow, 1977; Epstein, 1979; Hogan, 2009; Kendrick & Funder, 1988; Mischel, & Peake, 1982). The emergence of meta-analysis as a popular method for testing the consistency of predictive validities across a set of separate studies (e.g., situations) accelerated and transformed the debate into one of a more quantitative and methodological nature.

Basically, meta-analysis made it possible for organizational researchers to apply increasingly sophisticated quantitative tools to assess the predictive validity of test scores and, more importantly, the consistency of these estimates over studies. Within applied psychology, the most commonly used variant of meta-analysis has been the VG analysis, which more recently has adopted the label of psychometric meta-analysis (PMA; Borenstein, Hedges, Higgins, & Rothstein, 2009; Hunter & Schmidt, 2004). In the typical VG analysis, investigators first provide estimates of the criterion-related validity coefficients obtained (for the same, or similar, predictor-criterion variable pairs) from different study samples. Investigators then examine the cross-situational variability in those criterion-related validity coefficients (cf. Pearlman, Schmidt, & Hunter, 1980; Salgado et al., 2003; Schmidt & Hunter, 1977; Schmidt, Hunter, & Raju, 1988; Schmidt et al. 1993).

Unlike the meta-analytic techniques embraced in nearly all other areas of science, the VG technique seeks to estimate variability in validity coefficients after first adjusting (or, "correcting") the observed coefficients for statistical artifacts (e.g., measurement error, range restriction). These corrections are believed to remove irrelevant noise from the system, thus enhancing the comparability of these estimates across different situations (Schmidt & Hunter, 1977; Schmidt et al., 1993). However, VG procedures are not without their critics (Algera, Jansen, Roe, & Vijn, 1984; James, Demaree, & Mulaik, 1986; James, Demaree, Mulaik, & Ladd, 1992; Kemery, Mossholder, & Roth, 1987), many of whom have questioned whether the findings based on VG procedures may have yielded an inaccurate picture of both the *magnitude* and *consistency* of predictor-criterion pairs. Like these critics, we also have concerns with the conclusions reached using VG procedures, and it is in that spirit with which this chapter was written.

We have two basic goals for this chapter. First, we discuss the logic and rationale underlying the VG and Situational Specificity (SS) hypotheses and, based on the results from the extant literature, conclude that the SS hypothesis is alive and well in applied psychology. Second, we summarize five key concerns related to VG studies and the PMA procedures upon which they are based. These concerns include (1) the formulas that are used in VG analyses fail to explicitly (i.e., empirically) incorporate measured situational variables (e.g., authority structure, interpersonal interactions, social climate), despite evidence that such variables often moderate the types of predictor-criterion relationships cited in the VG literature (i.e., Ghiselli, 1959, 1966, 1973; Peters, Fisher, & O'Connor, 1982); (2) the formulas that are the basis of PMA, and thus the basis for all VG analyses, include critical (untested) assumptions, the tenability of which has been called into question (James et al., 1992; Köhler, Cortina, Kurtessis, & Gölz, 2015); (3) many VG studies have relied on dubious estimates of statistical artifacts (e.g., estimates of criterion reliability) when estimating corrected validity coefficients, and these estimates may have resulted in biased inferences about both mean validities and the variance (or lack thereof) around those means (cf. DeShon, 2003; LeBreton, Burgess, Kaiser, Atchley, & James, 2003; LeBreton, Scherer, & James, 2014; Murphy & DeShon, 2000; Putka & Hoffman, 2015; Viswesvaran, Ones, & Schmidt, 1996, 2005); (4) the appropriateness of inferences based on corrected (or partially corrected) correlation coefficients (LeBreton, Scherer, & James, 2014); and finally (5) reliance on meta-analytically derived effect sizes to guide selection decisions, especially given the negative evaluations of VG by U.S. courts (Biddle, 2010; Landy, 2003).

Thus, our chapter is structured as follows. First, we provide a brief introduction to the logic and rationale underlying VG. Second, we summarize the evidence suggesting that SS is alive and well in applied psychology. Third, we offer a review and critique of the procedures of PMA that form the basis for VG analyses. Finally, we conclude with general recommendations relevant for employment selection research and practice.

VALIDITY GENERALIZATION VERSUS SITUATIONAL SPECIFICITY

Validity Generalization

Validity studies conducted in the mid- to late-20th century offered modest hope for the utility of personality and KSAs as predictors of crucial outcome variables (i.e., job performance) in applied settings. Of particular interest were validity coefficients for cognitive ability tests, which tended to be modest in magnitude and often inconsistent across job types (Ghiselli, 1959, 1966, 1973). As a result of this inconsistency, many psychologists adhered to the basic hypothesis that the criterion-related validity evidence for any given selection test was situationally specific (Murphy, 2000; Schmidt & Hunter, 1998). Stated alternatively, in order to determine the extent to which inferences drawn from test scores were related to outcomes (e.g., job performance; Binning & Barrett, 1989), psychologists must understand the subtle differences or constraints that differed across situations (e.g., specific/unique job requirements identified as part of a job analysis, differential reward structures that might influence performance, culture or climate of the organization, work characteristics, etc.; James et al., 1986, 1992; Murphy, 2000). In addition, the belief that criterion-related validities were situationally specific was consistent with the more general movement toward situational specificity of behavior (including Person by Situation interaction and contingency models of behavior; cf. Endler & Magnusson, 1976; Grote & James, 1991; House & Mitchell, 1974; Kerr, Schriescheim, Murphy, & Stogdill, 1974; Mischel, 1968; Vecchio, 1987; Vroom, 1973; Wright & Mischel, 1987).

VG developed out of a desire to try to increase the precision (i.e., accuracy) of validity coefficient estimates for similar or identical predictor-criterion pairs. Like other forms of meta-analysis, VG is based on a sample-size weighted average effect size. Unlike other forms of meta-analysis, VG moves beyond a simple summary/description of effect sizes to draw inferences about the consistency (or lack thereof) in the observed effect sizes. More specifically, whereas a traditional meta-analysis describes/summarizes the overall relationship between a predictor and criterion for a set of samples, VG goes one step further to infer the degree to which additional factors contribute to the consistency of this relationship across samples. Typically, VG analyses are undertaken separately for different job types or job classes (i.e., clerical, mechanical, managerial; Schmidt & Hunter, 1977). A variety of factors may contribute to the inconsistency of criterion-related validity across samples. The factors to be considered are statistical artifacts, such as unreliability of predictor and criterion scores, range restriction in predictor scores, and sampling error (i.e., Hunter & Schmidt, 2004; Schmidt & Hunter, 1977; Schmidt et al., 1988; Schmidt et al., 1993). Thus, a VG analysis may be thought of as the *inferential* variant of the traditional, *descriptive* meta-analysis (Murphy, 2000).

The primary (but not the only) assumptions underlying a VG analysis include that (a) the true validity for a particular predictor-criterion pair is equal across populations but that (b) statistical artifacts that differ across studies (e.g., predictor and/or criterion reliability, range restriction, and sampling error) distort and restrict the magnitude of the observed validity. In an attempt to identify and effectively model the impact of these biasing statistical artifacts, the following structural equation—in which Greek symbols represent population parameters—is generally used in VG analysis:

$$r_k = \rho_k \alpha_k^{1/2} \varphi_k^{1/2} \xi_k + e_k ,$$

where ρ_k is the population correlation between the unrestricted true scores for the predictor and criterion in situation k (i.e., the true validity); r_k represents the observed validity coefficient—i.e., the correlation between a predictor X_k and a criterion Y_k for a random sample of n_k individuals from population (i.e., organization, situation) k; α_k is the unrestricted population reliability for the criterion in situation k; φ_k is the unrestricted population reliability for the predictor in situation k; ξ_k reflects the degree of range restriction in the predictor in situation k; and e_k is the sampling error inherent in r_k.

Once the statistical artifact population estimates are inserted into the equation and ρ_k is estimated for each k, the next step is to estimate the variance among the ρ_k, referred to as $V(\rho)$, and

determine whether or not this estimated variance coefficient is small enough to justify generalization of the validity across situations. The estimate for $V(\rho)$ is calculated based on the following estimation equation (see James et al., 1992):

$$\hat{V}(\rho) = \left[V(r) - V(\dot{r})\right] / \Pi$$

where $\hat{V}(\rho)$ is the estimate of variance in population (true) validities; $V(r)$ is the between-situation variance in the observed validities; $V(\dot{r})$ is the expected between-situation variance in validities associated with statistical artifacts; and Π is an additional correction for mean reliabilities and range restriction across situations. In essence, the amount of variance attributable to statistical artifacts is subtracted from the total observed variance, and the remaining variance, termed "residual variance," represents the true variance in validities that is unaccounted for (i.e., by statistical artifacts).

A primary step of VG is to determine whether or not cross-situational consistency in validities has been achieved. Basically, if the estimate of $V(\rho)$ is approximately equal to 0, then the ρ_k are deemed to be truly invariant across situations (i.e., generalizable), whereas an estimate of $V(\rho)$ greater than 0 is used as evidence consistent with a potential situational moderator. To this end, two rules have emerged that elaborate on the term "approximately equal" by imposing predetermined, theoretically justified critical values, and $V(\rho)$ must not extend above these values in order for cross-situational consistency to be established.

One rule is the "75% Rule," in which 75% of the total variance in validity estimates (i.e., $V(\rho)$) must be accounted for by statistical artifacts to effectively rule out the SS hypothesis, suggesting a construct is a universal and invariant predictor of the criterion of interest. The remaining variance in validity estimates (i.e., 25% of the variance in validity estimates) is attributed to additional (unmeasured) artifacts (i.e., clerical and programming errors; Hermelin & Robertson, 2001; Schmidt & Hunter, 1977). The importance of the 75% rule for informing decisions about VG versus SS was noted by Hunter and Schmidt (2004):

> If 75% or more of the variance is due to artifacts, we conclude that all of it is, on the grounds that the remaining 25% is likely to be due to artifacts for which no correction has been made.
>
> (p. 401)

This rule has been criticized for being insensitive to potential situational moderators (James et al., 1986) and, while considered outdated, it is still used alone or in combination with more advanced techniques (Geyskens, Krishnan, Steenkamp, & Cunha, 2009).

Given concerns over the 75% Rule, a second rule based on formal statistical tests of the heterogeneity of $V(\rho)$ has emerged. This rule emphasizes the development of a "credibility interval," in which the lower bound of the validity distribution is compared to a minimal validity coefficient value (e.g., .00, .01, .10). If the credibility interval does not contain the minimal value, one can say with a certain amount of confidence that the validity of the scores will generalize to other populations. Researchers frequently use 80% and 90% credibility intervals to draw inferences regarding the transportability of a validity coefficient to other situations (the concept of transportability is discussed later).

Evidence for Validity Generalization and Situational Specificity: A Continuum Perspective

Historically, VG and SS were framed as two mutually exclusive outcomes. That is to say, the criterion-related validity evidence of a particular selection test was said to *either* generalize *or* be situationally specific. We believe a more fruitful path forward is to recognize that VG and SS may be better conceptualized as forming the anchors of a single generalization–specificity continuum.

At one end of the continuum is found the VG hypothesis, which implies that a single (nonzero) population validity is invariant across all situations. The VG hypothesis may be formally

stated as a compound hypothesis (a) after correcting for statistical artifacts the estimate of $|\rho| > 0$ and (b) the estimate of $V(\rho) = 0$. Thus, the VG hypothesis states that there is a single, invariant (or fixed) "true" population correlation between the predictor and the criterion. Any deviations that are observed within a sample from this fixed value may be attributed entirely to measured (e.g., sampling error, measurement error, range restriction) and/or unmeasured artifacts (e.g., clerical errors). Evidence to support the VG hypothesis is furnished by demonstrating that 75% of the variance in local estimates is attributed to various forms of statistical artifacts, with the presumption being that the remaining 25% is attributed to other artifacts that are not quantifiable (e.g., clerical errors). Consequently, 100% of the variance in observed validities may be attributed to noise in the system, and there is nothing unique about situations (and by extension, there are no moderators—situational or otherwise).

At the other end of the continuum, we find the SS hypothesis, which implies that non-trivial variability in test validities is not attributed to measured and unmeasured artifacts. The strong SS hypothesis is agnostic with respect to the estimate of the mean validity (i.e., it could be zero, positive, or negative), but instead is focused solely on the true variability in validities. To understand the true variability in validities necessitates an understanding of the agonists of this variability (i.e., moderator variables, situational or otherwise).

Finally, residing in the middle of the continuum we find what might be labeled a "weak" SS hypothesis (or "weak" VG hypothesis, depending on one's theoretical proclivities); this hypothesis embraces the notion that the mean validity may likely be different from zero but also predicts significant variability around the mean. This middle-of-the-road hypothesis may also be considered consistent with the concept of transportability (discussed later in the chapter). Thus, to properly understand local validity estimates, one must also understand the critical differences arising across the situations where those local estimates were obtained—differences that are not entirely explained by measured and unmeasured statistical artifacts. These differences may be driven by moderator variables (situational or otherwise). In the context of test validation, one might formally state the weak SS hypothesis as a compound hypothesis (a) after correcting for statistical artifacts the estimate of $|\rho| > 0$ and (b) the estimate of $V(\rho) > 0$. Thus, this hypothesis states that there may be multiple (or variable) "true" population correlations existing between the predictor and the criterion.

In general, research in applied psychology has revealed limited support for the VG hypothesis but greater support for the SS hypotheses. Evidence consistent with the SS hypotheses exists because typical VG analysis rarely frees up all of the between-sample variance in validity coefficient estimates, sometimes freeing up very little variance for certain predictor-criterion pairs and/or job types (i.e., Murphy, 2000; Ones, Viswesvaran, & Schmidt, 2003; Salgado et al., 2003). For example, several VG analyses performed on tests of cognitive ability have revealed the moderating role of job complexity on correlations between ability/KSAs and performance criteria (i.e., Hunter & Hunter, 1984; Levine, Spector, Menon, Narayanan, & Cannon-Bowers, 1996; Russell, 2001; Salgado et al., 2003; Schmidt et al., 1993). Hunter and Hunter (1984) found that cognitive ability demonstrated a higher validity for predicting job performance and training success for occupations involving greater task complexity (i.e., the validity of cognitive ability tests is not invariant but fluctuates across situations as those situations vary in levels of task complexity). Likewise, Salgado and colleagues (2003) found that the empirical validity for general mental ability varied as a function of job type, with correlations ranging from .12 for police officers to .34 for sales occupations when predicting supervisor ratings of job performance. Relatedly, Schmidt and colleagues (1993) found that validity estimates for various measures of cognitive ability (i.e., general, verbal, quantitative, reasoning, perceptual speed, memory, and spatial and mechanical) varied (at least in part) as a function of job type, where the standard deviation of the validity estimates for reasoning ability predicting job performance was .04 for jobs involving stenography, typing, and filing and .19 for production and stock clerks.

Similar support for SS hypotheses has been obtained for personality traits, especially in the case of team-oriented organizations (Barrick & Mount, 1991, 1993; Mount, Barrick, & Stewart, 1998; Stewart, 1996; Stewart & Carson, 1995). For example, Barrick and Mount (1993) found that the validity of key personality traits (conscientiousness and extraversion) as predictors of job performance (as rated by supervisors) varied over managers as a function of managers'

perceived level of autonomy on the job. Validities tended to increase in proportion with the amount of perceived autonomy. Additionally, Mount and colleagues (1998) showed that some Big Five traits were more valid than other Big Five traits, but the dominant trait varied as a function of the degree to which situations demanded social and interpersonal interactions. To illustrate, agreeableness and extraversion had stronger validities for predicting performance for employees working in situations emphasizing team-oriented jobs (e.g., highest mean validities were .24 and .20, respectively) compared to the validities that were observed for employees working in clerical and "cubicle" jobs/situations that emphasized dyadic interactions (i.e., newspaper employees in the circulation department; banking employees in loan operations; telemarketing representatives; highest mean validities were .16 and .16, respectively). In contrast, the opposite was true for conscientiousness. Specifically, dyadic jobs yielded greater validity estimates (e.g., highest mean validity was .29) than did team-oriented jobs (e.g., highest mean validity was .19). Moreover, even when validities are examined for one specific job type (e.g., sales), validities vary for the extraversion–sales effectiveness relationship across organizations, with only 54% of their variance being accounted for by statistical artifacts (Barrick & Mount, 1991; Stewart, 1996).

Along these lines, *trait activation theory* (Tett & Burnett, 2003) posits that work situations send cues to employees about what personality traits may be relevant for a given situation. Thus, features of a situation may serve as triggers of (or inhibitors for) the expression of personality-based work behaviors. A number of studies have supported the basic tenets of trait activation theory, including its relevance for personality (e.g., agreeableness) as a predictor of outcomes such as innovation and creativity (Hunter & Cushenberry, 2015) and for better understanding the construct validity paradox that has troubled assessment center researchers for many years (Lievens, Chasteen, Day, & Christiansen, 2006; Lievens, Schollaert, & Keen, 2015).

Similarly, the strength of a situation (i.e., how much a situation restricts or inhibits behavior; Mischel, 1968) may moderate the magnitude of correlations between individual differences and work-related outcomes. For example, in a meta-analysis of the relationship between trait conscientiousness and job performance, Meyer, Dalal, and Bonaccio (2009) found that this relationship was significantly moderated by the strength of the work situation. For example, this correlation was weaker for jobs nested in very strong situations (.09 for nuclear equipment operation technicians working in a highly regulated work context) and was stronger for jobs nested in weaker situations (.23 for barbers working in a less regulated and more creative environment). More recently, Meyer and colleagues (2014) developed a measure of work-related situational strength and found that it significantly moderated the relationship between contextual work behaviors and the traits of conscientiousness and agreeableness. For example, agreeableness demonstrated a stronger relationship to organizational citizenship behaviors in weaker situations (i.e., where employees had greater discretion over their work activities; see also Meyer, Dalal, & Hermida, 2010, for a more detailed review of the situational strength concept and its relationships).

The above studies (and the concepts of situational strength and trait activation) were meant to be illustrative, not exhaustive. Overall, the theoretical models driving applied psychology are not simple, bivariate models (i.e., X correlates with Y). Rather, these models are inherently complex, multivariate, and regularly invoke mediating and moderating mechanisms. With respect to moderating mechanisms, situations continue to play central roles in our models, and the hypotheses derived and tested therefrom (and, it has long been recognized that test validation is simply a specific form of hypothesis testing; Binning & Barrett, 1989; Landy, 1986).

Overall, the results for both cognitive and non-cognitive selection tests indicate that (a) the mean correlation between job-relevant tests of KSAs or personality traits and work-related outcomes is often non-zero; (b) there is often considerable variance around these non-zero mean validities; and (c) in many instances, this non-trivial variance in validities may be attributed to situational variables that moderate the strength of the validities. This pattern of findings is consistent with the logic underlying the SS hypotheses and implies that idiosyncratic characteristics of the testing situations (e.g., situational strength, trait-activating situations) may be exerting a non-trivial influence on the predictive validity of many employment tests. Further buttressing the arguments of the SS hypothesis is the general finding that most recent meta-analytic reviews test for and report evidence of moderation (Aytug, Rothstein, Zhou, & Kern, 2012; Geyskens et al., 2009). Obviously, not all of these reviews considered

situational moderators, but many did. Finally, consistent with the SS hypotheses is the recommendation by the developers of VG to rely on random effects versus fixed effects models (fixed effects models are consistent with the VG hypothesis and random effects models are consistent with the SS hypotheses). In summary, when viewed as a continuum ranging from strong VG to strong SS, the extant literature indicates that the validities of most employment tests fall toward the middle or the SS end of the generalization–specificity continuum.

PSYCHOMETRIC META-ANALYSES AND THE FUTURE OF VALIDITY GENERALIZATION

Although the adoption of random effects meta-analytic procedures and the continual search for moderators has largely put to rest the VG versus SS debate, a number of important concerns continue to exist regarding the procedures of PMA that underlie all VG analyses (cf. Bornstein, et al., 2009; Hunter & Schmidt, 2004). These concerns are of a sufficient magnitude to warrant a review and discussion of how they might impact the future use of PMA/VG procedures (or the interpretation of previous studies relying on PMA/VG procedures).

Concern 1: Failure to (Explicitly) Model Situational Attributes

A serious concern with current PMA/VG methods is that situational variables are actually never included as part of the formal statistical model. Instead, a number of prominent situational attributes have been systematically omitted from meta-analytic summaries that have relied on PMA/VG procedures. Thus, potential moderators of predictor-criterion relationships in meta-analyses have not been formally tested, and the structural models linking predictors to criteria may be considered mis-specified (James, Mulaik, & Brett, 1982). Potential situational moderators include organizational contextual variables such as authority structure, standardization of job tasks and procedures, reward processes, leadership styles, organizational culture, and organizational climate. Thus, the pervasiveness of situational moderators has likely been *underestimated* in studies relying on the PMA/VG procedures; thus, our previous discussion of the selection literature should be interpreted accordingly.

Returning to Equations 1 and 2, it is clear that no substantive situational variables are actually taken into account in a PMA/VG analysis. Only basic statistical properties of the measurement scores are "corrected." Because situational variables are not included as substantive variables in meta-analytic summaries, it is impossible to ascertain which situational variables might (or might not) influence a particular predictor-criterion relationship. Although more recent applications of PMA/VG have included formal tests for moderators, these moderators have often been methodological in nature (e.g., student vs. field samples, objective vs. subjective criteria) rather than substantive in nature (e.g., interpersonal or dyadic interactions; social climate).

The PMA/VG estimation approach is based on residualization (where artifact variance is removed *before* testing for moderation) and is disconcerting to proponents of SS hypotheses, who certainly could argue that situational variables should be measured and formally tested as moderators before these variables are simply rejected as sources of error. Instead, current applications of PMA/VG largely take moderators into consideration on a *post hoc* basis, after the estimate of $V(\rho)$ is found to differ from zero (Cortina, 2003).

The residualization approach to PMA analysis offers a problematic test of SS hypotheses. This problem can be broken down into concerns related to statistical power and concerns related to knowledge of quantifiable differences between and within studies that could act as moderating effects. Situational factors may impact validity estimates; however, many PMA studies include a limited number of primary studies (k) and/or the primary studies that are included may have used a small sample size (N), and both of these factors have been linked to insufficient power for detecting moderation effects (i.e., Alexander & DeShon, 1994; Cortina, 2003; James, Demaree, Mulaik, & Mumford, 1988; James et al., 1986; Murphy, 2000; Spector & Levine, 1987; Steel & Kammeyer-Mueller, 2002). When the number of studies (or number of studies used

to explore a specific effect) in a PMA is low, the power to detect moderating effects is similarly low as PMA/VG relies on any existing variation between study effect sizes as the indication of moderation. Thus, adequate power (i.e., sufficiently large sample sizes and number of studies in the PMA/VG study) is requisite for any reasonable test of the SS hypothesis. Insufficient power to detect moderation should preclude an interpretation in favor of VG or against SS.

Most moderators studied in meta-analyses are explored in a *post hoc* manner, involve methodological (vs. truly substantive situational) moderators, and rely on the extent to which those conducting the PMA wish to code the studies for these methodological moderators. As a consequence, many moderator variables are included because they are conveniently coded, not because they are derived from strong psychological theories. Indeed, many of the moderators explored in selection contexts are likely of little interest to most organizations. As an example, around 90% of U.S. firms employ fewer than 20 employees and 98% employ fewer than 100 individuals (U.S. Census Bureau, 2012), yet many meta-analyses are conducted on studies where the samples are from large companies, college students, or government employees. Important contextual variables, such as spans of control, reward structures, unemployment risk, benefits, flexibility, justice climate, perceived organizational support, and a host of other variables are likely to vary widely not only between large and small work organizations but also within the subgroup of smaller work organizations. Data from individuals within these contexts are rarely collected and, even if they were, this contextual information may not be reported in individual validation studies. Thus, such contextual variables cannot be explored in moderator analyses in PMA. Selection tools may exhibit high levels of consistency when most samples studied in the PMA come from similar contexts (e.g., large organizations or government samples), but that does not allow researchers to assume that the consistent effects uncovered from those substantively similar contexts can be transported to situations that were systematically excluded by the PMA (e.g., smaller organizations).

Similarly, the moderator variable of interest must vary between studies to be explored with PMA/VG. A moderation effect could be reported in every study explored within a PMA. However, unless subgroup means and variances are reported in each of those studies, there is no way to extract the necessary information to explore those subgroup effects with PMA. Additionally, many variables are continuous rather than categorical (such as span of control, unemployment risk, role ambiguity, or leader trust and support). Moderating effects for continuous variables in primary studies are explored after first computing a cross-product term. There currently exists no way to summarize and test such continuous moderating effects that may be reported within the individual studies included as part of a PMA (DeSimone & Schoen, 2015). Although "moderator variables (interactions) never studied in any individual study can be revealed by meta-analysis" (Hunter & Schmidt, 2004, p. 26), it is also true that meta-analysis does not have a test that is analogous to moderation tests used in primary studies (DeSimone & Schoen, 2015; Podsakoff, MacKenzie, Ahearne, & Bommer, 1995).

In summary, the advocacy for using PMA/VG as methods for uncovering moderators (see Hunter & Schmidt, 2004, p. 26) coupled with the strong statements regarding "proof" of cross-situational consistency (i.e., no moderators; Hunter & Schmidt, 2004, pp. 404–405) paints a confusing picture. PMA/VG can be used to detect some forms of moderation, but only when sufficient variation exists between studies on the construct of interest. In addition, moderators in PMA/VG studies are often included and tested because they are conveniently coded, not because they represent critically relevant constructs derived from strong psychological theory. Finally, the current PMA/VG techniques may only be used when the moderator is categorical, no techniques have been developed that will accommodate continuous moderator variables in a PMA/VG analysis.

Concern 2: Untested Statistical Assumptions

Statistical Artifacts and Validities Must Be Independent of Situational Variables

Several researchers have exposed a critical, implicit assumption underlying the use of PMA/VG procedures—namely, that the effects of situational variables and statistical artifacts on validity coefficients must be *independent* (i.e., orthogonal) of one another (Burke, Rupinski, Dunlap, &

Davison, 1996; James et al., 1986; James et al., 1992; Raju, Burke, Normand, & Langlois, 1991; Thomas & Raju, 2004). However, this assumption has rarely been discussed or formally tested in PMA/VG studies.

Consider the example offered by James and colleagues (1992). They argued that variations among organizations (i.e., situations) in the restrictiveness of organizational climate would likely engender variations in criterion reliability. Restrictiveness of climate encompasses various organizational features that create strong versus weak work situations, including authority structures, standardization of job tasks and procedures, and reward structures (James et al., 1992). A highly restrictive climate (i.e., strict rules, guidelines, steep hierarchical structure, reward system not based on individual merit) would likely contribute to a decreased expression of individual differences among employees on performance because of a tendency toward compliance and conformity. This variance restriction should, in turn, attenuate criterion reliability and any relationship between these variables and job functioning (i.e., true validity) compared to what might be expected in a less restrictive climate (i.e., open communication, fewer restrictions and rules, reward system based on individual merit).

If a situational variable such as restrictiveness of climate jointly affects the magnitudes of validities *and* criterion reliabilities, then the VG model is likely to include a covariance between validities and the reliabilities. Covariation between validities and a statistical artifact such as criterion reliability violates the fundamental assumption that *these factors are statistically orthogonal to one another*. Covariation between validities and criterion reliabilities implies that removing variance in validities associated with variance in criterion reliabilities likely results in removing variance due to true situational factors (e.g., restrictiveness of climate). This is because variation in the situational variable (climate) serves as a *common cause* for variation in validities and criterion reliabilities. To remove variance due to reliability is to remove variance due to its causes—the situational variable. It follows that one is likely to increase his/her chances of (erroneously) rejecting the SS hypothesis by incorrectly attributing variance in validities to statistical artifacts, when in fact that variance is attributed to situational features (climate) that impacted both the validities and the artifacts (e.g., reliabilities).

In response to a concern of interdependencies among validities and statistical artifacts, two alternative models were introduced. One model, proposed by James and colleagues (1992), addressed this assumption of independence directly. Specifically, their model removed the assumption of independence by including covariance terms between validities and each of the statistical artifacts included in the PMA/VG correction equations. The second model, proposed by Raju and colleagues (1991), attempted to circumvent the problem engendered by lack of independence. Their approach corrected for unreliability, attenuation, and sampling error within each individual sample prior to averaging validities (i.e., predictor-criterion correlations) across studies. Therefore, violation of the assumption within studies is no longer an issue, although violation of the assumption across studies remains unresolved (Thomas & Raju, 2004).

Thomas and Raju (2004) tested and compared the accuracy of these two approaches. Although no comparison was made between results obtained by application of the model presented by James and colleagues (1992) versus the traditional PMA/VG estimation equations, results from the method developed by Raju and coauthors (1991) surpassed the traditional VG model in accuracy. Furthermore, the two models (i.e., James et al., 1992 and Raju et al., 1991) demonstrated comparable properties in accurately estimating validity coefficients. The method proposed by Raju and colleagues provided slightly more stable estimates (i.e., lower variance in estimates across samples). One consequence of the latter finding is that it provides additional support for the SS hypothesis (e.g., the residual variances of these estimates, which can be interpreted as arising from situational influences, tend to increase using these procedures). Of course, neither model identifies which, nor in what way, situational variables moderate validities.

Statistical Artifacts Must Be Independent of One Another

In addition to the assumption that situational factors are uncorrelated with effect sizes and statistical artifacts, PMA/VG procedures also invoke the assumption that the statistical artifacts are "independent of each other" (Hunter & Schmidt, 2004, p. 139). Thus, range restriction,

predictor reliability, and criterion reliability are all assumed to be statistically orthogonal from one another. The tenability of this assumption was recently challenged by Köhler and colleagues (2015), who conducted two large-scale meta-analytic reviews of different types of reliability coefficients (with data from 518 and 347 studies, respectively). Contrary to the statistical assumptions that form the basis of all PMA/VG procedures, reliability coefficients obtained from primary studies were often not independent of one another.

In Study 1, Köhler and colleagues (2015) summarized the degree of correlation between different types of reliability coefficients based on articles published in the *Journal of Applied Psychology* and the *Academy of Management Journal*. They found that the degree of correlation between different types of reliability coefficients ranged in magnitude from small to large. For example, the correlation between predictor reliabilities estimated using coefficient alpha and criterion reliabilities estimated using intra-rater correlations was nearly orthogonal at −0.03. In contrast, the correlation between predictor reliabilities estimated using coefficient alpha and criterion reliabilities estimated using inter-rater correlations was −0.45. It is also important to note that the reported correlations were obtained using a highly range restricted set of data (i.e., studies from only two of the top academic journals were included, and the various forms of reliability were, on average, quite high). PMA/VG studies are often based on collections of effect sizes sampled from a broader array of sources (e.g., greater number of journals and/or unpublished studies placed in the "file drawer" (possibly due to low reliabilities)). Thus, the correlations between reliability coefficients provided in Study 1 by Köhler and coauthors likely underestimate the magnitude of these associations, and thus underestimate the extent to which the fundamental assumptions underlying PMA/VG have been violated.

In Study 2, Köhler and colleagues conducted a meta-analysis on the relationships between perceived organizational support (POS) and a number of its antecedents and consequences. Their results further confirmed that this fundamental assumption of PMA/VG analyses may be routinely violated (i.e., predictor and criterion reliabilities were correlated across studies). These authors noted that "correlations between reliabilities are quite substantial" (p. 376), with values ranging from −0.80 to +0.34. Consequently, they counseled researchers to take into account the correlation between reliabilities, lest they overcorrect the observed mean effect size (see Köhler et al., 2015, p. 381).

Thus, although the assumption that reliabilities are independent has gone largely untested in individual PMA/VG studies, the results from Köhler and coauthors suggest that this assumption may be untenable. Violating this assumption impacts (i.e., biases) not only the estimation of between-study variation, $V(\rho)$, but also the estimate of the true score validity, ρ (Köhler et al., 2015). Estimates of true score validity that are upwardly biased are especially problematic in selection contexts (see our Concern 5 below).

In summary, the statistical foundation for all PMA/VG studies includes a core set of assumptions regarding the independence of study artifacts from one another (e.g., predictor reliability, criterion reliability, predictor range restriction) and the independence of situational variables from both statistical artifacts and observed criterion-related validities. Although attempts have been made to develop better estimating models and procedures (James et al., 1992; Raju et al., 1991), most PMA/VG studies continue to use the estimating equations presented by Schmidt and Hunter (see Aytug et al., 2012) that assume artifacts meet these two critical orthogonality assumptions. However, there is growing awareness in the extant literature that both of these assumptions may frequently be violated, resulting in misleading estimates of the mean true validity and the between-study homogeneity in validities (i.e., violating these assumptions results in biased estimates used to infer VG vs. SS).

Concern 3: Questionable Estimates Used to Correct for Artifacts

Our third concern related to PMA/VG procedures relates to the specific values that are used to represent the statistical artifacts in "correction" equations. Even if data meet the necessary statistical assumptions (which appears increasingly unlikely; see Concern 2), the correction equations that form the basis of PMA/VG analyses will only yield accurate (i.e., unbiased) estimates

of the population parameters (i.e., true validity correlation) when accurate (i.e., unbiased) estimates of the statistical artifacts are inserted into proper correction equations. Unfortunately, there is growing concern that a number of previously published PMA/VG studies may have relied on inaccurate estimates of statistical artifacts. The most contested application of the PMA/VG procedures has arisen when the criterion variable has been measured using supervisory performance ratings, arguably the most commonly used criterion in test validation and personnel decision making.

In a highly cited meta-analytic review, Viswesvaran and colleagues (1996) examined the reliability of both supervisory and peer performance ratings. The authors found that the sample-size weighted mean estimates of inter-rater reliability were quite low. More specifically, the mean estimates of inter-rater reliability were found to be .52 and .42 for data obtained using supervisor and peer ratings of performance, respectively. These estimates of reliability stand in stark contrast to estimates based on test-retest reliabilities and internal consistency reliabilities. For example, supervisory ratings demonstrated a mean test-retest reliability of .81 and an even stronger mean internal consistency reliability of .86. Temporal stability data were not available for peers, but the peer data mirrored the supervisor data with respect to internal consistency, with an average reliability of .85.

Although ratings data appeared to be both internally consistent and reasonably stable, Viswesvaran and coauthors argued that the preferred estimate of reliability was furnished by the inter-rater reliability coefficients. Consequently, with a few notable exceptions (e.g., Meriac, Hoffman, Woehr, & Fleisher, 2008), the inter-rater reliability estimates provided by Viswesvaran and colleagues have become the default values used for artifact distributions to make statistical corrections in PMA/VG studies based on performance ratings.

To be clear, we do not have concerns with the breadth of the study by Viswesvaran and colleagues (1996), as we are confident they did a thorough and competent review. As such, we are confident that the mean estimates of inter-rater reliability reported by Viswesvaran and coauthors reflect the values reported in the literature. In addition, we do not have concerns with the use of correction equations (assuming of course that the necessary statistical assumptions have been met; see Concern 2). Instead, our concern is actually more fundamental and may be broken into three subcomponents: (1) the appropriateness of using inter-rater correlations to estimate the reliability of performance ratings, (2) the bias that exists in sample estimates of statistical artifacts, and (3) the questionable use of measures with such poor psychometric properties for the inferential decisions suggested by PMA/VG.

Appropriateness of Inter-rater Correlations to Index Reliability

First, we examine whether the inter-rater correlation is the most appropriate statistic for estimating the reliability of performance ratings? This statistic defines reliability as the correlation between two parallel (in a psychometric sense) raters. There is a growing consensus that the raters who furnish the data (e.g., performance ratings of a target individual provided by three different supervisors) do not meet the stringent requirements necessary to be treated as psychometrically "parallel" measures of performance (Murphy & DeShon, 2000; Putka & Hoffman, 2015; Putka, Hoffman, & Carter, 2014; Putka, Le, McCloy, & Diaz, 2008).

For example, Putka and colleagues (2014) noted that ratings obtained from two supervisors may not be truly parallel measures. They nested their compelling arguments in the extant literature (e.g., Trait Activation Theory; leader-member exchange theory) and described how estimates of criterion reliability based on interrater reliabilities may result in overcorrected (or undercorrected) validities. Specifically, they provided a new correction equation that would allow for two raters to provide distinct (yet valid) information about employees' job performance. They noted that their new equation "reduces to the traditional correction if those unique perspectives completely reflect performance-irrelevant, accidental error" (p. 546). If the supervisory or peer ratings are not parallel evaluations of employees' performance, then the application of the traditional correction equation should not be used, and instead we direct the interested reader to the work of Putka and coauthors.

Biased Sample Estimates of Statistical Artifacts

Second, even if one assumes that the inter-rater correlations are based on essentially parallel ratings (cf. LeBreton et al., 2003; Schmidt, Viswesvaran, & Ones, 2000), one must still obtain *unbiased sample estimates of these inter-rater reliability coefficients*. More specifically, if the variability in job performance has been restricted (e.g., due to interventions such as valid and effective recruitment programs, the use of valid selection tools, the use of effective training interventions, attrition stemming from lack of person–organization fit, a restrictive climate/culture, or other potential causal mechanisms), then (like any correlation coefficient) observed estimates of inter-rater reliability will be downwardly biased (cf. Burke, Landis, & Burke, 2014; Huffcutt, Culbertson, & Weyhrauch, 2014; LeBreton et al., 2003; Sackett, Laczo, & Arvey, 2002). Thus, prior to "correcting" correlations for criterion unreliability, it is imperative to obtain an unbiased estimate of inter-rater reliability by correcting the attenuated reliability estimate for range restriction. Indeed, as the developers of PMA/VG procedures lamented, "In the typical validation study, the criterion reliability, as well as the test validity, is available only on the restricted group. Both coefficients should be corrected first for restriction in range. The validity coefficient should then be corrected for attenuation" (Schmidt, Hunter, & Urry, 1976, p. 475). Interestingly, rather than calculate the unrestricted estimates of inter-rater reliability, Viswesvaran and colleagues opted to meta-analytically summarize the restricted (i.e., downwardly biased) estimates of inter-rater reliability. Individuals who are familiar with the correction equations should appreciate how using downwardly biased estimates of inter-rater reliability will yield upwardly biased estimates of criterion-related validity and, obviously, restrict the possibility of finding situational moderators.

Recently, Huffcutt and colleagues (2014) advocated using a multistage artifact correction process. Specifically, they noted that statistical artifacts (e.g., inter-rater reliability estimates of job performance) could be biased by other statistical artifacts (e.g., measurement error, range restriction). They suggested that prior to making corrections for statistical artifacts as part of a PMA/VG study, one should first seek to obtain the most accurate (i.e., unbiased) estimates of those artifacts. Huffcutt and coauthors reanalyzed data summarizing the criterion-related validity of the employment interview and documented how their approach yields less biased (i.e., more accurate) estimates of the true validity.

On a similar note, Burke and colleagues (2014) recommended that corrections based on supervisory ratings of job performance should be made using a range of potential estimates of inter-rater reliability; this range of estimates should be selected so as to reflect situations that controlled for extraneous variables likely to attenuate estimates of inter-rater reliability. They offered a point-estimate inter-rater reliability (based on supervisor ratings) of .80 rather than the estimate of .52 that has permeated previous PMA/VG studies (and which has likely yielded inflated/overcorrected estimates of population validities). We also agree with the observation of Burke and coauthors that multilevel issues are becoming more common in meta-analyses (and thus will require different estimation formulae; see also comments by Sackett, 2014).

Suspending Psychometric Standards in PMA/VG Studies

Finally, we are troubled that, with rare exceptions (e.g., Meriac et al., 2008), applied psychologists wishing to measure job performance via ratings have largely ignored traditional standards for measurement and instead embraced psychometrically questionable measures (i.e., supervisory performance ratings). We echo the sentiments presented by LeBreton and colleagues (2014) that applied psychologists should not apply different standards to evaluate the quality of predictor measures versus criterion measures. As these authors noted, many of the leading psychometricians and applied psychologists of the 20th and 21st centuries have emphasized the importance of using measurement systems yielding reliable assessments of the target constructs, and this is true for *both* predictor and criterion constructs:

> [desirable reliabilities] usually fall in the .80s or .90s.
>
> (Anastasi, 1968, p. 78)

> a test should have a minimum reliability coefficient of at least .94. Some have been more liberal in this regard, allowing a minimum of .90.
>
> (Guilford & Fruchter, 1973, p. 91)

> Relevancy [is] the first requirement for a criterion. . .. Reliability is the second requirement of a criterion.
>
> (Smith, 1976, p. 746)

> Most texts in industrial psychology contain lengthy lists of requirements for criteria. . .. [these lists] might be reduced to three requirements: reliability, validity, and practicality.
>
> (Landy, 1985, pp. 150–151)

> the minimum accep level of reliability for psychological measures in the early stages of development is .70 (Nunnally, 1978). Higher levels may be required of measures . . . used in advanced field research and practice.
>
> (Nunnally & Bernstein, 1994, p. 839)

> A relevant, reliable, and uncontaminated criterion measure(s) must be obtained or developed.
>
> (SIOP Principles, 2003, p. 14)

> if criteria are to be useful, they must be measureable in a consistent manner.
>
> (Ployhart, Schneider, & Schmitt, 2006, p. 167)

In summary, in order to justify using the inter-rater reliability values reported by Viswesvaran and his colleagues, one must avow that (a) inter-rater correlations provided in primary studies meet the statistical assumptions requisite for interpretation as a reliability coefficient and (b) the estimates of inter-rater reliability have not been attenuated by other statistical artifacts. However, even if one were willing to concede that these criteria have been met, we would argue that performance ratings with reliabilities in the .40s and .50s should not be used as the basis for making critical personnel decisions (e.g., which tests are deemed "valid"; which employees should be hired, fired, promoted, rewarded, or punished).

Critical decisions that impact the lives of individuals should only be based on reliable measures (and this is true for both predictor and criterion constructs). The absence of such assessments from the tool chest of applied psychology is not sufficient justification for embracing measurement systems where 50% to 60% of the observed variance is error variance; if assessments with such questionable psychometric properties were acceptable, applied psychologists would still be using projective tests to select employees (see Lilienfeld, Wood, & Garb, 2000, for a review of projective tests).

Concern 4: Imprecise Inferences Drawn from "Corrected" Validities

Our fourth concern relates to the potential for applied psychologists/practitioners to arrive at misleading inferences about selection systems from PMA/VG studies. In particular, we urge applied psychologists/practitioners to be wary of using corrected correlations for evaluating the practical benefit of a selection test. Consider the classic validity model presented in Figure 4.1 (used in LeBreton et al. (2014) and adapted from Binning and Barrett (1989)).

Within the context of a local validation study, Inference 3 may be conceptualized as the observed correlation between a predictor measure (e.g., Watson-Glaser critical thinking inventory) and a criterion measure (e.g., ratings of computer programmer job performance furnished by a supervisor). Inferences 2 and 4 may be conceptualized as representing the reliability of the predictor and criterion measures. If certain statistical assumptions are tenable, then reliability estimates may be computed using scores obtained from the predictor and the criterion. Finally, Inference 1 represents the hypothetical/conceptual/theoretical relationship between the latent constructs that are assessed via the predictor test and the criterion measure. In our example, these constructs might be labeled "general intelligence" and "performance."

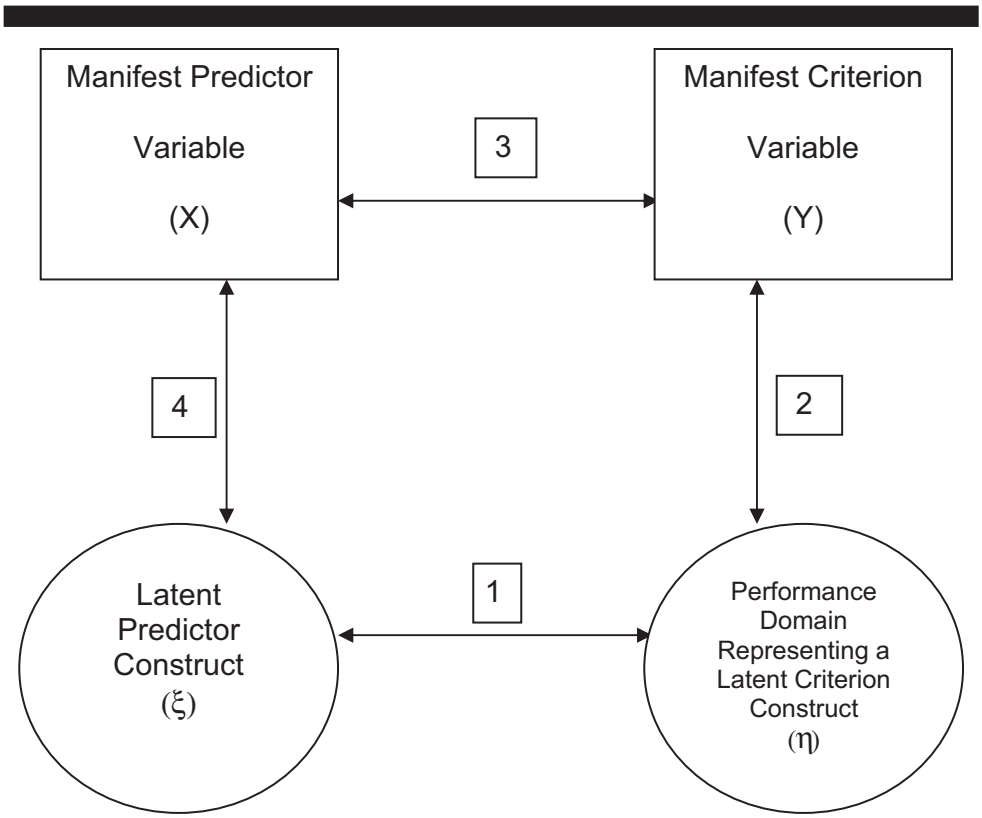

FIGURE 4.1 Basic Construct Validation Model

Note: Figure 4.1 originally appeared as Figure 1 in: LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Validity generalization in a land of suspended judgment. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*, 478–500 (copyrighted by Cambridge University Press). This figure is reprinted with permission.

Within the context of a PMA/VG study, Inference 3 may be conceptualized as the sample-weighted (or unweighted) mean observed correlation taken over samples and (often) based on different assessments within and between samples. For example, Study 1 might use the Watson-Glaser and Supervisory Performance Ratings (furnished by a single supervisor); Study 2 might use the Wonderlic Personnel Test and Supervisory Performance Ratings (perhaps averaged over two supervisors); Study 3 might use Raven's Progressive Matrices and objective indicators of sales performance; etc. Inference 3 simply represents the average validity taken over these studies. Depending on the particular approach to PMA/VG, Inferences 2 and 4 may represent (a) the average observed reliabilities of the predictor and criterion measures or (b) an estimate of these reliabilities obtained using extant artifact distributions. Inference 1 again represents the hypothetical/conceptual/theoretical relationship between the latent constructs that are assessed via the different predictor tests and the different criterion measures.

Meaning of Corrected Coefficients

Corrected coefficients are hypothetical estimates of what the relationship between predictor and criterion might look like if certain assumptions have been met. Most notably that "one had access to an infinitely long predictor and an infinitely long criterion (i.e., perfectly reliable measures representing a one-to-one correspondence between the [observed measures] and [the constructs they purport to assess])" (LeBreton et al., 2014, p. 491). We believe it is important

that both researchers and practitioners recognize the information that is (and is not) conveyed using corrected versus uncorrected coefficients. To be clear, there is nothing inherently "good or bad" about corrected coefficients. Indeed, corrected coefficients are routinely estimated as part of many applications of structural equations modeling (SEM; James et al., 1982). However, additional assumptions, as noted in Concern 2, are required when making these corrections in a PMA/VG study versus a primary study that implements an SEM analysis. Further complicating the interpretation of PMA/VG studies has been the estimation of a partially corrected "operational validity."

Operational validity is a term used to denote a coefficient that has been asymmetrically subjected to corrections. Specifically, the correlation (or mean correlation) is corrected for measurement error in the criterion (e.g., performance ratings) but is not corrected for measurement error in the predictor. However, when operational validities are estimated, applied psychologists often interpret these validities as "suggestive of how we should expect a selection test to perform "operationally" or "in practice" (LeBreton et al., 2014, p. 491). Consider statements such as:

> [operational validities are appropriate] because in actual test use we must use observed test scores to predict future job performance and cannot use applicants' (unknown) true scores.
>
> (Hunter & Schmidt, 2004, p. 126)

> We generally are not *per se* interested in a measured fallible indicator of performance; we want to know how well we predict the underlying construct reflected by that fallible measure.
>
> (Sackett, 2014, p. 502)

> But employee selection must be based on observed scores among applicants, thus the relevant relationship is the operational validity of the predictor set for the criterion construct.
>
> (Viswesvaran et al., 2014, p. 514)

Although we agree with Hunter and Schmidt (2004) that it is inappropriate to use applicants' (unknown) true scores on a selection test, we find it troubling that these authors (and many in our field) have no qualms about validating that imperfect and flawed test against the applicants' (unknown) true scores on the criterion (i.e., their perfect and unflawed scores on the latent criterion construct). The argument for making selection decisions using operational validities is based on the presumption that it is unfair to penalize the evaluation of a predictor measure by the measurement error tainting the criterion. However, the criterion-related validity estimate (i.e., correlation coefficient) represents a *joint relationship* between a predictor and a criterion. Indeed, evidence supporting the "validity" of inferences from test scores proceeds under the presumption that we have a highly reliable and relevant criterion. If our criterion is irrelevant and/or unreliable, then why bother looking for tests to predict that irrelevant and/or unreliable criterion?

Computing an operational validity places a disparate and asymmetrical emphasis on the predictors that make up a selection system. This approach to test validation is inconsistent with extant validation frameworks that have emphasized the importance of accumulating validity evidence for both predictors and criteria. Indeed, if a criterion is 50% random noise, why bother trying to predict it? The consequence of relying on operational validities for test validation has enabled applied psychologists to ignore the quality of criteria (hence the tendency to make corrections using criterion reliabilities in the 0.40s and 0.50s), which simply further inflames the criterion problem bemoaned for decades in applied psychology (Austin & Villanova, 1992; LeBreton et al., 2014; Wallace, 1965).

Operational validities seek to estimate not the strength of relationship one might expect to see "in operation" or "in practice" between a predictor and criterion, but instead to estimate the relationship between observed predictor scores and the latent (i.e., theoretical/hypothetical/conceptual) criterion construct. Returning to our example above, the operational validity for Study 1 represents the correlation between observed scores on the Watson-Glaser and a "perfect" criterion that was obtained by collecting supervisory performance ratings from an infinite number (or to "approximate" perfect reliability, at least a very large number) of (psychometrically parallel) supervisors. There is a very low likelihood that an organization will have access to

a large number of parallel supervisors for every single employee (cf. Murphy & DeShon, 2000; Putka & Hoffman, 2015; Putka et al., 2008; see also Concern 3).

Let's assume in Study 1 that the observed correlation between the Watson-Glaser and supervisors' performance ratings is .25. If we were to correct this observed correlation for measurement error in only the criterion (e.g., using the .52 estimate recommended by Viswesvaran et al. 1996), the operational validity increases to .35. Does this number really capture the quality of prediction obtained using this selection test to predict this criterion in this particular context?

Is it possible that an organization could, *in actual practice*, expect to see such an impressive validity? It depends. How many organizations typically estimate job performance as a unit-weighted average of performance ratings provided by 65 supervisors for each employee? Why 65 supervisors? Applying the Spearman-Brown prophesy equation using the .52 estimate reported by Viswesvaran and his colleagues, we find that it would take the ratings of 65 psychometrically parallel supervisors to obtain a criterion reliability of .99 (≈ 1.00), the value assumed to be tenable when undertaking the calculation of an operational validity.

We conclude that corrected coefficients (especially those that are asymmetrically corrected for only criterion unreliability) convey limited practical value. We are not alone in this judgment:

> corrected *r*s are of little practical value. . .. The prediction of one variable from another and the accompanying error of estimate must necessarily be based on obtained, or fallible, rather than true scores.
>
> (McNemar, 1962, p. 153)

> when one is faced with making inferences about behavior in the real world, it is not particularly useful to know how predictive a test would be if criterion measures were perfect.
>
> (Womer, 1968, p. 65)

> correcting for artifacts is not the proper goal of meta-analysis. The purpose of meta-analysis is to teach us what is, not what might be some day in the best of all possible worlds when all of our variables might be perfectly measured.
>
> (Rosenthal, 1984 as quoted in DeShon, 2003, p. 386)

> "corrections" confuse the essential distinction between what *might be* and *what is*. The observance of that distinction is the primary factor separating science from mere supposition. The forgetting of that distinction is the hallmark of validity generalization.
>
> (Seymour, 1988; italics in original, p. 352)

> Practitioners should be especially wary of validity estimates adjusted for unreliability . . . these adjustments are intended to provide theoretical estimates of the magnitude of a validity coefficient under conditions of perfect measurement. . . . There is nothing operational about "operational validity."
>
> (DeSimone, 2014, p. 530)

In summary, applied psychologists (especially those working in practice) are encouraged to focus their attention on the uncorrected correlations when interpreting the results from primary validity studies (or weighted mean uncorrected correlations when interpreting PMA/VG results). In contrast, psychologists interested in understanding hypothetical/theoretical/conceptual relationships that (might) exist between latent constructs may be better served by examining the corrected coefficients. Of course, this presumes that all necessary statistical assumptions have been met for estimating the corrected coefficients. Like the authors cited in the paragraph above, we see limited "practical" value in operational or corrected validities; instead, the value of corrected coefficients is in estimating what one (might) see in the theoretical/conceptual world where measurement error does not exist (or is very, very small; e.g., the world where each employee is rated by 65 supervisors who provide psychometrically parallel ratings).

Concern 5: Use of PMA Effect Sizes in Applied Contexts

It is one thing to compute a "corrected" correlation as an estimate of the hypothetical relationship between predictor and criterion constructs, but it is quite another thing to use that

hypothetical value as an excuse for not undertaking a local validity study. From a legal perspective, HR practitioners may be especially interested in knowing how cases relying on PMA/VG studies have been received by the Supreme Court of the United States (SCOTUS). Said differently, if PMA/VG evidence suggests a test yields a non-zero validity in predicting job performance, can a company safely rely on this information without conducting a local validity study?

First, we would recommend that practitioners who are planning to use effect size estimates obtained from PMA/VG studies should consider the magnitude of the reported effect. The courts have questioned the use of selection tools with low validities. Although the courts have been reticent to set a specific minimum cutoff for criterion-related validity coefficients, it does appear that tests with validities below .30 are questioned more heavily than are tests with validities higher than .30 (Biddle, 2010).

In addition, as discussed under Concern 4, practitioners looking to PMA/VG studies should be mindful of distinguishing between uncorrected and corrected validity coefficients. The corrected coefficients (frequently misrepresented as the true value, ρ, rather than the estimate, $\hat{\rho}$) furnish a theoretical estimate of what the effect size might be if everything in the situation was perfect (e.g., no measurement error, no range restriction). The corrected validity is a hypothetical ideal that does not exist in reality and can be easily contested in the courts (see Seymour, 1988). All things being equal, a better representation of what one might find in a local validity study is the value provided by the sample weighted uncorrected validity (often represented by $\bar{r}$). Given that the statistical assumptions underlying artifact corrections appear to be increasingly untenable (see Concern 2), the corrected validity may actually represent an overcorrected (i.e., inflated or upwardly biased) estimate of the true validity. That the corrected validity could be inflated is immediately relevant if one assumes the courts consider tests with validities above .3 as more valid than tests with validities below .3. Corrected validities, depending on the number and type of corrections used, may yield values nearly twice as large as their uncorrected counterparts (especially when assessments of inter-rater reliability are used to correct supervisor ratings of performance; see our Concern 3). Indeed, LeBreton and colleagues (2014) demonstrated that when corrections based on dubious estimates of criterion reliability are simultaneously applied to multiple predictor variables, it is possible to explain nearly 100% of the variance in job performance using only four or five selection tests. The implication is that situational variables (e.g., perceptions of climate, culture, justice, fairness, leadership, team cohesion, training interventions) are, thus, determined to be irrelevant to job performance.

We are aware of no case heard by the SCOTUS where PMA/VG studies have helped to win a case. For example, in a number of cases heard by the SCOTUS in the context of discrimination, the court has not looked favorably on PMA/VG evidence or PMA/VG expert testimony (see Biddle, 2010; Landy, 2003; Outtz, 2011). Thus, especially in the real-world selection contexts where organizations are faced with the possibility of discrimination lawsuits, relying only on PMA/VG studies appears ill-advised.

Outside of concerns regarding discrimination, advocates of PMA/VG today largely argue for the concept of transportability rather than true VG. Regardless of reported evidence for the existence of moderators, practitioners trying to interpret and use results from PMA/VG must remain cautious in interpreting the weaker transportability inference. Kemery and colleagues (1987) demonstrated that a 90% credibility interval that does not include 0 (i.e., consistent with transportability) could still include a large proportion of situations (e.g., jobs, work contexts) where the true validity was in fact 0. In short, just because a credibility interval does not include 0, one cannot unconditionally conclude that the selection test/tool in question is transportable to all situations.

Based on concerns of the potential impacts of discrimination and the possibility of low validities in certain jobs/context even when PMA/VG evidence is supportive of transportability, we recommend that practitioners augment any PMA/VG results with results obtained from a local validation study. Of course, local validity studies are not without potential limitations (e.g., sampling error associated with smaller sample sizes). There, however, are ways to combine the results of a local validity study with the results summarized in a meta-analysis (see Biddle, 2010; Brink & Crenshaw, 2011).

In summary, those hoping to draw conclusions from PMA/VG for selection purposes should be cautious. Meta-analysis can be a useful tool for summarizing research. However, PMA/VG studies should not be viewed as substitutes for a well-conducted local validity study. Those hoping to use effect size estimates from PMA/VG studies should carefully scrutinize the information reported. For selection purposes, we recommend interpreting/using the uncorrected validities rather than operational validities or fully corrected validities as the effect size estimates. We also recommend a careful analysis of the credibility interval. If the credibility interval is wide, even if the credibility interval does not include 0, then there are likely moderating effects between the studies included in the analysis. In addition, given the empirical evidence supporting situational specificity, the fact that most PMA/VG studies include tests for moderation (Aytug et al., 2012), and proponents of PMA/VG are now recommending random versus fixed effect models (Hunter & Schmidt, 2000; 2004), the tide appears to be strongly turning in a direction that further justifies the use of local validation studies. Finally, those looking to apply the results from PMA/VG studies should remember that the overall summary effect size is the average across populations; however, an effect size reported in a particular moderation analysis might be more representative of (i.e., consistent with) a practitioner's local situation/context.

CONCLUSIONS

Proponents of VG, and more recently PMA, have raised awareness of a number of important points. Studies with small sample sizes are subject to considerably less than desirable effects from sampling error. Measurement error and range restriction do attenuate the magnitude of validities. And the corrections used in PMA for measure unreliability are similar to those used in individual SEM studies without controversy. Finally, the quantitative summary of effect sizes, which was not a feature of qualitative reviews, is yet another strength of this method. However, like any statistical tool, PMA/VG is not without its controversies or limitations.

The first purpose of this chapter was to discuss one of these controversies, namely the extent to which criterion-related validities are situationally specific or generalize across situations. On balance, we found considerable empirical evidence consistent with the situational specificity hypotheses. This was not entirely surprising, given that the majority of theoretical models that make up the canon of the social sciences have adopted contingency, systems, or interactional perspectives.

The second purpose of this chapter was to catalog a list of concerns and limitations that have emerged with respect to the use of PMA/VG procedures, including:

1. Methodological concerns related to the systematic omission of situational variables from formal tests of VG
2. Statistical concerns related to the untenable nature of the assumptions underlying PMA/VG procedures
3. Methodological and statistical concerns related to the specific point-estimates used to estimate "corrected" validities
4. Theoretical concerns related to the drawing of improper inferences using corrected coefficients
5. Practical concerns related to use of corrected and/or operational validities (i.e., partially or asymmetrically "corrected" validities) as justification for not undertaking a local validation study

In summary, we conclude that PMA/VG procedures are useful statistical tools for elucidating the impact of statistical artifacts on validity estimates. However, we also conclude that the results of PMA/VG studies are incomplete and often misleading because they have failed to confirm that critical statistical assumptions were met or failed to include relevant situational variables from the actual tests of VG. To date, the overwhelming evidence favors hypotheses that situational moderators (i.e., variables that affect the magnitude of the validity of various tests in predicting employee performance) are not only possible but also quite likely.

NOTE

1. Portions of this chapter are based, in part, on a chapter co-authored by Larry James and Heather McIntyre, which appeared in the first edition of this Handbook. We are grateful to Heather McIntyre and Leslie James for permission to use the previous chapter as a starting point for this new and updated contribution to the second edition of this Handbook. Lawrence R. James was professor emeritus at Georgia Institute of Technology; this chapter is published posthumously.

REFERENCES

Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin, 2*, 308–314.

Algera, J. A., Jansen, P. G. W., Roe, R. A., & Vijn, P. (1984). Validity generalization: Some critical remarks on the Schmidt-Hunter procedure. *Journal of Occupational Psychology, 57*, 197–210.

Anastasi, A. (1968). *Psychological testing* (3rd ed.). New York, NY: Macmillan.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology, 77*, 836–874.

Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. (2012). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analysis. *Organizational Research Methods, 15*, 103–133.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.

Barrick, M. R., & Mount, M. K. (1993). Autonomy as a moderator of the relationships between the big five personality dimensions and job performance. *Journal of Applied Psychology, 78*, 111–118.

Biddle, D. A. (2010). Should employers rely on local validation studies or validity generalization (VG) to support the use of employment tests in Title VII situations? *Public Personnel Management, 39*, 307–326.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478–494.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, UK: John Wiley & Sons.

Brink, K. E., & Crenshaw, J. L. (2011). The affronting of the Uniform Guidelines: From propaganda to discourse. *Industrial and Organizational Psychology, 4*, 547–553.

Burke, M. J., Landis, R. S., & Burke, M. I. (2014). 80 and beyond: Recommendations for disattenuating correlations. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*, 531–535.

Burke, M. J., Rupinski, M. T., Dunlap, W. P., & Davison, H. K. (1996). Do situational variability act as substantive causes of relationships between individual difference variables? Two large-scale tests of "common cause" models. *Personnel Psychology, 49*, 573–598.

Buss, A. R. (1979). The trait-situation controversy and the concept of interaction. *Personality and Social Psychology Bulletin, 5*, 191–195.

Cortina, J. M. (2003). Apples and oranges (and pears, oh my!): The search for moderators in meta-analysis. *Organizational Research Methods, 6*, 415–439.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York, NY: Irvington.

DeShon, R. P. (2003). A generalizability theory perspective on measurement error corrections in validity generalization. In Kevin R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 365–402). Mahwah, NJ: Lawrence Erlbaum Associates.

DeSimone, J. A. (2014). When it's incorrect to correct: A brief history and cautionary note. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*, 527–531.

DeSimone, J. A., & Schoen, J. L. (2015). *Moderation effects not detectable by meta-analytic techniques*. Presented at the Society of Industrial and Organizational Psychology Annual Meeting, Philadelphia.

Endler, N. S., & Magnusson, D. (1976). Toward an interactional psychology of personality. *Psychological Bulletin, 83*, 956–974.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37*, 1097–1126.

Geyskens, I., Krishnan, R., Steenkamp, J-B. E. M., & Cunha, P. V. (2009). A review and evaluation of meta-analysis practices in management research. *Journal of Management, 35*, 393–419.

Ghiselli, E. E. (1959). The generalization of validity. *Personnel Psychology, 12*, 397–402.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests*. New York, NY: Wiley.

Ghiselli, E. E. (1973). The validity of aptitude tests. *Personnel Psychology, 26*, 461–477.

Grote, G. F., & James, L. R. (1991). Testing behavioral consistency and coherence with the situation-response measure of achievement motivation. *Multivariate Behavioral Research, 26*, 655–691.

Guilford, J. P., & Fruchter, B. (1973). *Fundamental statistics in education and psychology* (5th ed.). New York, NY: McGraw-Hill.

Hermelin, E., & Robertson, I. T. (2001). A critique and standardization of meta-analytic validity coefficients in personnel selection. *Journal of Occupational and Organizational Psychology, 74*, 253–277.

Hogan, R. (2009). Much ado about nothing: The person-situation debate. *Journal of Research in Personality, 43*, 249.

House, R. J., & Mitchell, T. R. (1974). Path-goal theory of leadership. *Journal of Contemporary Business, 3*, 1199–1237.

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2014). Multistage artifact correction: An illustration with structured employment interviews. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*, 548–553.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*, 275–292.

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

Hunter, S. T., & Cushenbery, L. (2015). Is being a jerk necessary for originality? Examining the role of disagreeableness in the sharing and utilization of original ideas. *Journal of Business and Psychology, 30*, 621–639.

James, L. R., Demaree, R. G., & Mulaik, S. A. (1986). A note on validity generalization procedures. *Journal of Applied Psychology, 71*, 440–450.

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

James, L. R., Demaree, R. G., Mulaik, S. A., & Mumford, M. D. (1988). Validity generalization: A rejoinder to Schmidt, Hunter, and Raju (1988). *Journal of Applied Psychology, 73*, 673–678.

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

Kemery, E. R., Mossholder, K. W., & Roth, L. (1987). The power of the Schmidt and Hunter additive model of validity generalization. *Journal of Applied Psychology, 72*, 30–37.

Kendrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist, 43*, 23–34.

Kerr, S., Schriesheim, C. A., Murphy, C. J., & Stogdill, R. M. (1974). Toward a contingency theory of leadership based upon consideration and initiating structure. *Organizational Behavior and Human Performance, 12*, 62–82.

Köhler, T., Cortina, J. M., Kurtessis, J. N., & Gölz, M. (2015). Are we correcting correctly? Interdependence of reliabilities in meta-analysis. *Organizational Research Methods, 18*, 355–428.

Landy, F. J. (1985). *Psychology of work behavior* (3rd ed.). Homewood, IL: Dorsey Press.

Landy F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183–1192.

Landy, F. J. (2003). Validity generalization: Then and now. In Kevin R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 155–195). Mahwah, NJ: Lawrence Erlbaum Associates.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*, 80–128.

LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for criterion reliability in validity generalization: A false prophet in a land of suspended judgment. *Industrial and Organizational Psychology, 7*, 478–500.

Levine, E. L., Spector, P. E., Menon, S., Narayanan, L., & Cannon-Bowers, J. (1996). Validity generalization for cognitive, psychomotor, and perceptual tests for craft jobs in the utility industry. *Human Performance, 9*, 1–22.

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91*, 247–258.

Lievens, F., Schollaert, E., Keen, G. (2015). The interplay of elicitation and evaluation of trait-expressive behavior: Evidence in assessment center exercises. *Journal of Applied Psychology, 100*, 1169–1188.

Lilienfeld, S. O., Wood, J. M., & Garb, H. M. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*, 27–66.

McDaniel, M. A., Kepes, S., Banks, G. C. (2011). The uniform guidelines are a detriment to the field of personnel selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 494–514.

McNemar, Q. (1962). *Psychological statistics* (3rd ed.). New York, NY: John Wiley and Sons.

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042–1052.

Meyer, R. D., Dalal, R. S., & Bonaccio, S. (2009). A meta-analytic investigation into situational strength as a moderator of the conscientiousness-performance relationship. *Journal of Organizational Behavior, 30*, 1077–1102.

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management, 36*, 121–140.

Meyer, R. D., Dalal, R. S., Jose, I. J., Hermida, R., Chen, T. R., Vega, R. P., Brooks, C. K., & Khare, V. P. (2014). Measuring job-related situational strength and assessing its interactions with personality and voluntary work behavior. *Journal of Management. 40*, 1010–1041.

Mischel, W. (1968). *Personality and assessment.* New York, NY: John Wiley.

Mischel, W., & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89*, 730–755.

Mount, M. K., Barrick, M. R., & Stewart, G. L. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human Performance, 11*, 145–165.

Murphy, K. R. (2000). Impact of assessments of validity generalization and situational specificity on the science and practice of personnel selection. *International Journal of Selection and Assessment, 8*, 194–206.

Murphy, K. R., & DeShon, R. P. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873–900.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

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

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2003). Personality and absenteeism: A meta-analysis of integrity tests. *European Journal of Personality, 17*, S19–S38.

Outtz, J. L. (2011). Abolishing the uniform guidelines: Be careful what you wish for. *Industrial and Organizational Psychology, 4*, 526–533.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65*, 373–406.

Peters, L. H., Fisher, C. D., & O'Connor, E. J. (1982). The moderating effect of situational control of performance variance on the relationships between individual differences and performance. *Personnel Psychology, 35*, 609–621.

Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: Contemporary practice and theory* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Podsakoff, P. M., MacKenzie, S. B., Ahearne, M., & Bommer, W. H. (1995). Searching for a needle in a haystack: Trying to identify the illusive moderators of leadership behaviors. *Journal of Management, 21*, 422–470.

Putka, D. J., & Hoffman, B. J. (2015). The "reliability" of job performance ratings equals 0.52. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 247–275). New York, NY: Routledge.

Putka, D. J., Hoffman, B. J., & Carter, N. T. (2014). Correcting the correction: When individual raters offer distinct but valid perspectives. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*, 543–548.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology, 93*, 959–981.

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology*, *76*, 432–446.

Rosenthal, R. (1984). *Meta-analysis procedures for social research*. Beverly Hills, CA: Sage.

Russell, C. J. (2001). A longitudinal study of top-level executive performance. *Journal of Applied Psychology, 86*, 560–573.

Sackett, P. R. (2014). When and why correcting validity coefficients for interrater reliability makes sense. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*, 501–506.

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. *Personnel Psychology, 55*, 807–825.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003). Criterion validity of general mental ability measures for different occupations in the European community. *Journal of Applied Psychology, 88*, 1068–1081.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529–540.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Rothstein-Hirsh, H. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38*, 697–798.

Schmidt, F. L., Hunter, J. E., & Raju, N. S. (1988). Validity generalization and situational specificity: A second look at the 75% rule and Fisher's z Transformation. *Journal of Applied Psychology, 73*, 665–672.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61*, 473–485.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. (1993). Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78*, 3–12.

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

Seymour, R. T. (1988). Why plaintiffs' counsel challenge tests, and how they can successfully challenge the theory of "validity generalization." *Journal of Vocational Behavior, 33*, 331–364.

Smith, P. C. (1976). Behavior, results, and organizational effectiveness: The problem of the criteria. In. M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 745–775). Chicago, IL: Rand-McNally College Publishing.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author. Reprinted with permission.

Spector, P. E., & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology, 72*, 3–9.

Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology, 87*, 96–111.

Stewart, G. L. (1996). Reward structure as a moderator of the relationship between extraversion and sales performance. *Journal of Applied Psychology, 81*, 619–627.

Stewart, G. L., & Carson, K. P. (1995). Personality dimensions and domains of service performance: A field investigation. *Journal of Business and Psychology, 9*, 365–378.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*, 500–517.

Thomas, A., & Raju, N. S. (2004). An evaluation of James et al.'s (1992) VG estimation procedure when artifacts and true validity are correlated. *International Journal of Selection and Assessment, 12*, 299–311.

U.S. Census Bureau. (2012). *Latest SUSB annual data: U.S. & states, totals* [Data file]. Retrieved from http://www.census.gov/econ/susb/

Vecchio, R. P. (1987). Situational leadership theory: An examination of a prescriptive theory. *Journal of Applied Psychology, 72*, 444–451.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–574.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108–131.

Viswesvaran, C., Ones, D. S., Schmidt, F. L., Le, H., & Oh, I-S. (2014). Measurement error obfuscates scientific knowledge: Path to cumulative knowledge requires corrections for unreliability and psychometric meta-analyses. *Industrial and Organizational Psychology, 7*, 507–518.

Vroom, V. H. (1973). A new look at managerial decision making. *Organizational Dynamics, 1*, 66–80.

Wallace, S. R. (1965). Criteria for what? *American Psychologist, 20*, 411–417.

Wright, J. C., & Mischel, W. (1987). A conditional approach to dispositional constructs: The local predictability of social behavior. *Journal of Personality and Social Psychology, 53*, 1159–1177.

Womer, F. B. (1968). *Basic concepts in testing.* Boston, MA: Houghton Mifflin.