Industrial and Organizational Psychology, 10(3), pp 421–456 September 2017.
 doi:10.1017/iop.2017.26

Focal Article

Meta-Analysis and the Myth of Generalizability

Robert P. Tett and Nathan A. Hundley
University of Tulsa

Neil D. Christiansen
Central Michigan University

Rejecting situational specificity (SS) in meta-analysis requires assuming that residual variance in observed correlations is due to uncorrected artifacts (e.g., calculation errors). To test that assumption, 741 aggregations from 24 meta-analytic articles representing seven industrial and organizational (I-O) psychology domains (e.g., cognitive ability, job interviews) were coded for moderator subgroup specificity. In support of SS, increasing subgroup specificity yields lower mean residual variance per domain, averaging a 73.1% drop. Precision in mean rho (i.e., low SD(rho)) adequate to permit generalizability is typically reached at SS levels high enough to challenge generalizability inferences (hence, the "myth of generalizability"). Further, and somewhat paradoxically, decreasing K *with increasing precision undermines certainty in mean* r *and Var(*r*) as meta-analytic starting points. In support of the noted concerns, only 4.6% of the 741 aggregations met defensibly rigorous generalizability standards. Four key questions guiding generalizability inferences are identified in advancing meta-analysis as a knowledge source.*

Keywords: meta-analysis, validity generalization, situational specificity, moderator subgroup, quantitative literature review

There is little doubt that meta-analysis has greatly benefited the science of work behavior. The major take-home message from the earliest applications in industrial and organizational (I-O) psychology (Schmidt, Gast-Rosenberg, & Hunter, 1980; Schmidt & Hunter, 1977, 1978, 1984; Schmidt, Hunter, & Caplan, 1981; Schmidt, Hunter, & Pearlman, 1981; Schmidt, Hunter, Pearlman, & Shane, 1979; Schmidt, Ocasio, Hillery, & Hunter, 1985) was that, of all the variability observed across studies in validity estimates

Robert P. Tett, University of Tulsa; Nathan Hundley, University of Tulsa; Neil D. Christiansen, Central Michigan University.

The authors gratefully acknowledge the following individuals for their helpful feedback on earlier drafts: David Fisher, Fred Oswald, Mitch Rothstein, Paul Sackett, Piers Steel, and Frances Wen. No endorsement of this work, in whole or in part, is implied.

Correspondence concerning this article should be addressed to Robert P. Tett, Department of Psychology, 800 S. Tucker Ave., University of Tulsa, Tulsa, OK 74104. E-mail: robert-tett@utulsa.edu

for cognitive ability tests, the lion's share is due to sampling error and other artifacts. Thus, the doctrine of *situational specificity* gave way to one of *validity generalization*, which diminished the need (and value) of local validation and strengthened the credibility of research findings through aggregation. Given the celebrated gains of validity generalization in the realm of ability testing, it is understandable that mean validity is the most cited meta-analytic output (Carlson & Ji, 2011).

Past debates on generalizability inferences in meta-analysis (e.g., James, Demaree, & Mulaik, 1986; Sackett, Tenopyr, Schmitt, Kehoe, & Zedeck, 1985; Schmidt, Hunter, Pearlman, & Rothstein-Hirsh, 1985) tended to resolve toward rejection of situational specificity (in support of generalizability), albeit with unsettled questions. More recent perspectives (e.g., James & McIntyre, 2010; Murphy, 2003; Steel & Kammeyer-Mueller, 2008) echo the challenges of generalizing meta-analytic findings. Sackett (2003) states:

> Validity generalization is still wrongly viewed by many, not as a theory about the process of drawing inferences from cumulative data, but as a general statement that the bulk of the variability in research findings is due to statistical artifacts. (p. 111)

Review of meta-analytic findings published over the past 35 years confirms the need to revisit this issue as, in many cases, substantial variance in validity estimates remains after accounting for artifacts, rendering the strong focus on mean effect sizes potentially tenuous if not misguided. It is our hope that renewed discussion of generalizability will promote the value of meta-analysis as a knowledge-generating framework.

Our specific aims are to (a) raise awareness of the potential confusion arising from use of the terms "validity generalization" and "situational specificity" in interpreting meta-analytic results; (b) emphasize generalizability of mean rho as the key meta-analytic inference; (c) show, by review of 24 published meta-analytic studies, how generalizability inferences tend to be treated in I-O psychology, as well as the complex effects of increasing moderator specificity on those inferences; and (d) offer recommendations regarding how meta-analytic findings are reported and how such findings are interpreted with respect to generalizability. In the process, we (e) provide evidence that substantial residual variance is attributable to moderators, in support of situational specificity; (f) show how generalizability of mean rho is achieved with adequate precision typically under highly specified moderator conditions (ergo, "the myth of generalizability"); and (g) identify and explain a further paradox of low-K specificity, in which increasing conditional specificity serving precision in generalizability is undermined by corresponding increases in second-order sampling error in deriving mean r and Var(r) as analytic starting points. We begin with an overview of

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

Hunter–Schmidt (HS) meta-analytic methods, those most often used in I-O psychology.

How Meta-Analysis Works[1]

HS meta-analysis takes a distribution of observed validity estimates (r_{xy}s), drawn from one or more populations of such estimates, and manipulates the resulting sample-weighted mean r and Var(r) to generate two important outputs: (a) mean rho and (b) Var(rho). Mean rho is derived as mean r corrected for systematic attenuation from correctible artifacts (r_{xx}, r_{yy}, range restriction, dichotomization[2]), and, quite independently, Var(rho) is derived by subtracting Var(e), composed of sampling and other artifact error variances, from Var(r) and then adjusting this residual variance (Var(res)) upward to account for differential artifact effects. As Var(e) approaches Var(r), Var(res) decreases, leaving little or no room for moderators to operate. A low Var(rho), following the upward adjustment, favors the inference that mean rho is generalizable. That is, a user may safely assume that mean rho (and mean r) applies to his or her situation because the relationship is claimed to hold across all conditions represented in the aggregation.

Figure 1 depicts the major meta-analytic elements, processes, and variance sources. Notably, there are two levels of variance in meta-analysis that can affect generalizability inferences: (Level 1) Var(r) due to (a) study-level (first-order) sampling error, (b) differential effects of other correctable artifacts, (c) differential effects of uncorrectable artifacts (e.g., human error in calculating r_{xy}), and (d) differences in moderator standing; and (Level 2) second-order variance due to (a) second-order sampling error, in turn a function of K and N per r_{xy}, and (b) bias in generating mean r and Var(r) as the starting points for the first-level derivations. We address first-level processes before returning to consider how second-level sampling of r_{xy}s (further) affects generalizability of meta-analytic findings. For now, let us assume that mean r and Var(r) are derived from large-K aggregation (i.e., $K > 100$) of r_{xy}s, each based on relatively large N (min = 150).[3]

In addition to mean rho and Var(rho), several related indices are produced as aids in judging generalizability. The square root of Var(rho),

[1] Meta-analytic methods are described in detail by Hunter and Schmidt (2004) and others (e.g., Card, 2012; Hedges & Olkin, 2005; Lipsey & Wilson, 2001). Here, we focus on aggregation of r_{xy}; our main points likely apply to d as well. We focus on HS procedures over others (e.g., Raju, Burke, Normand, & Langlois, 1991) due to their prominence in I-O psychology.

[2] Dichotomization is best corrected at the study level prior to deriving mean r (Hunter & Schmidt, 2004).

[3] These numbers are not intended as uniquely defensible standards for meta-analysis but rather to support an assumption of low second-order sampling error to promote initial examination of first-level processes. For the noted K and N, the maximum 95% confidence interval width, assuming homogeneity (see Whitener, 1990), is ± .016 around mean r.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

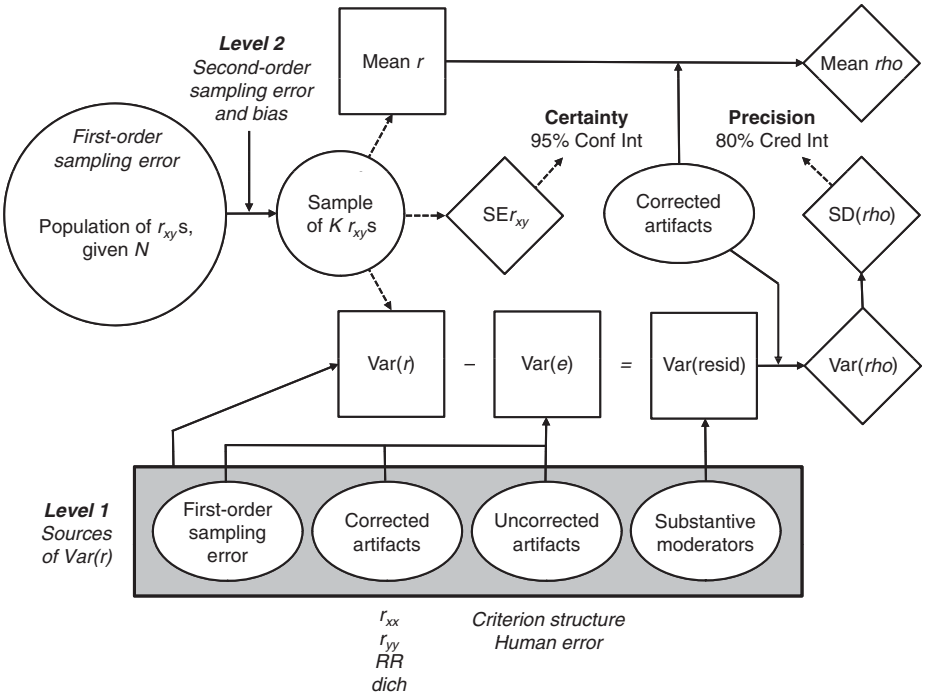

Figure 1. Meta-analysis in a nutshell.

SD(rho), is often used to create a credibility interval around mean rho. A credibility interval does not reflect variability due to sampling error and corrected artifacts because such variability has been removed in deriving SD(rho). It informs as to the likely presence or absence of moderators of rho, which speaks directly to the question of generalizability (Schmidt & Hunter, 1977; Whitener, 1990). The most commonly reported is the 80% credibility interval (80% CI), calculated as mean rho $\pm$ 1.28 $\times$ SD(rho). The 80% CI specifies the range within which 80% of population validities are expected to fall. The lower boundary of the 80% CI is the 90% credibility value (90% CV), the point below which 10% of population validities are expected to fall.

As a further marker for the possible operation of moderators, "%VE" is often derived as the percentage of Var(r) due to Var(e). As Var(e) approaches or exceeds Var(r), %VE approaches or exceeds 100%, and mean validity is inferred as situationally generalizable. If Var(e) $>$ Var(r) (i.e., observed r_{xy}s are less variable than expected due to artifacts), %VE is often truncated at 100%. A cutoff of 75% is frequently invoked in judging generalizability. The "75% rule" (Schmidt & Hunter, 1977) holds that, if %VE $>$ 75%, situational generalizability of mean rho may be inferred, as the remainder of Var(r) can be attributed to uncorrected artifacts.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

Selected moderators are tested as either subgroup differences or study-level correlations. The former are far more common. Hunter and Schmidt (2004) offer a nonstatistical test based on an analysis of variance model. Subgroup moderators are supported to the degree subgroups of r_{xy}s yield (a) different mean rhos and (b) SD(rho)s that average lower than SD(rho) for the combined (i.e., broader) aggregation. This latter marker for moderation—shrinking SD(rho) with increasing moderator specificity—is key to evaluating generalizability in our survey of I-O meta-analyses, to which we return later on.

Validity Generalization and Situational Specificity

The major aim of most *r*-focused meta-analyses is to discover the true linear relationship, rho, between two variables by correcting mean *r* for systematic attenuation due to measurement artifacts. A further aim is to identify the consistency of rho across conditions after filtering out noise in estimation due to sampling error and cross-study differences in one or more correctable artifacts. These two aims are combined in generalization: mean rho is often a major subject of generalization, and moderators specify the conditions to which mean rho can be generalized.

Two key concepts relating to generalizability in meta-analysis are validity generalization (VG) and situational specificity (SS). Taken at face value, VG and SS seem like polar opposites: If one has a validity estimate that generalizes across all conditions (e.g., jobs, demographics, measurement formats), this would appear to imply simultaneously both VG and a lack of SS. Conversely, if one observes substantial SS in population validity, it is not obvious how validity can be understood as generalizable.

Despite the strong (negative) surface connection between VG and SS, they are formally distinct concepts in HS meta-analysis (e.g., Schmidt & Hunter, 1977): VG is inferred when a lower bound estimate of population validity exceeds some minimal value (most often 0), and SS is inferred when Var(*r*) is not completely attributable to sampling error and other artifacts. Thus, one may observe both VG and SS, as when population validity is found to be at least .10 in the large majority of applications and yet also varies, even substantially, as a possible function of one or more moderators (whether or not they are identified and/or tested).

Most HS meta-analysts are well aware of the VG/SS distinction and report results separately for those two inferences. There is, nonetheless, opportunity for confusion among consumers of meta-analytic findings, as noted by Murphy (2000). Perhaps reflecting some of that confusion, Carlson and Ji (2011) reported that, out of 1,489 citations of meta-analytic findings in top I-O psychology outlets, only 7 (.47%) noted variability in effect sizes, far underrepresenting cases where SD(rho) > 0. Part of the problem may

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

be that meta-analysts themselves tend to downplay or ignore their own SS results.[4] We argue that, contrary to the way it tends to be treated in many meta-analyses and as cited by others, SS is more critical than VG in drawing meta-analytic inferences.

To be clear, both VG and SS inferences are useful. Consider a case where mean rho is .25 and the 80% CI ranges from .10 to .40 (based on SD(rho) = .117). It is helpful to know that population validity can be expected to be at least .10 in 90% of conditions represented in the aggregation (whether overall or in a specific moderator subgroup), as the predictor may be a safe bet for inclusion in an assessment battery[5] and for citing as relevant in subsequent research. It is also useful to know, independent of the VG inference, that rho varies between .10 and .40 in 80% of cases, as researchers and consumers seek to estimate what validity is likely to be in a given situation, and the .30-unit width of the interval precludes precision in that judgment. The VG inference, despite its merit, suffers from several limitations relative to the SS inference, as follows.

First, VG dichotomizes the continuum of correlation strength (James & McIntyre, 2010): In terms of the "90% > 0" rule, either the 90% CV falls above 0, conferring VG, or it does not, failing to confer VG. As occurs in other types of inquiry (e.g., correlation), dichotomization entails a loss of information (e.g., degrading a normally distributed test score to pass/fail). The VG inference offers a basic (and easily communicated) on/off heuristic regarding a specific lower-bound value, but the SS inference, represented by the 80% CI, more fully captures generalizability in terms of the full correlation continuum.[6]

Second, whether or not unidirectional (positive or negative) validity is generalizable is irrelevant in some applications of meta-analysis. Consider the general personality factor of Agreeableness (A) in relation with job performance. Good reasons can be offered for expecting a positive r_{xy} in jobs where caring for others (i.e., high A) is especially valued and a negative r_{xy} where being tough skinned (i.e., low A) is favored (cf., Tett & Christiansen, 2007; Tett, Jackson, Rothstein, & Reddon, 1999). Meta-analytic means for A in predicting work criteria under the assumption of unidirectionality (cf. Tett et al., 1999) are often close to zero, with 80% CIs extending in both directions (e.g., Barrick & Mount, 1991; Hough, Ones, & Viswesvaran, 1998; Judge, Bono, Ilies, & Gerhardt, 2002; Vinchur, Schippmann, Switzer, & Roth, 1998). In such cases, the 90% CV rule is moot and mean rho is especially

[4] We offer evidence for this in our brief survey of I-O meta-analyses, presented below.

[5] Validity expected in actual applications should be centered on uncorrected mean *r*, not mean rho.

[6] Half the 80% CI is identified by the lower bound 90% CV. The 80% CI is, thus, less a replacement than an extension of the 90% CV.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

diminished as a focus of generalization. The 80% CI, in contrast, is relevant in all cases regardless of directionality and, where bidirectionality is evident, drives pursuit of directional moderators.

Third, the VG inference sets a low standard for generalizability (Murphy, 2003). In terms of our earlier example, finding that population validity $> .10$ more than 90% of the time is helpful, but one cannot infer from this that rho $= .25$ is generalizable. Concluding that mean validity "generalizes" would be a misrepresentation of the data, given that 80% of rhos are estimated to fall between .10 and .40 (and 10% below .10 and another 10% above .40).

Finally, the VG inference focuses attention on the weaker end of the rho distribution. The stronger end is important because it shows validity under potentially favorable conditions. That rho in our example is greater than .40 in 10% of conditions represented in the aggregation should encourage efforts to identify what those conditions are so we might improve predictions and explanations of targeted criteria. Most meta-analysts are keen to identify moderators, and many such efforts are supported. Credibility intervals within moderator subgroups, however, are still often wide enough to compromise meaningful generalizability of mean rho.

In sum, meta-analysis offers uniquely powerful estimates of the strength and direction of relationship between targeted variables. Finding that 90% of rhos fall above zero (or on one side of it) is useful when otherwise so much noise in findings across studies undermines certainty affording practical and theoretic advance. Generalizing validity, however, means more than where the 90% CV falls in relation to zero; its broader meaning entails stability in rho across conditions, as conferred by how close SD(rho) is to 0, yielding a narrow credibility interval, no matter whether (mean) rho is strong or weak (Carlson & Ji, 2011).

A high SD(rho) limits generalizability by the possibility that substantive moderators may be responsible for some large proportion of residual variance. This feature of generalizability, quantified by proximity of SD(rho) to 0, may be denoted as *precision*.[7] Low precision (i.e., high SD(rho)) does not render mean rho uninterpretable. It does, however, render the mean less useful in application to particular settings because *it then is not clear which rho is the one most applicable to a given situation*. Low precision impedes generalizability, favoring SS albeit not guaranteeing it.

[7] Nunnally (1978) and Cortina (1993) discuss "precision" in terms of variability of inter-item *r*s in estimating internal consistency reliability. McDaniel, Rothstein, and Whetzel (2006), citing Sterne and Egger (2005), use it in discussing funnel plots in trim-and-fill methods of detecting publication bias. Its meaning here is similar to those other uses, but close mathematical parallels to either should not be inferred.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

A Fly in the Ointment

SD(rho) > 0 indicates SS only to the degree that all of Var(*e*) is subtracted from Var(*r*). Unfortunately, some sources of Var(*e*) are difficult to estimate: variance due to differences in uncorrectable (or hard-to-correct) artifacts (e.g., criterion factor structure, errors in calculating and communicating r_{xy}). Arguments for rejecting SS (e.g., Schmidt & Hunter, 1977) emphasize that SD(rho) may be > 0 not due to moderators but due to unclaimed Var(*e*). This is the basis for the 75% rule and is an important assumption because it confers rejection of SS despite SD(rho) > 0.

By their nature, these artifact variances are difficult to estimate directly. An indirect test of the composition of SD(rho) is offered, however, by examining the effects of moderator specificity on SD(rho) within and across multiple meta-analytic studies. If residual Var(*r*) is due solely to unclaimed error, then SD(rho) should show no reduction with increasing specificity in moderator subgrouping. Conversely, as per Hunter and Schmidt's (2004) second criterion for inferring moderation, if SD(rho) drops with increasing specificity, then this would offer support for the effects of moderators. This question is pursued in a later section.

Second-Level Variance

Our main points up to now apply to the Level 1 variance sources captured by Var(*r*) in deriving Var(rho) and SD(rho). Level 2 warrants consideration because it addresses the robustness of mean *r* and Var(*r*) as starting points for first-level calculations. Two major sources of second-level error are (a) random error in sampling r_{xy}s from the population of r_{xy}s and (b) systematic sampling bias. As to the former, when *K* is small and/or *N* per input study is small, mean *r* is uncertain (as per SE(r_{xy}) and the 95% confidence interval around mean *r*) and so mean rho is uncertain; small *K* and small *N* also make Var(*r*) uncertain and so also Var(rho) and SD(rho).

Regarding sampling bias, mean *r* and Var(*r*) are products of both the meta-analyst's choice of r_{xy} estimates to include in the aggregation and the availability of r_{xy}s. The "bias" in r_{xy} choice is generally not problematic, with the understanding that such choices largely determine the actual population of conditions to which generalizations can be made. Results based on r_{xy}s from only civilian settings, for example, are of questionable generalizability to military settings (whether or not military findings were deliberately excluded).

Other biases in selecting r_{xy} can be more problematic. Publication bias, "the possibility that not all completed studies on a topic are published in the literature and that these studies are systematically different from published studies" (McDaniel, Rothstein, & Whetzel, 2006, p. 927), has been a growing

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

target of research. It has traditionally and most often been pursued as "fail-safe *N*" in the context of "the file-drawer problem" (e.g., Rosenthal, 1979), but several new approaches have been taken more recently (e.g., Coburn & Vevea, 2015). With few major exceptions (e.g., Dalton, Aguinis, Dalton, Bosco, & Pierce, 2012), evidence suggests that publication bias is a problem in meta-analysis (Hedges & Vevea, 1996; Kepes, Banks, McDaniel, & Whetzel, 2012; Kromrey & Rendina-Gobioff, 2006; van Assen, van Aert, & Wicherts, 2015), albeit less so under some conditions than others (Coburn & Vevea, 2015; Ferguson & Brannick, 2012; McDaniel et al., 2006). Besides editorial screening, a similar bias occurs when researchers use a "Texas shooter" strategy to identify from an array of available results just those judged worthy of reporting (Biemann, 2013).

To the degree such biases operate, as a whole and differentially across domains, mean r and Var(r) become uncertain as starting points for the main analyses from which mean rho and Var(rho) are derived. For example, if stronger r_{xy}s are selected for publication (by "Texas shooters" and/or reviewers and editors), mean rho will be overestimated. SD(rho), in addition, will be underestimated, and so also inferences of SS. As methods for detecting selection biases and evidence regarding the extent of the problem continue to be developed, the importance and nature of such biases will better inform meta-analytic generalizations going forward.

To recap, there are two broad levels of variance affecting generalizability inferences in meta-analysis: Level 1 variance sources include first-order sampling error, other artifact error (claimed and unclaimed), and moderators. Level 2 variance sources include second-order sampling error and sampling biases in deriving mean r and Var(r) as starting points for Level 1 procedures. Two key meta-analytic outcomes bearing on generalizability at the two levels are SD(rho) and SE(r_{xy}), respectively. SD(rho) (Level 1) speaks to precision in generalizing mean rho (lower SD(rho) = higher precision), whereas SE(r_{xy}) (Level 2) speaks to certainty in mean r and Var(r) as foundations for generalizability inferences (lower SE(r_{xy}) = higher certainty).

A Brief Survey of Meta-Analyses in I-O Psychology

To get a sense for how issues of generalizability are dealt with in meta-analysis applied in the field of I-O psychology, we surveyed 24 such studies published in four predictor content domains (cognitive ability, personality, work attitudes, and leadership) and three method domains (job interview, assessment centers, and situational judgment tests). Our specific aims in conducting the survey were to (a) assess the effects of increasing aggregation specificity on SD(rho) and SE(r_{xy}) within and across I-O research domains and (b) estimate normative practices regarding emphasis on VG and SS inferences.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

Key to understanding generalizability inferences in meta-analysis is the effect of increasingly specified conditions represented in a given aggregation. This is broadly referred to as situational specificity as discussed above and includes a variety of moderator classes: (a) Main situational moderators include general setting (e.g., civilian vs. military, public vs. private sector), job family (e.g., professional vs. nonprofessional), specific jobs nested within a family (e.g., computer programmer), job complexity, and so on. Other moderator classes include (b) predictor constructs (e.g., the five-factor model of personality and nested facet traits; general mental ability (GMA) and its facets), (c) criterion constructs (e.g., personnel data vs. job performance and nested task vs. contextual facets), (d) general methods (e.g., situational vs. behavioral interviews; assessment center overall ratings and nested exercise ratings), (e) measurement features (e.g., follower- vs. leader-rated LMX, nine-item vs. 15-item OCQ), (f) research design and purpose (e.g., concurrent vs. predictive, research vs. administrative), (g) demographics (e.g., age, gender), and (h) sources (e.g., peer-reviewed articles vs. conference papers). Each prospective moderator offers opportunity for greater specification of the conditions affecting the generalizability of the given mean rho.[8]

Building on Hunter and Schmidt's (2004) analysis-of-variance approach to identifying moderators, the expectation favoring SS is that SD(rho) will shrink with increasing moderator specificity. This is because group mean differences should account for a portion of overall variance, leaving smaller mean Var(*r*) and so also mean Var(rho) and mean SD(rho). We tracked this effect by comparing SD(rho) across meta-analytic aggregations differentiated in terms of "specificity level," operationalized as follows.

First, using moderators reported in the 24 source articles as input, we created a multilevel taxonomy of sequentially nested moderator subgroups per moderator class. This taxonomy is presented in Table 1. Note that nesting level (e.g., 0 to 3) is shown at the top of each moderator class. Aggregations typically combine moderators from multiple classes (e.g., predictor and criterion facets) and often from multiple subcategories within a given class. Specificity level per aggregation was indexed by adding the values (i.e., 0 to 3) for each moderator subgroup involved in that aggregation. The scheme yields specificity level 1 when aggregating, for example, GMA validity in predicting all work criteria. Specificity values increase with each narrowing of conditions represented in the aggregation. For clarification, consider the following examples.

[8] All but the last moderator class ("sources") can be understood broadly as "situational" by defining the conditions of generalization (e.g., targeting predictor X and criterion Y using method A in setting B). "Sources," which speaks more to publication bias, is involved in only 22 of the 741 aggregations (3%). Excluding it from the specificity analyses would have a trivial impact on current findings.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

Table 1. Seven Classes of Subgroup Moderators with Specificity Levels

Criterion constructs and methods	Predictor constructs and methods	
0 1 2 3	0 1 2 3	0 1 2 3
All work-related criteria	All predictor constructs/methods	
Job performance/proficiency (general)	*Predictor construct*	*Predictor method*
Task performance	Cognitive ability/GMA	Job interview
Task performance facets (e.g., analyzing)	GMA facets (e.g., verbal reasoning)	Situational
Contextual performance/OCB	Personality	Job related
Contextual performance facets (e.g., supporting)	Extraversion	Psychological
Adaptive performance	E facets (e.g., sociability)	Structured
Managerial/leadership performance	Conscientiousness	Unstructured
M/L performance facets (e.g., decision making)	C facets (e.g., order)	Individual
Rated performance (subjective)	Agreeableness	Board
Other than rated performance (objective)	A facets (e.g., nurturance)	Averaging
Counterproductive work behavior	Emotional stability/neuroticism	Consensus
CWB facets (e.g. theft)	ES/N facets (e.g., anxiety)	Cognitive ability test available
Training performance/proficiency	Openness to experience	Cognitive ability test unavailable

Table 1. Continued

Criterion constructs and methods	Predictor constructs and methods	
Personnel data/career advancement	O facets (e.g., creative thinking)	Cognitive ability test availability unknown
Productivity	Masculinity–femininity	Assessment center (OAR)
Withdrawal	Competency	Particular exercises (e.g., in-basket)
Withdrawal cognitions	Management/leadership	Situational judgment test
Turnover intention	M/L facet (e.g., leading/influencing)	Knowledge based
Turnover	LMX	Behavior/tendency based
Tenure	Drive	Constant content (SJT-DV)
Absenteeism	Job knowledge	Personality composite
Lateness	Interpersonal skill	Heterogeneous composite
Status/level/promotions	Teamwork	
Salary	Attitudes	
Safety/accidents	Job satisfaction	
Safe/unsafe behavior	Organizational commitment	
Safe	Safety climate	
Unsafe	Other	
Driving	e.g., Role clarity, trust	
Non-driving		
Accidents		
Subjective		
Objective		
External		
Management potential		

Table 1. Continued

Situational features	Meaurement features	Research design features	Sources
0 1 2	0 1	0 1 2 3	0 1 2 3
Job type/family	Rating source	Main design	All sources
Professional	Self-rated	Cross sectional	Articles
Scientists/engineers	Other rated	Longitudinal	Top tier
Teachers	Follower rated	Exploratory/confirmatory	Other ranked
Nurses	Leader rated	Exploratory	Unranked
Nonprofessional	IV/DV shared/common	Confirmatory	Theses/dissertations
Laborers	IV/DV independent/noncommon	Job analysis	Presentations
Management/leadership	Measurement length	Formal job analysis	Book chapters
Low	Multi-item	Armchair job analysis	Unpublished
Middle	Single item	No job analysis	Technical reports
High	Instrument	Unknown	Government
Nonmanagerial	LMX-7	Validation strategy	Proprietary
Police	LMX-MDM	Concurrent/incumbents	Test vendor manuals
Sales	LMX-Other	Predictive/recruits	
Skilled or semi-skilled	9-item OCQ	With feedback	
Computing and account recording clerical	15-item OCQ	Without feedback	
Steno typing and filing clerical	Global job satisfaction	Experimental	

Table 1. Continued

Situational features	Meaurement features	Research design features	Demographics
Miscellaneous/mixed	Facet job satisfaction	Purpose	0 1
Male dominated	CA-WPT	Research	Age
Female dominated	CA-Other	Administrative	Adult
Gender neutral	CWB-Bennett and Robinson	Promotion	Young adult
General setting	CWB-Other	Early identification	Sex
Civilian		Selection	Male
Military			Female
Business			
College			
Public sector			
Private sector			
Collectivistic culture			
Individualistic culture			
Job complexity			
High			
Medium			
Low			
Observer rated			
Objectively rated (e.g., ONET)			

Barrick and Mount (1991) aggregated validity estimates by general personality factor crossed with job family or criterion type. Their meta-analysis for Conscientiousness in predicting assorted work outcomes of professionals, for example, was coded as specificity Level 3, given that Conscientiousness falls at Level 2 in the predictor taxonomy (nested within the personality domain) and professionals fall at Level 1 of job type/family within the situational taxonomy. Greater specificity is evident in aggregations reported by Martin, Guillaume, Thomas, Lee, and Epitropaki (2016) on LMX in relations with different types of job performance (e.g., task vs. contextual) under various conditions (e.g., leader- vs. follower-rated LMX, common vs. noncommon rating sources, three different LMX measures) crossed as far as permitted by available input sources. Specificity levels in Martin et al. (2016) range from 5 (e.g., LMX = 3, under predictors; task performance = 2, under criteria) to 9 (involving sequentially nested combinations of several predictor, criterion, and measurement moderators).

Overall, specificity values range from 1 to 9 in the 741 aggregations reported in the 24 studies. Clearly, this method of assessing conditional specificity leaves room for improvement. A major limitation is the untested assumption of equal intervals within and across classes. For example, it is not clear that specificity gained moving from Extraversion to one of its facets (e.g., sociability) is equal to specificity gained moving from GMA to, say, verbal ability or moving from unpublished sources to technical reports. The coding also ignores the possibility of interactions among moderators in their effects on SD(rho). For example, the impact of increased specificity moving from Extraversion to sociability might vary by criterion type. Notwithstanding these limitations, the proposed system permits initial evaluation of moderator specificity in relations with SD(rho) and other meta-analytic outputs.

Study Selection

Hundreds of meta-analyses have been published in I-O psychology since the method was introduced in the late 1970s. Our selection of studies within each of the seven domains was limited to applications of standard HS procedures targeting relationships at the individual level of analysis (e.g., excluding team- and organization-level studies). We further targeted (mostly) influential articles (e.g., from *Journal of Applied Psychology*, *Personnel Psychology*, *Psychological Bulletin*) and sought to assemble sets of studies within each domain that collectively covered a variety of criteria and moderators. Finally, where possible, we sought representation across the decades. Critically for current aims, our choice of studies, although not random, was blind to meta-analytic treatment of VG and SS. We expect that our sample is fairly representative of peer-reviewed, individual-level meta-analyses in I-O psychology regarding generalizability estimates and inferences.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

Coding

Each study reported results for multiple aggregations. For each aggregation, we recorded (a) the predictor and criterion and classified those variables by general domain (e.g., ability, personality, job performance, withdrawal); (b) main meta-analytic output, including *K*, total *N*, mean *r*, mean rho, SD(rho) and artifacts used to derive it,[9] %VE, and 80% CI; missing data were calculated directly, if possible, from available data (e.g., 80% CI = mean rho ± 1.28 × SD(rho)); given our focus on SD(rho), aggregations lacking SD(rho) and ways to derive it were dropped; and (c) moderator subgroup combinations from which specificity level could be derived (described above). We further noted, per article, (d) several other features relevant to generalizability inferences, including use of the 90% CV rule, 75% rule, *Q* (a chi-square test for homogeneity; Cochran, 1954), degree of emphasis placed on SS in light of SD(rho) > 0, and entry of mean rhos into tertiary analyses (e.g., multiple regression, path analysis).[10]

Regarding %VE, some authors reported a maximum of 100% when Var(*e*) exceeded Var(*r*). To avoid underestimating mean %VE (Schmidt et al., 1985), we calculated it directly where possible and otherwise estimated it as the reciprocal of the mean of reciprocals (Barrick & Mount, 1991) from cases where %VE > 100%. A total of 35 missing values were estimated as the overall mean (i.e., 151%) derived from 27 cases with %VE > 100%.[11] SE(r_{xy}) was calculated using the reported mean *r*, total *N*, and *K*. We used the homogeneity case equation (Whitener, 1990) to avoid directly confounding SE(r_{xy}) with SD(rho). Reliance on the heterogeneity case equation would increase SE, strengthening later arguments regarding limits to generalizability. Assuming homogeneity thus renders current arguments regarding SE(r_{xy}) conservative. For degree of emphasis on SS, we coded each article as either 0 = SS completely ignored; 1 = SS noted in passing (e.g., in defining the 80% CI), but otherwise ignored; 2 = SS discussed briefly (e.g., two or three sentences); 3 = SS discussed in moderate detail (e.g., in interpreting observed SD(rho) > 0); or 4 = SS emphasized as limiting the generalizability of mean rho.

Results

Table 2 summarizes selected features of the 24 studies included in the review. Results for the entire set include the following: (a) Eight studies (33%) used

[9] Our aim in coding psychometric artifacts corrected in deriving SD(rho) was to examine and possibly adjust for effects of those corrections on SD(rho) across cases. However, point-biserial *r*s between correction for each artifact (0/1) and SD(rho) are weak overall and present no discernable pattern of effects across or within domains. Accordingly, we did not control for differences in reliance on artifacts in deriving SD(rho).

[10] The relevance of tertiary analyses is discussed later on.

[11] We considered using means based on available values from the same domain, but insufficient cases permitted this.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

Table 2. Summary of 24 Meta-Analytic Studies in I-O Psychology

Area/domain/article	Journal	Main variables: Predictors	Main variables: Criteria	N MAs	Spec range[b]	Mean SD(rho)	90% rule	75% rule	90% CV	80% CI	Q	Tertiary analysis	SS emph.[c]
Content													
Cognitive ability													
Schmidt et al. (1979)	PPsych	GMA, facets	Job performance	14	4–5	.22	Yes	Yes	Yes	No	No	No	2
Rothstein & McDaniel (1992)	JBP	Cognitive ability, facets	Job performance	72	4–7	.07	No	No	No	No	No	No	2
Salgado et al. (2003)	PPsych	GMA, facets	Job perf, training success	12	2–3	.17	No	No	Yes	No	No	No	2
Gonzalez-Mulé et al. (2014)	JAP	GMA	CWB, OCB	33	2–4	.17	Yes	No	No	Yes	No	Yes	2
Personality													
Barrick & Mount (1991)	PPsych	FFM	Assorted criteria	65	3–4	.09	No	Yes	Yes	No	No	No	1
Bartram (2005)	JAP	Great 8 pers facets, ability	Great 8 perf dimensions	8	4–6	.11	No	Yes	Yes	No	No	No	2
Zimmerman (2008)	PPsych	FFM, job complexity	TOI, turnover	14	1–5	.08	Yes	Yes	No	Yes	No	Yes	2
Beus et al. (2015)	JAP	FFM	Safety behaviors, accidents	41	4–5	.07	No	No	No	Yes	No	Yes	2
Work attitudes													
Tett & Meyer (1993)	PPsych	Job sat, org commitment	TOI/WC, turnover	39	4–7	.11	Yes	Yes	No	Yes	Yes	Yes	2
Judge et al. (2001)	Psych Bull	Job sat	Job performance	29	3–5	.20	Yes	No	No	Yes	Yes	No	2
Judge et al. (2002)	JAP	FFM	Job satisfaction	5	4	.18	Yes	No	No	Yes	No	Yes	2
Choi et al. (2015)	JAP	FFM	Org commitment facets	20	4–5	.12	Yes	No	No	Yes	No	Yes	3

Table 2. Continued

Area/domain/article	Journal	Main variables: Predictors	Main variables: Criteria	N MAs	Spec range[b]	Mean SD(rho)	90% rule	75% rule	90% CV	80% CI	*Q*	Tertiary analysis	SS emph.[c]
Leadership													
Lord et al. (1986)	JAP	Personality facets	Leadership perceptions	10	4	.10	Yes	Yes	Yes	No	No	No	4
Lowe et al. (1996)	LQ	Leadership facets	Leadership effectiveness	35	5–6	.21	Yes	No	Yes	Yes	No	No	4
Judge & Piccolo (2004)	JAP	Leadership facets	Job performance	9	4–8	.26	Yes	No	No	Yes	No	Yes	2
Martin et al. (2016)	PPsych	LMX	Performance facets	133	5–9	.10	Yes	Yes	No	Yes	No	Yes	1
Methods													
Job interview													
Weisner & Cronshaw (1988)	J Occ Psych	Job interview	Work-related criteria	14	1–4	.16	No	No	No	No	No	No	3
McDaniel et al. (1994)	JAP	Job interview	Job perf, training perf	37	2–5	.17	No	No	Yes	No	No	No	2
Huffcutt et al. (2004)	IJSA	Job interview	Job performance	15	3–5	.09	No	Yes	No	Yes	No	No	2
Assessment centers													
Gaugler et al. (1987)	JAP	AC overall rating	Various work criteria	17	1–3	.12	Yes	No	Yes	No	No	No	3
Arthur et al. (2003)	PPsych	AC dimensions	Various job-related criteria	22	1–3	.11	Yes	No	Yes	No	No	Yes	3
Hoffman et al. (2015)	JAP	AC exercises	Job performance	21	3–4	.08	No	No	No	Yes	Yes	Yes	1
Situational judgment tests													
McDaniel et al. (2007)	PPsych	FFM, ability	SJT, job performance	54	2–5	.08	Yes	No	No	Yes	No	Yes	3
Christian et al. (2010)	PPsych	SJT	Job performance, facets	22	3–4	.09	No	No	No	Yes	No	No	1
			Totals	741	1–9	.12	58%	33%	38%	58%	13%	46%	M=2.2

[a]Moderators include those explicitly tested for subgroup moderation and construct specificities (e.g., FFM factors, performance facets); excluded are continuous moderators tested by moderator correlation.
[b]Specificity level range: see text for operationalization of specificity level per aggregation.
[c]SS emphasis: 0 = SS completely ignored; 1 = SS noted in passing but otherwise ignored; 2 = SS discussed briefly; 3 = SS discussed in moderate detail; 4 = SS emphasized as limiting the generalizability of mean rho.

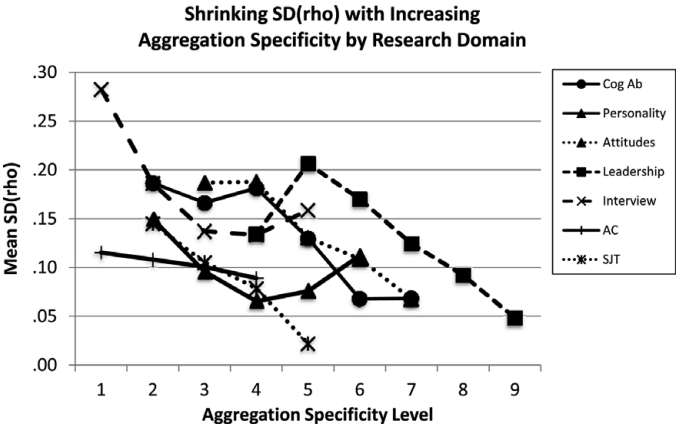

Figure 2. Shrinking SD(rho) with increasing aggregation specificity by research domain.

the 75% rule and three (13%) reported *Q* to judge the presence or absence of moderators; (b) 14 (58%) used the 90% rule in judging VG, 14 (58%) reported the 80% CI, and five (21%) used both the 90%CV and 80% CI. (c) Regardless of whether or not the 90% CV rule was invoked, 83.1% of the 741 aggregations met the 90% CV standard for VG. (d) Of the 410 cases with available %VE, 307 (74.9%) had values less than 75%. Finally, (e) regarding treatment of residual variance, four studies (17%) briefly noted its connection to moderators but ignored it when interpreting observed results, 13 (54%) briefly discussed the possibility of moderators in light of observed residual variance, five (21%) discussed the issue in moderate detail, and the remaining two (8%) emphasized that SD(rho) > 0 limits generalizability of mean rho.

Effects of moderator specificity on SD(rho), SE(r_{xy}), and %VE are shown in Table 3 for each of the seven domains and all sources combined. Several points bear noting here. First, with few exceptions, mean SD(rho) drops within each domain and overall as moderator specificity increases. Figure 2 plots this effect per domain. The far-right column of Table 3 includes three values per block. The upper value is the percentage drop in Var(rho) going from lowest to highest specified conditions. Values range from 40.5% (assessment center) to 97.7% (SJT), averaging 73.1%. Correspondingly, the pooled within-domain correlation between specificity level and SD(rho) (involving all 741 cases, controlling for between-domain differences in mean SD(rho)) $= -.39$ ($p < .001$)). Contrary to SD(rho) > 0 being due to unclaimed artifacts, current results suggest that residual variance is at least partially attributable to combinations of moderators, in general support of SS.

Second, although mean SD(rho) shrinks with increasing moderator specificity, it does not drop to 0 at even the highest specificity level per

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

Table 3. Mean SD(rho), Mean %VE, Mean *K*, and Mean SE(r_{xy}) by Aggregation Specificity Level

	Specificity level										
Domain/variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**	**Change from broadest to narrowest[a]**
Cognitive ability (4 studies)											
Count SD(rho)	0	3	28	20	24	32	24	0	0	131	
Mean SD(rho)		.19	.17	.18	.13	.07	.07			.12	86.6%
Count %VE	0	2	10	4	10	0	0	0	0	26	
Mean%VE		57.8	52.1	53.3	63.5					57.1	9.9%
Mean K		75.0	25.6	25.9	80.4	18.4	9.9			50.5	
Mean SE(rxy)		.009	.016	.018	.016	.031	.045			.016	5.19
Personality (4 studies)											
Count	0	1	44	35	40	8	0	0	0	128	
Mean SD(rho)		.15	.10	.07	.08	.11				.08	44.1%
Count %VE	0	1	44	22	12	8	0	0	0	87	
Mean%VE		41.0	60.0	58.6	44.2	52.8				56.6	28.9%
Mean K		17.0	36.5	11.7	10.5	26.4				25.5	
Mean SE(rxy)		.016	.018	.027	.022	.017				.021	1.01
Work attitudes (4 studies)											
Count	0	0	5	33	26	20	9	0	0	93	
Mean SD(rho)			.19	.19	.13	.11	.07			.14	86.9%
Count %VE	0	0	1	8	20	3	0	0	0	32	
Mean%VE			25.2	18.7	23.2	21.6				22.0	-14.3%
Mean K			214.4	65.5	28.2	21.6	14.4			42.7	
Mean SE(rxy)			.006	.012	.014	.012	.017			.013	2.59

Table 3. Continued

Domain/variable	Specificity level 1	2	3	4	5	6	7	8	9	Total	Change from broadest to narrowest[a]
Leadership (4 studies)											
Count	0	0	0	12	11	57	45	45	17	187	
Mean SD(rho)				.13	.21	.17	.12	.09	.05	.13	87.1%
Count %VE	0	0	0	10	3	27	44	42	17	143	
Mean%VE				44.1	14.6	16.3	18.9	27.6	57.8	27.3	31.0%
Mean K				16.9	51.8	27.5	23.8	20.2	15.4	22.5	
Mean SE(rxy)				.033	.010	.015	.020	.027	.038	.024	1.16
Job interview (3 studies)											
Count	1	9	24	27	5	0	0	0	0	66	
Mean SD(rho)	.28	.19	.14	.13	.16					.15	68.5%
Count %VE	1	7	8	12	1	0	0	0	0	29	
Mean%VE	14.0	27.8	53.6	52.2	68.0					45.9	385.7%
Mean K	150.0	54.0	45.0	25.3	27.0					42.0	
Mean SE(rxy)	.004	.011	.014	.023	.022					.017	5.38
Assessment center (3 studies)											
Count	2	20	28	10	0	0	0	0	0	60	
Mean SD(rho)	.12	.11	.10	.09						.10	40.5%
Count %VE	2	20	17	0	0	0	0	0	0	39	
Mean%VE	43.8	38.4	13.9							28.0	-68.2%
Mean K	182.5	29.8	17.1	9.4						32.1	
Mean SE(rxy)	.006	.017	.015	.020						.015	3.43

Table 3. Continued

Domain/variable	Specificity level 1	2	3	4	5	6	7	8	9	Total	Change from broadest to narrowest[a]
SJT (2 studies)											
Count	0	2	26	38	10	0	0	0	0	76	
Mean SD(rho)		.15	.11	.08	.02					.08	97.7%
Count %VE	0	2	19	23	10	0	0	0	0	54	
Mean%VE		11.8	18.9	23.8	94.9					34.8	704.4%
Mean K		106.5	28.6	10.3	4.0					19.1	
Mean SE(rxy)		.006	.020	.028	.036					.026	6.41
All (24 studies)											
Count	3	35	155	175	116	117	78	45	17	741	
Mean SD(rho)	.17	.14	.12	.12	.11	.13	.10	.09	.05	.12	92.1%
Count %VE	3	32	99	79	56	38	44	42	17	410	
Mean%VE	33.87	35.71	42.53	41.36	48.05	24.41	18.90	27.64	57.77	37.4	70.6%
Mean K	171.7	43.9	36.7	25.5	33.0	23.9	18.4	20.2	15.4	29.9	
Mean SE(rxy)	.005	.014	.017	.023	.019	.019	.027	.027	.038	.021	7.34

[a]Top value per block = % decrease in Var (rho), middle value = % increase in %VE, bottom value = proportional increase in SE(*rxy*).

domain. It is unclear how much of the remaining residual variance is due to unclaimed artifacts or to further (untested) moderators. A moderator strongly affecting personality-performance linkages is whether a given estimate was derived under confirmatory (e.g., job analysis-driven) or exploratory (i.e., hit-or-miss) conditions. Tett and colleagues (Tett, Jackson, & Rothstein, 1991; Tett et al., 1999) show that the former, on average, are about twice as strong as the latter. This effect might reasonably extend to other domains.[12] An important implication here is that combining confirmatory-based and exploratory-based estimates in a single aggregation can substantially underestimate mean rho applicable to confirmatory conditions alone due to dilution from the weaker exploratory estimates. In such cases, the upper 90% CV might offer a better subject of generalizability over mean rho for those researchers who have good reason to expect a stronger relationship. The effect of the confirmatory/exploratory distinction and other moderators on further reductions in SD(rho) awaits additional study.

Third, consistent with the shrinking mean SD(rho), there is a corresponding increase in mean %VE with increasing moderator specificity, although the pattern is less clear. Notably, %VE is available for only 55% of the 741 cases, rendering observed changes in %VE less reliable than those in SD(rho). The correlation between the two indices is –.66. The middle value in the far right column of Table 3 (per block) indicates the percent increase in %VE from the broadest to the narrowest aggregation for the given domain. Averaging across domains yields a mean 154% increase (i.e., proportional increase = 2.54).

Fourth, as might be expected, the reduction in SD(rho) and increase in %VE with increasing specificity is accompanied by a general reduction in K and, correspondingly, an increase in second-order sampling error reflected in SE(r_{xy}). The bottom value per block in the rightmost column of Table 3 shows the proportion increase in SE(r_{xy}) moving from the least to most specified aggregations. The range of proportional increases across the seven domains is 1.01 (personality) to 6.41 (SJTs), with mean = 3.60.

Discussion

A key question driving every meta-analytic investigation is that of generalizability. A common focus has been whether or not a lower bound estimate of population validity (90% CV) falls above 0, as a basis for inferring VG. Noted limitations in the technical VG inference (e.g., focus on weaker values) lead us to urge closer attention to SD(rho) and the 80% CI as critical in judging the generalizability of mean rho (Higgins, Thompson, & Spiegelhalter, 2009; Steel, Kammeyer-Mueller, & Paterson, 2015). In only seven of the

[12] None of the 24 reviewed studies directly tested the C/E moderator effect.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

reviewed studies (29%) is SD(rho) > 0 acknowledged as grounds for caution in generalizing mean rho, and in only two (8%) is this caution underscored. This may help explain why variability in meta-analytic effect sizes is so infrequently cited (Carlson & Ji, 2011).

Results show further that, as conditional specificity increases, SD(rho) tends to decrease. This suggests that SD(rho) > 0 represents a distribution of rhos, not just of fallible rho estimates. Specifically, our review revealed an average 73.1% drop in Var(rho) moving from low to high specificity within research domains, which suggests that a large share of Var(rho) is due to moderators, not unclaimed Var(*e*).

Our results suggest a fidelity-bandwidth trade-off between precision in generalizing mean rho (as SD(rho) approaches 0) and the breadth or scope of conditions to which generalizations can be made. Standards for precision in generalizing mean rho are discussed below. A critical question at this point is that, *if precision sufficient to generalize mean rho is achieved only at high levels of aggregation specificity, then in what sense is mean rho "generalizable"?* It is on these grounds that generalizability in meta-analysis might be judged a myth. Broad-level aggregations may yield SD(rho) low enough to garner high precision in generalizing mean rho across conditions (i.e., as essentially a universal truth). More often, however, aggregations at the broadest levels, as shown in Table 3, have SD(rho) too high to permit useable precision in generalizing mean rho to any particular work setting.

The myth of generalizability emerges from recognition of the inverse relationship between precision and scope of generalization: as precision increases, scope decreases. A further inverse relationship leads to a second challenge in meta-analytic generalization. As the available pool of input r_{xy}s is split into increasingly narrowed moderator categories, *K* per moderator subgroup drops (see Table 3). A paradox emerges in that, as SD(rho) shrinks in providing greater precision, the corresponding reduction in *K* increasingly undermines mean *r* and Var(*r*) as starting points for the main (Level 1) calculations. This is similar to a bandwidth–fidelity trade-off except, here, narrowing bandwidth from increasingly specified moderator conditions comes with *lower* fidelity due to corresponding decreases in *K*. Thus, paradoxically, higher precision comes with lower certainty.[13]

Consistent with the paradox, our survey results show that moving from broadest to narrowest aggregation conditions increases SE(r_{xy}) by a factor of 3.6, averaging across research domains. To get a better sense of the implications for generalizing mean rho, consider the case of cognitive ability.

[13] This is not a fundamental flaw in meta-analysis; it is the result of a practically limited number of input r_{xy}s available for aggregation. As available *K* increases, the noted paradox becomes less problematic.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

At the broadest aggregation level (specificity = 2), GMA is linked to job performance, training success (Salgado, Anderson, Moscoso, Bertua, & de Fruyt, 2003), and counterproductive work behaviors (CWBs) (Gonzalez-Mule, Mount, & Oh, 2014), without further specifications. Mean SE(r_{xy}) for these three cases = .009, yielding a 95% confidence interval with a relatively narrow width of ±.017. Ignoring possible second-level biases in sampling r_{xy}s, mean *r* and mean rho can be taken as reasonably robust estimates of targeted population values. However, mean SD(rho) for those broad aggregations = .19, yielding an 80% CI with width ±.239. Given that residual variance is not readily attributable to unclaimed artifacts (as per results in Table 3), generalizing mean rho from these aggregations to any actual specific work setting would be tenuous.

At the narrowest aggregation level in the cognitive ability domain (specificity = 7), facets of GMA are linked to job performance of, for example, women working in high-complexity, male-dominated jobs (Rothstein & McDaniel, 1992). Mean SD(rho) for the 24 cases = .07, narrowing the 80% CI to width ±.087, which is a 64% decrease. The improved precision in generalizing mean rho, however, comes at a price: mean SE(r_{xy}) has increased from .009 to .045, yielding a 95% confidence interval five times wider, at ±.087 (up from ±.017), around mean *r*. Thus, certainty in mean rho weakens despite improved precision around that estimate at higher aggregation specificity.

Our findings and the issues that drove our survey promote more careful attention to the concept and mechanics of generalizability in meta-analysis. We suggest that there are four critical questions guiding generalizability inferences.

Question 1: What Are We Generalizing?

In drawing inferences of generalizability from meta-analysis, it is important to be clear regarding what exactly is being generalized. Use of the 90% rule, as in technical VG inferences, draws attention, of course, to the 90% CV. What is being generalized here is a judgment of reasonably minimal validity. One might just as easily and informatively focus on the upper 90% CV (i.e., upper boundary of the 80% CI) as rho achievable under favorable (e.g., confirmatory) conditions. By focusing on mean rho, in the context of technical SS inferences, most meta-analyses implicitly promote the mean as what readers can take as a generalizable "truth" from the meta-analysis. Consumers need to be wary here, as a positive VG inference, bearing on the lower 90% CV, does not by itself imply rejection of SS in promoting generalizability of mean rho. Further defining the "what" of generalizability are the particular artifact corrections employed: Mean rho, reflecting such corrections, is a theoretical ideal, whereas mean *r*, lacking corrections, permits more realistic application.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

Question 2: To Where Are We Generalizing?

Just as important as the "what" of generalization is the "where." This is a function of factors operating at each of the two levels. Operating at Level 1 are all the various conditions defining the given aggregation (e.g., features listed in Table 1). Conditional specificity is important to the degree the rho targeted for generalization (e.g., mean, lower CV, upper CV) actually varies between moderator subgroups. Failed moderators introduce false specificity in generalizations. Note, however, that comparison between subgroups can extend beyond the means, depending on what is being generalized (as per Q1). Subgroups might have identical mean rhos, but different SD(rhos) (i.e., because further moderators are operating differentially across subgroups). These will yield different CVs even when mean rhos are the same. Accordingly, care is needed in generalizing across subgroups beyond considering only their means.

Operating at the second level are second-order sampling error, and choices and biases determining the sample of r_{xy}s from which mean r and Var(r) derive. Similar to what happens in primary-level research (i.e., with people as units of analysis), conditions targeted for generalization from meta-analysis may vary from the conditions actually represented in the r_{xy} sample: The intended population, for example, might not exclude military settings, yet available r_{xy}s may derive only from civilian settings.

Question 3: How Precise Is Our Generalization?

Precision of generalization is conferred by a low SD(rho) as a function of (a) the specificity of the aggregation conditions (i.e., as per moderator subgroups) and (b) uncorrected artifacts, whether available or not. As SD(rho) increases from 0, generalizing mean rho becomes less precise. How much imprecision should be tolerated is difficult to nail down. Oswald and McCloy (2003) cautiously suggest SD(rho) cutoffs of .125, .075, and .025 as large, moderate, and small, corresponding to 95% intervals with widths of ±.25, ±.15, and ±.05 correlation units, respectively. Thus, where SD(rho) = .125, 95% of rhos are understood to fall within ±.25 around mean rho. Such imprecision severely limits meaningful generalizability of mean rho. We suggest that Oswald and McCloy's SD(rho) < .025 offers a reasonable, albeit rigorous, standard, yielding an 80% CI with width ±.032.

Question 4: How Certain Is Our Generalization?

As noted above in discussing the paradox of low-K specificity, a further issue to consider is SE(r_{xy}), which reflects second-order sampling error in estimating population mean r as a foundation for mean rho. Standards for SE(r_{xy}), as far as we are aware, currently do not exist. Statistical standards (e.g., $p < .05$, power > .80) are commonly used as inferential guideposts, and we

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

suggest that an SE standard may prove useful in inferring generalizability of mean *r* and mean rho. Borrowing from Oswald and McCloy (2003), we tentatively offer as a criterion for inferential certainty that SE(r_{xy}), as with SD(rho), also not exceed .025, yielding a 95% confidence interval with width ±.05 around mean *r*.[14] Softer standards in each case may be defensible but only with reduced precision and/or certainty in generalization. Small-*K* aggregations limit both certainty in representing the intended r_{xy} distribution and the power to detect moderators (e.g., Hedges & Pigott, 2004; Hunter & Schmidt, 2004; Sackett, Harris, & Orr, 1986).

To summarize, generalizability inferences from meta-analysis can be framed in terms of (a) what is being generalized, (b) the conditions specifying the scope of generalization, (c) precision of generalization as per SD(rho), and (d) certainty of generalization as per SE(r_{xy}). As meta-analysis methods bring rigor to literature review, the four questions promote rigor in the interpretation of meta-analytic results bearing on generalizability.

Application of the SD(rho) and SE(r_{xy}) Standards

Applying the SD(rho) (Level 1) and SE(r_{xy}) (Level 2) standards to the 24 studies in our survey identifies 113 of the 741 cases (15.2%) with SD(rho) < .025 and 551 cases (74.4%) with SE(r_{xy}) < .025. Cases meeting both conditions number 34 (4.6%)—a surprisingly small proportion in light of the strong focus on mean rho by those citing meta-analytic findings (Carlson & Ji, 2011). The data permit a more detailed analysis, as follows.

Given that moderator specificity relates negatively with SD(rho) and positively with SE(r_{xy}) (see Table 3), we might expect results of broader aggregations (at lower specificity) to more likely fail the L1 precision standard (SD(rho) is too large) but pass the L2 certainty standard (SE(r_{xy}) is small); whereas, at the other end, narrower aggregations should more likely pass the precision standard (small SD(rho)) but fail the certainty standard (SE(r_{xy}) is too large). The top part of Figure 3 shows the percentages of cases passing the .025 standards at each of the nine specificity levels. Corresponding percentages are shown as well for the .075 and .125 standards, from Oswald and McCloy (2003). Clearly evident is how the precision and certainty criteria operate in opposite directions as a function of aggregation specificity. Notably, the specificity–SE(r_{xy}) relationship would be expected even if moderators have zero influence on rho. SD(rho), on the other hand, is not a direct function of *K*; shrinking SD(rho) with increasing specificity offers empirical support for SS, operating especially at lower levels of aggregation specificity.

[14] SE(r_{xy}) derived using the heterogeneous case equation includes SD(rho), which varies by the artifacts included for correction. This complicates identification of standards for SE(r_{xy}). We offer the .025 criterion as a plausible focal point for discussion of certainty standards in generalizing mean rho.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

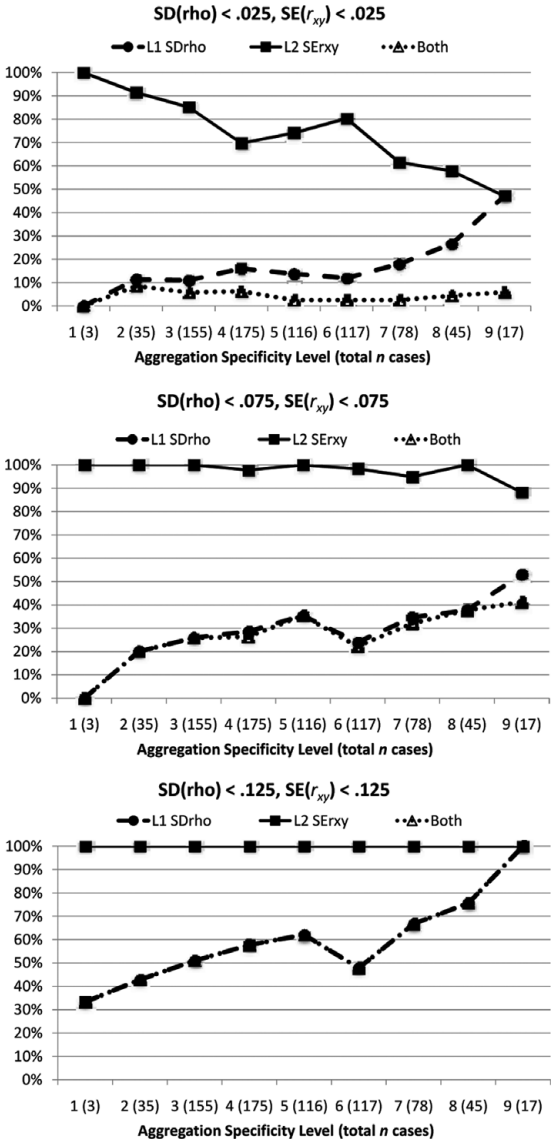

Figure 3: Percentages of cases meeting three Level 1 and Level 2 generalizability criteria at each of nine moderator specificity levels.

Also plotted is the percentage of cases that pass both criteria, per specificity level. Consider the top part of Figure 3, which shows results based on the relatively rigorous .025 standards. Of the total 741 cases, only 34 (4.6%) meet both conditions, and the highest percentage of cases meeting both is 8.6% at specificity Level 2. The middle and lower parts of the figure show increasing percentages of cases meeting weaker standards: 28.2% of the 741 cases meet both .075 standards and 57.6% meet both .125 standards. Thus,

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

even using relatively liberal precision and certainty standards, 42% of the 741 mean rhos fail to qualify for generalization.

The myth of generalizability is evident here in two respects. First, overall pass rates for generalizability are low, especially at the more robust levels of precision and certainty. Second, pass rates generally improve as aggregation specificity increases. The paradox of low-*K* specificity is evident as well, in that increasing precision (i.e., shrinking SD(rho)) is met with loss of certainty (i.e., increasing SE(r_{xy})) as specificity increases. That 34 cases met both high-precision and high-certainty standards shows the paradox is not absolute. Indeed, *K* may not be small (e.g., > 50) even for sequentially nested moderator subgroups. *K* does naturally tend to drop, however, with increasing specificity, given that input estimates, in practice, are of limited supply. Effects on SE(r_{xy}) should not be ignored when making and evaluating generalizability inferences in meta-analysis.

Utility Implications of Imprecision in Generalizing Validity Estimates

To see the practical merits of precision in generalizability, consider application of mean *r*[15] in estimating utility (i.e., cost savings from use of a given predictor) using the Brogden–Cronbach–Gleser model (cf., Hoffman & Thornton, 1997, p. 461). With high precision (i.e., low SD(rho)), the user need only confirm that the local setting is one represented in the aggregation. Mean *r* in that case will yield a reliable estimate of utility (per selection ratio, SD_y$, and testing cost). Decreasing precision (i.e., increasing SD(rho)) creates ambiguity in expected savings for the given predictor. Consider the moderate-precision case where SD(rho) = .075, yielding an 80% CI with width of ±.10. Assume further that mean r = .25, $\$SD_y$ = $20,000, selection ratio = .20, cost of testing = $20/applicant, and N applicants = 100. Using the upper 90% CV (i.e., r = .35), utility = $9,588/hire and $191,760 for all 20 hires. Using the lower 90% CV (i.e., r = .15), utility = $4,212/hire and $84,240 for all 20 hires. The $107,520 difference in total utility (i.e., $5,376/hire) between the upper and lower estimates is far from trivial; users should be very interested to know which estimate of *r* applies to their particular situation, whether between .15 and .35 or outside that interval. Where SD(rho) = .025 (i.e., precision is high), the overall difference in utility between the upper and lower 80% CI estimates = $35,840 (i.e., $1,792/hire). Although substantially lower, this difference in savings might still drive identification of *r* most likely applicable to local conditions.

[15] Practical applications call for generalizability of mean *r*, not mean rho.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

Tertiary Analyses Using Meta-Analytic Mean Correlations

Meta-analytic mean *r*s and rhos are often used as input into further analyses, including regression to examine multivariate and incremental validity (e.g., Hoffman, Kennedy, LoPilato, Monahan, & Lance, 2015), and path analyses in testing model fit (e.g., Tett & Meyer, 1993). Of the 24 studies examined here, 11 (46%) entered means as input into such analyses. The attraction of mean rho is understandable (i.e., reduced sampling error and control of artifacts). Such applications raise concerns, however, when SD(rho) > 0 for one or more of the input estimates. Consider regression with three predictors, A, B, and C, each in relation to Y.

The regression coefficients in a three-predictor equation are a function of six bivariate *r*s (AY, BY, CY, AB, AC, and BC). When SD(rho) = 0 for all six mean *r*s, regression coefficients can (also) be expected to be consistent across conditions. However, as SD(rho) increases per mean *r* in the input matrix, the regression coefficients will tend to be even less stable than the bivariate *r*s. This directly extends to analyses of incremental validity, whether in terms of regression coefficients or changes in R^2 (i.e., the squared semipartial *r*). Because the estimation of *R* involves all six mean *r*s in the matrix, the variability in *R* across situations is compounded and the standard deviation of these estimates will be *much* larger than the SD(rho)s in the input matrix. Thus, regression results using meta-analytic mean *r*s are less generalizable than the mean *r*s are themselves, as SD(rho) increases per input mean *r*.

Similarly, when using mean *r*s to run path analysis, resulting fit indices apply to the model specified using those means. Given SD(rho) = 0 for all input *r*s, fit indices may be considered generalizable to the conditions represented in the corresponding aggregations. To the degree SD(rho) > 0 per input *r*, however, observed fit based on just the selected values is of questionable generalizability. Fit, both per model in absolute terms and comparatively across models, is likely to vary substantially depending on one's choice of input values (e.g., mean rho, lower 90% CV, upper 90% CV).

For all tertiary analyses involving mean *r*s or rhos, some combinations may be unlikely (e.g., involving the lower 90% CV for some linkages and the upper 90% CV for others) or mathematically impossible. But if choices are to be made among available input values that might narrow the range of viable combinations, it is not clear on what grounds such choices might be made. Viswesvaran and Ones (1995) suggest adding select moderators to meta-analytically based path analysis. What to do when SD(rho) > 0 within moderator subgroups, however, is far from clear, and the problem grows multiplicatively where SD(rho) > 0 per input correlation. Relying on only mean rho offers convenience but at best an incomplete test of targeted path models with respect to generalizability.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

A related concern is that mean rhos for different relationships, especially if drawn from different meta-analytic studies, inevitably represent different aggregation conditions, creating an apples and oranges problem when entered into the same tertiary analysis. Thus, even if SD(rho) is small for all input values, generalizability may be strained by heterogeneity across inputs in the conditions to which generalizations can be aimed.

Recommendations on Reporting and Interpreting Meta-Analytic Inferences

Our analyses lead us to offer the following suggestions regarding how meta-analytic results are presented and used with respect to generalizability.

1. Meta-analysts are urged to be clear about four outputs per aggregation: (a) the subject of generalization, whether it is mean rho, mean *r*, the lower or upper boundary of the 80% CI, or some other logically or empirically relevant value; (b) the conditions to which generalizations can reasonably be made, as per study inclusion and exclusion criteria, possible bias in source availability, and moderators; (c) precision of generalizability (e.g., SD(rho) $<$ .025, .075, or .125); and (d) certainty of generalizability (e.g., SE(r_{xy}) $<$ .025, .075, or .125).
2. Meta-analysts are encouraged to be cautious in entering mean rhos into tertiary analyses to the degree SD(rhos) and SE(r_{xy})s are $>$ 0. Doing so compounds imprecisions and uncertainties, yielding results even less generalizable than the input estimates. Tertiary analyses, if run, might be repeated using different combinations of estimates from the rho distributions. Managing results under such conditions can be expected to pose interpretive challenges, particularly with respect to generalizing tertiary analysis results.
3. Consumers of (I-O) meta-analytic findings are urged to consider SD(rho) and SE(r_{xy}) as equal in importance to mean rho when drawing inferences (Higgins et al., 2009; Tett & Christiansen, 2007). Large SD(rho)s and SE(r_{xy})s signal limitations in generalizing mean rho. The larger SD(rho) is, the more attention needs to shift away from mean rho and to the 80% CI, where upper and lower boundaries point to rhos achievable under select conditions. Identifying those conditions is a logical next step.
4. Meta-analytic consumers seeking to apply mean rho in particular situations need to be careful to consider how similar those situations are to the conditions defining the source aggregation. The .025, .075, and .125 standards offer initial guidance in judging generalizability. Where SD(rho) $>$ 0, practitioners might have reason to believe the targeted situation is one where rho is especially strong (or weak). Local validation offers viable confirmation only to the degree that *N* is large. Reliance on

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

multiple, preferably converging lines of evidence (e.g., predictive, concurrent, job analysis, expert linkage judgments) will often be the safer bet.

Conclusions

Generalizability is a key aim in science, and aggregation of past findings both promotes and frames generalizability inferences. That SD(rho) in each of the seven targeted research domains shrinks with increasing moderator specificity suggests prominent and replicable contributions of moderators to Var(rho), in support of SS. Aggregations where SD(rho) > 0 call for further research to identify the conditions affecting rho. Meta-analysis thus serves less a terminal role in scientific discovery, as it is commonly treated with a focus on mean rho and the 90% CV, than an intermediary role, offering disciplined summary of a literature in the identification of promising directions for further research in light of SD(rho) and the 80% CI. This is not to suggest a return to the old days of errant searches for generalizability via (small-*N*) local validation, but rather it is an acknowledgment that the world (of work) is complex; that targeted phenomena are affected by multiple, intersecting (nonartifactual) forces; and that meta-analysis, contrary to the way it is often expected to work, does not guarantee simple answers.

Our findings reveal a critical trade-off in meta-analysis between precision and scope of generalization: Reaching adequate precision typically requires a level of moderator specificity that challenges the inference that mean rho is truly "generalizable." The "myth of generalizability" is most clearly evident in the 95% of the 741 aggregations that failed plausible, albeit rigorous, generalizability standards (L1: SD(rho) $< .025$; L2: SE(r_{xy}) $< .025$). Beyond the myth, increasing specificity, thus conferring increased precision, also decreases certainty in mean *r* and Var(*r*) as starting points for precision calculations, thereby creating a "paradox" of low-*K* specificity.

Despite the myth, the paradox, and the complexity of its answers, meta-analysis is still a highly valuable tool, not only for what it generates in mean rhos, SD(rhos), and moderator effects, but also for how it advances our knowledge of what we know and how we come to know it. It is, as Sackett (2003, p. 111) suggests, more "a theory about the process of drawing inferences from cumulative data [than] a general statement that the bulk of the variability in research findings is due to statistical artifacts." Building on Sackett's point, we can ask several key questions framing generalizability inferences in meta-analysis: "What is being generalized to where?" and "How precise and how certain are those inferences?" Collectively, answering those questions extends, to some degree, the rigor of meta-analytic methods to the interpretation and use of meta-analytic findings.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

An important aim in offering this article is to spur discussion of concepts and inferential processes in meta-analysis, such as VG and SS, and to clarify the meaning of generalizability per se. What standards for precision and certainty are to be settled on as most defensible, whether .025, .075, .125 or something else, and what conditions might drive that choice, are key questions going forward. Also of interest is pursuit of refinements to the measurement of aggregation/moderator specificity and replication of its observed negative relationship with SD(rho), based on a broader representation of meta-analyses in I-O psychology and other disciplines. Exchanges and advances along such lines, we hope, will promote the yields from meta-analysis as a source of generalizable knowledge.

References[16]

*Arthur, W. Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*(1), 125–154.

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

*Bartram, D. (2005). The Great Eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185–1203.

*Beus, J. M., Dhanani, L. Y., & McCord, M. A. (2015). A meta-analysis of personality and workplace safety: Addressing unanswered questions. *Journal of Applied Psychology, 100*(2), 481–498.

Biemann, T. (2013). What if we were Texas sharpshooters? Predictor reporting bias in regression analysis. *Organizational Research Methods, 16*(3), 335–363.

Carlson, K. D., & Ji, F. X. (2011). Citing and building on meta-analytic findings: A review and recommendations. *Organizational Research Methods, 14*(4), 696–717.

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York: Guilford.

*Choi, D., Oh, I., & Colbert, A. E. (2015). Understanding organizational commitment: A meta-analytic examination of the roles of the five-factor model of personality and culture. *Journal of Applied Psychology, 100*(5), 1542–1567.

*Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*(1), 83–117.

Coburn, K. M., & Vevea, J. L. (2015). Publication bias as a function of study characteristics. *Psychological Methods, 20*(3), 310–330.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129.

Cortina, J. M. (1993). What is coefficent alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Dalton, D. R., Aguinis, H., Dalton, C. M., Bosco, F. A., & Pierce, C. A. (2012). Revisiting the file drawer problem in meta-analysis: An assessment of published and nonpublished correlation matrices. *Personnel Psychology, 65*(2), 221–249.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*(1), 120–128.

*Gaugler, B. B., Rosenthal, D. B., Thornton III, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493–511.

[16] Asterisks indicate the 24 meta-analytic studies included in the current review.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

*Gonzalez-Mulé, E., Mount, M. K., & Oh, I. (2014). A meta-analysis of the relationship between general mental ability and nontask performance. *Journal of Applied Psychology, 99*(6), 1222–1243.

Higgins, J. P., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 172*, 137–159.

Hedges, L. V., & Olkin, I. (2005). *Statistical methods for meta-analysis* (2nd ed.). Orlando, FL: Academic Press.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9*(4), 426–445.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21*(4), 299–332.

*Hoffman, B. J., Kennedy, C. L., LoPilato, A. C., Monahan, E. L., & Lance, C. E. (2015). A review of the content, criterion-related, and construct-related validity of assessment center exercises. *Journal of Applied Psychology, 100*(4), 1143–1168.

Hoffman, C. C., & Thornton III, G. C. (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology, 50*(2), 455–470.

Hough, L. M., Ones, D. S., & Viswesvaran, C. (1998, April). Personality correlates of managerial performance constructs. Paper presented in R. C. Page (Chair), *Personality determinants of managerial potential performance, progression and ascendancy*. Symposium conducted at the 13th Annual Conference of the Society for Industrial Organizational Psychology, Dallas, TX.

*Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U. (2004). The impact of job complexity and study design on situational and behavior description interview validity. *International Journal of Selection and Assessment, 12*(3), 262–273.

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

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., & McIntyre, H. H. (2010). Situational specificity and validity generalization. In J. L. Farr & N. T. Tippins (eds.), *Handbook of employee selection* (pp. 909–920). New York: Routledge.

*Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin, 127*(3), 376–407.

*Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*, 765–780.

*Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: A meta-analytic test of their relative validity. *Journal of Applied Psychology, 89*(5), 755–768.

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods, 15*(4), 624–662.

Kromrey, J. D., & Rendina-Gobioff, G. (2006). On knowing what we do not know: An empirical comparison of methods to detect publication bias in meta-analysis. *Educational and Psychological Measurement, 66*(3), 357–373.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

*Lord, R. G., de Vader, C. L., & Alliger, G. M. (1986). A meta-analysis of the relation between personality traits and leadership perceptions: An application of validity generalization procedures. *Journal of Applied Psychology, 71*(3), 402–410.

*Lowe, K. B., Kroeck, K. G., & Sivasubramaniam, N. (1996). Effectiveness correlates of transformation and transactional leadership: A meta-analytic review of the MLQ literature. *The Leadership Quarterly, 7*(3), 385–425.

*Martin, R., Guillaume, Y., Thomas, G., Lee, A., & Epitropaki, O. (2016). Leader–member exchange (LMX) and performance: A meta-analytic review. *Personnel Psychology, 69*(1), 67–121.

*McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*(4), 599–616.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology, 59*(4), 927–953.

*McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*(1), 63–91.

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*(4), 194–206.

Murphy, K. R. (2003). *Validity generalization: A critical review*. Mahwah, NJ: Erlbaum.

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

Oswald, F. L., & McCloy, R. A. (2003). Meta-analysis and the art of the average. In K. Murphy (ed.), *Validity generalization: A critical review* (pp. 311–338). Mahwah, NJ: Erlbaum.

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

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

*Rothstein, H. R., & McDaniel, M. A. (1992). Differential validity by sex in employment settings. *Journal of Business and Psychology, 7*(1), 45–62.

Sackett, P. R. (2003). The status of validity generalization research: Key issues in drawing inferences from cumulative research findings. In K. Murphy (ed.), *Validity generalization: A critical review* (pp. 91–114). Mahwah, NJ: Erlbaum.

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to type I error. *Journal of Applied Psychology, 71*(2), 302–310.

Sackett, P. R., Tenopyr, M. L., Schmitt, N., Kehoe, J., & Zedeck, S. (1985). Commentary on forty questions about validity generalizations and meta-analysis. *Personnel Psychology, 38*, 697–798.

*Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & de Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology, 56*(3), 573–605.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology, 65*, 643–661.

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. (1978). Moderator research and the law of small numbers. *Personnel Psychology, 31*, 215–231.

Schmidt, F. L., & Hunter, J. E. (1984). A within setting test of the situational specificity hypothesis in personnel selection. *Personnel Psychology, 37*, 317–326.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for two jobs in the petroleum industry. *Journal of Applied Psychology, 66*, 261–273.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences and validity of aptitude tests in selection: A red herring. *Personnel Psychology, 38*, 697–798.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Rothstein-Hirsh, H. (1985). Forty questions about validity generalization and meta-analysis. *Journal of Applied Psychology, 66*, 166–185.

*Schmidt, F. L., Hunter, J. E, Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology, 32*, 257–281.

Schmidt, F. L., Ocasio, B. P., Hillery, J. M., & Hunter, J. E. (1985). Further within-setting empirical tests of the situational specificity hypothesis in personnel selection. *Personnel Psychology, 38*, 509–524.

Steel, P. D. G., & Kammeyer-Mueller, J. (2008). Bayesian variance estimation for meta-analysis: Quantifying our uncertainty. *Organizational Research Methods, 11*(1), 54–78.

Steel, P., Kammeyer-Mueller, J., & Paterson, T. A. (2015). Improving the meta-analytic assessment of effect size variance with an informed Bayesian prior. *Journal of Management, 41*(2), 718–743.

https://doi.org/10.1017/iop.2017.26 Published online by Cambridge University Press

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. Rothstein, A. J. Sutton, & M. Borenstein (eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 223–240). Chichester, UK: Wiley.

Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye Hollenbeck, Murphy, and Schmitt (2007). *Personnel Psychology, 60*, 967–993.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703–742.

Tett, R. P., Jackson, D. N., Rothstein, M., & Reddon, J. R. (1999). Meta-analysis of bi-directional relations in personality-job performance research, *Human Performance, 12*, 1–29.

*Tett, R. P., & Meyer, J. P. (1993). Job satisfaction, organizational commitment, turnover intention, and turnover: Path analyses based on meta-analytic findings. *Personnel Psychology, 46*, 259–293.

van Assen, Marcel A. L. M., van Aert, Robbie C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20*(3), 293–309.

Vinchur, A. J., Schippmann, J. S., Switzer, F. S., & Roth, P. L. (1998). A meta-analytic review of job performance for salespeople. *Journal of Applied Psychology, 83*, 586–597.

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology*, *48*(4), 865–885.

*Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61*(4), 275–290.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75*, 315–321.

*Zimmerman, R. D. (2008). Understanding the impact of personality traits on individuals' turnover decisions: A meta-analytic path model. *Personnel Psychology, 61*, 309–348.