ANNUAL REVIEWS

Annual Review of Organizational Psychology and Organizational Behavior

Meta-Analysis in Organizational Research: A Guide to Methodological Options

Scott B. Morris

Department of Psychology, Illinois Institute of Technology, Chicago, Illinois, USA; email: scott.morris@iit.edu

Annu. Rev. Organ. Psychol. Organ. Behav. 2023. 10:225–59

First published as a Review in Advance on November 28, 2022

The *Annual Review of Organizational Psychology and Organizational Behavior* is online at orgpsych.annualreviews.org

https://doi.org/10.1146/annurev-orgpsych-031921-021922

Keywords

meta-analysis, validity generalization, effect size, research synthesis, psychometric artifacts

Abstract

Meta-analysis provides a powerful tool for integrating findings from the research literature and building statistical models to explore trends and inconsistencies in the research base. Meta-analysis starts with a process for translating results from each study into an effect size that represents all findings in a common metric. Statistical models are then applied to estimate the mean, variance, and moderators of effect size. This article explores several key decision points in conducting a meta-analysis, including issues in obtaining a common metric, accounting for psychometric artifacts, and choosing an appropriate statistical model. It provides recommendations for choosing among alternate approaches and reporting results to ensure transparency.

INTRODUCTION

Meta-analysis is a powerful methodology for summarizing results across a body of literature, and it has become a highly influential way of communicating the state of the science on a topic. This high-level view of the literature comes at a cost—the meta-analyst is typically far removed from the actual data, increasing the opportunities for multiple forms of error to creep into the analysis. More specifically, meta-analysis rests on the back of a scientific enterprise for reporting research findings that has historically provided incomplete data, lack of transparency in research practices, and potentially skewed samples of research findings (Banks et al. 2016; Kepes et al. 2012, 2022). Meta-analytic procedures frequently rely on statistical adjustments to counteract some of the limitations of primary study findings, often with insufficient evidence that the data correspond to the assumptions of the statistical corrections. Rigorous meta-analytic research demands careful attention to the nature of the data used to obtain effect size estimates and transparency in the assumptions we are making about these data. Fortunately, modern meta-analysis methodology provides many options to adapt analytic strategies to align with the demands of a particular context and to conduct sensitivity analyses regarding the underlying assumptions. This article reviews best practices and emerging strategies for operationalizing effect sizes and analyzing trends and inconsistencies in the research base, with the goal of increasing the rigor and transparency of meta-analytic research.

COMMON METRIC EFFECT SIZE

The first step in conducting a meta-analysis is to represent research findings as an effect size statistic. Common effect size indices include the correlation coefficient (r) for bivariate relationships, the standardized mean difference (d) for group comparisons and the odds ratio (OR) for relations between dichotomous variables. Various other effect size statistics are regularly used in primary research studies (e.g., R^2, η^2, ω^2, raw and standardized regression coefficients) but are less common in meta-analysis. Increasingly, researchers are developing context-specific effect size indices that are tightly aligned with particular research questions. Examples include the dmod statistic for quantifying categorical moderators and differential prediction (Dahlke & Sackett 2018, Nye & Sackett 2017), indices of the magnitude of indirect effect in mediation analysis (van Zundert & Miočević 2020), and group differences in variance (Senior et al. 2020).

Because effect sizes serve as the basic data for meta-analysis, the operationalization of effect size should be approached with the same care that is given to the design of data collection methods in primary research. The choice of effect size should consider several important features that are useful for the purpose of conducting a meta-analysis. First, the effect size should be a pure reflection of the magnitude of a phenomenon, such as the amount of change induced by an intervention or the strength of the relationship between a predictor and an outcome (Kelley & Preacher 2012). Importantly, effect sizes do not reflect the precision or accuracy with which these data estimate the population parameter of interest, and should not be sensitive to the sample size. Unlike tests of statistical significance, which reflect a combination of magnitude and precision, we can separate these two features of the data by reporting an effect size estimate along with a confidence interval reflecting the degree of uncertainty in that estimate (Cumming 2012, Morris & Shokri 2021).

Second, for the purpose of meta-analysis, an effective effect size must present results in a common metric that is comparable across studies. Averaging effect sizes will be meaningful only if they all estimate the same thing on the same scale. Determining a common metric is often more nuanced than choosing a standardized effect size statistic, an issue that is addressed in more detail in the section titled Standardized Effect Size Indices.

Third, the sampling error variance of the effect size statistic must be accurately estimated. Sampling error variance is inversely related to the precision of the effect size estimate; it reflects the

extent to which we would expect the effect size to vary from sample to sample solely due to random sampling of participants from the population of interest. The sampling error variance affects study weights and meta-analytic summary results (e.g., confidence intervals, between-study variance estimates); therefore, appropriate sampling error variance estimates are critical for accurate meta-analytic results.

When selecting an effect size statistic for a meta-analysis, it is important to recognize that effect sizes serve various purposes: communicating the practical significance of findings, as a key determinant in power analysis, and as a way of providing a common metric for comparing results across studies. It is this last function that is particularly relevant to meta-analysis, where a common metric is an essential prerequisite of quantitative synthesis. Features of an effect size that may be useful for other purposes (e.g., power analysis) may not necessarily make it a good candidate for meta-analysis. For example, Murphy et al. (2009) discuss the usefulness of percent variance statistics (e.g., R^2) for power analysis, but these statistics are generally not recommended as effect sizes in meta-analysis (Schmidt & Hunter 2015).

Standardized Effect Size Indices

The need for a common metric arises from the fact that, in most research domains, the pool of studies will use a variety of measures, each with a unique scale. Because we lack a common set of measures on which to compare findings, Oswald & McCloy (2003) describe meta-analysis as a massive missing data problem that we attempt to solve through standardization. Standardization provides a common metric in a very specific sense. Hedges (1982) noted that if the outcome variables are linearly equatable, then the standardized mean difference will be equivalent across different metrics for the outcome variables. That is, we assume that the scale of the outcome in any study can be obtained by applying a linear transformation to the scale of any other outcome included in the meta-analysis. This transformation will have a parallel effect on the numerator and denominator of the effect size; therefore, differences in scaling will cancel out in the calculation of the effect size. Thus, the standardized effect size can be considered scale free. However, for this procedure to work properly, the standard deviation (SD) for each measure needs to be estimated on the same population. Standardization will fail to produce a common metric if there are differences across studies in the composition of samples or other conditions that influence score variance.

Kelley & Preacher (2012) note that standardized effect sizes are inherently multidimensional—a larger effect size can be caused by a larger mean difference or by a smaller SD. Thus, differences between studies in the composition of the sample may introduce an additional source of study-to-study variability in effect size, variance that is not due to differences in effect magnitude but rather due to more or less homogeneous samples. For example, in examinations of differential validity, differences between groups in range restriction can introduce artificial differences in correlations (Roth et al. 2014). Although standardization is the norm, some researchers caution against the use of standardized effect sizes, arguing that they are inherently ambiguous because they can be affected by differences in research design, sample homogeneity, and other confounding factors (Baguley 2009).

Standardization is not needed for all statistics. For example, differences or ratios of proportions are directly interpretable in the original metric. Additionally, when a pool of studies involves a common outcome variable, unstandardized effect sizes are perfectly appropriate and can avoid some of the complexities introduced by standardization. Methods have been developed for meta-analysis of unstandardized mean differences (Bond et al. 2003, Borenstein et al. 2021) and unstandardized regression coefficients (Becker & Wu 2007, Raju et al. 1986).

When computing the standardized mean difference, there is often a choice between several alternate SD estimates. In a typical between-groups experiment, the effect size might be

standardized using the SD of the control group (Glass et al. 1981) or the pooled SD across groups (Hedges 1982). If homogeneity of variance can be safely assumed to hold, pooling SDs across independent groups is preferred, because doing so will produce a more precise estimate of effect size, resulting in a narrower confidence interval on the mean effect size and greater power to detect moderator variables. However, in situations where variance may differ between groups, use of a pooled SD will introduce an additional source of ambiguity in the interpretation of the effect size. In the presence of heterogeneity, the pooled SD is unlikely to be equivalent across studies (e.g., due to unequal subgroup size or difference in the factors that produce heterogeneity). In such cases, standardization using a reference group that is common across studies is more likely to produce effect sizes that are directly comparable.

Alternate Research Designs

Obtaining a common metric through standardization can be particularly challenging when the collection of studies employs a variety of research designs or when estimates are taken from different analytic models. A good example of this issue arises when meta-analyzing research involving a mix of independent groups and repeated-measures designs (Morris & DeShon 2002). For example, in a study comparing pre-post change in a treatment versus control group, the result could be standardized using the SD of pretest or posttest scores (or an average of both), change scores, or covariate-adjusted scores. Standardized effect sizes computed using these alternate SDs will generally not be comparable to one another. In particular, the SD of change scores will often be substantially smaller than the SD of raw scores, resulting in the false appearance of larger effect size in studies utilizing the change-score metric. Study design is often coded as a moderator variable, and several studies have demonstrated larger effect sizes in repeated-measures designs (Finkelstein et al. 1995, Lipsey & Wilson 1993, Olian et al. 1988). Establishing a common metric is essential for such analyses to be interpretable as true differences and not simply an artifact of effect size scaling.

When alternate standardization metrics are available, the choice among them should be informed by both the interpretability of the resulting effect sizes in relation to the research question and the likelihood of obtaining a common metric across studies (Morris & DeShon 2002). For example, when evaluating an organizational intervention (e.g., the effectiveness of a training program), the magnitude of the effect is easiest to conceptualize in terms of the variability of baseline scores. A result of $d = 1.0$ indicates that an average person at baseline is expected to move to 1 SD above the mean, or the eighty-fourth percentile, as a result of the intervention. Alternatively, if the research question were about group differences in the growth rate, it may be more natural to define the effect size in terms of change score metric, such that $d = 1.0$ would indicate that an average person in the treatment group showed more change than 84% of the control group. In most situations, the baseline metric will be more easily understood.

Another consideration in choosing an effect size metric is the need to identify a metric that is compatible across studies. To this end, one strategy is to identify a referent population, such as the population of untreated individuals, and choose the SD from each study that best reflects this common referent population. Control group and pretest SDs are often the most comparable across all studies in a meta-analysis and thus are the best choice for computing standardized effect sizes (Becker 1988, Morris 2008).

Similar issues arise in other contexts as well, including factorial designs (Cortina & Nouri 2000, Morris & DeShon 1997, Olejnik & Algina 2000) and multilevel analyses (Feingold 2009, Hedges 2009). Whenever the pool of studies involves a mix of research designs, careful attention should be paid to the choice of SD to ensure that effect sizes reflect a common population across

CHOOSING AN SD FOR THE STANDARDIZED MEAN DIFFERENCE

1. Standardize using a reference group that is common across studies.
2. Use the pooled SD if it is reasonable to assume that variances are homogeneous; otherwise, use the SD of the control group or a common reference group.
3. For repeated-measures designs, use the pretest or baseline SD.

studies. A summary of recommendations is provided in the sidebar titled Choosing an SD for the Standardized Mean Difference.

In addition to identifying a common basis for standardization, different research designs will affect the sampling error variance of the effect size estimate. Common formulas for the variance of the standardized mean difference are based on the comparison of two independent groups, and may misrepresent the precision of effect sizes from other designs. Alternative estimates of sampling error variance have been discussed in the context of repeated-measures (Morris 2008, Morris & DeShon 2002) and multilevel designs (Hedges 2009).

In summary, it is important to understand that computing a standardized effect size like the correlation coefficient or Cohen's *d* will not magically make studies comparable. Standardization may be necessary to account for scaling differences across measures, but standardization using the sample SDs can also introduce sources of incompatibility while removing others. Furthermore, in many situations, more than one standardization metric will be available, and choices in these situations should be intentional and clearly documented in research reports.

CORRECTION FOR STATISTICAL ARTIFACTS

Statistical artifacts refer to features of a research design that cause the estimated effect size to differ from the value of the statistic in the population of interest. No study is perfect, and limitations of the research design can systematically distort the estimated effect size. Statistical and psychometric theories are able, if certain assumptions hold, to estimate the magnitude of these distortions and adjust the results to reflect what might have been found had an ideal research design been used. Such corrections have become standard practice in research on the validity of employee selection procedures, but are also broadly applicable to many areas of research (Wiernik & Dahlke 2020).

A ubiquitous artifact is sampling error, which refers to deviation of a sample statistic from the population value because the study is based on a limited sample of the population. Estimates of sampling error are central to statistical significance tests and confidence intervals, and also play a key role in meta-analysis. Sampling error is distinct from other artifacts considered in meta-analysis because it reflects unsystematic error and because its effect on an individual study cannot be known—there is no way to correct an individual study for sampling error. Fortunately, as long as a collection of samples is representative of the population of interest, sampling errors will not bias the effect size estimate; that is, the expected value of the sample effect size equals the population value. However, sampling error does increase the variance of effect sizes across studies. Consequently, a key step in meta-analysis is to estimate the magnitude of variance due to sampling error and to distinguish it from true heterogeneity across studies. This topic is discussed more fully below in the section titled Quantifying Heterogeneity.

Unlike sampling error, psychometric artifacts cause bias in the effect size estimate. Unreliability in measures tends to attenuate observed relationships, relative to what would be found if study variables were measured without error (Schmidt & Hunter 1996). Similarly, lower correlations

tend to be found when a continuous variable is forced into a small number of response levels, especially when a continuous variable has been artificially dichotomized into high and low groups (Aguinis et al. 2009, Schmidt & Hunter 2015). Range restriction occurs when study participants are selected in a way that limits the observed range of predictor and criterion scores, leading to underestimates of the relationship in the unrestricted population (Dahlke & Wiernik 2020, Sackett & Yang 2000). This is common in research on employee selection, where candidates with low scores on the predictor tend to be screened out by the selection process and therefore cannot be included in validation research. Each of these artifacts can be corrected by first estimating the impact of the artifact on the study results, and then applying a statistical correction formula to reverse this effect.

Of course, many other statistical artifacts can affect research results, but they are generally not correctable through meta-analytic procedures. Samples may not be representative of the population of interest for reasons beyond the selection mechanisms addressed by range restriction corrections. Measures may lack construct validity or be subject to systematic biases (e.g., criterion contamination, common method bias). A multitude of threats to internal validity can bias results in quasi-experimental and correlational research designs. Given the multitude of ways a study can go astray, it is useful to keep in mind that statistical adjustments are no substitute for good research design.

Note also that adjustments for psychometric artifacts generally increase the sampling error in corrected estimates. Even though corrections typically increase the magnitude of the effect size, they simultaneously widen the confidence interval (Oswald et al. 2015). Therefore, sampling error formulas designed for observed effect sizes will underestimate the sampling error variance of corrected effect sizes, and appropriate adjustments to study weights and confidence intervals must be applied when conducting psychometric meta-analysis (Schmidt & Hunter 2015). Methods exist to adjust the standard error and confidence intervals for corrected correlations (Bobko 1983, Fife et al. 2013, Li et al. 2013, Raju & Brand 2003, Schmidt & Hunter 2015, Wiernik & Dahlke 2020) and standardized mean differences (Wiernik & Dahlke 2020). A simple approximation is to first compute the confidence interval on the observed effect size and then apply the psychometric adjustment to the two endpoints (Schmidt & Hunter 2015). For more details on performing psychometric artifact corrections, see Schmidt & Hunter's (2015) seminal text on psychometric corrections and Wiernik & Dahlke (2020) for a broader summary of artifact corrections for correlations, observed group differences, and experimental effects.

Rationale for Psychometric Corrections

There are two main reasons to correct for statistical artifacts. First, when applied appropriately, corrected effect sizes reverse the biasing effect of psychometric artifacts and provide a better estimate of the true relationship (Schmidt & Hunter 2015, Wiernik & Dahlke 2020). Second, correcting for artifacts addresses artificial differences between studies in their research procedures, increasing the compatibility of results across studies (Sackett 2014).

Estimating unbiased effect size. Statistical artifacts are often described as producing bias in research findings. This is meant in a very particular way: Results obtained using a particular measure or sample do not, on average, estimate the effect size that would be obtained under ideal conditions that perfectly reflect the context about which the research desires to make inferences. As such, whether a particular design feature is considered a source of bias depends on the inference being made by the researcher. Different corrections, including no correction at all, all reflect reasonable estimates; they simply refer to different inferential spaces. An uncorrected sample

correlation for a particular measure is an unbiased estimate of the population correlation for that measure; however, it is a biased estimate of the correlation for a perfectly reliable measure.

This perspective on statistical artifacts has several implications. First, when determining what corrections to apply, careful attention should be paid to the inferences being made in a particular research context, and how the data informing the artifact corrections relate to that inference (LeBreton et al. 2017). The following discussion outlines some of these considerations. Second, it is often possible to frame research questions in multiple ways, and reasonable persons may disagree about how best to conceptualize a particular research question. Therefore, it is important to report research findings in a way that permits maximum flexibility in interpreting results in relation to alternate conceptualizations of the problem. At a minimum, researchers should always report uncorrected estimates alongside corrected values, and it can often be helpful to report results based on several alternate corrections related to different assumptions (Burke et al. 2014).

Correcting for inflated between-study variance. Correcting for statistical artifacts allows better estimates of the heterogeneity of a phenomenon by removing artificial differences due to variations in research methodology. If studies use different measures, samples, or research designs, the impact of these factors on effect size can artificially create the appearance of heterogeneity. For example, studies using a more reliable outcome measure would be expected to produce larger effect sizes than those using a less reliable measure. Without correcting for statistical artifacts, these differences might lead to the false conclusion that the relationship is situation specific (LeBreton et al. 2017, Murphy 2000, Schmidt et al. 1985), or they might be confused with substantiative moderators of the relationship.

Estimating operational effect size. The two goals of artifact correction do not always lead to the same choices. A researcher might want to account for differences in reliability when examining the variance of effect sizes, but in regard to the mean effect size, there might be greater interest in the uncorrected correlation for an operational measure. One approach to this inconsistency is to first conduct the meta-analysis using the fully corrected estimates and then, in a second step, to estimate the operational validity by reapplying the attenuating effect based on the average level of the artifact, effectively reversing the artifact correction (Schmidt & Hunter 2015). This approach also provides flexibility to compute operational validities for different contexts with different levels of the artifact (e.g., adjusting for alternate reliabilities for different outcome measures or contexts).

Concerns About Psychometric Artifact Corrections

Since the early days of meta-analysis, there have been debates about the appropriateness of psychometric corrections (Schmidt et al. 1985). Concerns about these methods can be organized into three broad types. First, a correction may lack adequate conceptual justification. A meta-analysis should always provide a clear and explicit rationale for the specific corrections linked to the relevant inference in a particular setting (Le et al. 2009). Second, all statistical corrections involve assumptions about both the nature of the mechanism underlying the artifact and the distribution of artifacts across studies. Before applying corrections, researchers should be aware of these assumptions, gather evidence on potential violations, and consult research on the accuracy of corrected effect sizes when violations occur. A third type of critique that has been leveled against psychometric corrections is based on the specific values of the artifacts used in the corrections. Most meta-analyses rely on published research reports, which often do not provide the information needed for statistical corrections. Obtaining plausible values for artifacts often involves going to sources beyond the pool of studies included in the meta-analysis, and the relevance of the external sources may be brought into question.

Before applying an artifact correction, meta-analysts should ask themselves the following questions:

1. Does the logic of the artifact correction match the nature of the statistical artifact in a particular setting and the inferences made by the researcher?
2. Are the assumptions of the statistical correction reasonable in a particular context? If not, how might the violation of assumptions affect the accuracy of the correction?
3. Are the values used for the artifact correction representative of the studies to which they are applied?

The following sections examine these questions in the context of the most common psychometric corrections, namely measurement error and range restriction, and provide recommendations for determining the most appropriate corrections when conducting a meta-analysis.

Measurement Error

From the perspective of classical test theory, measurement error refers to random fluctuations in a person's score on a measure. This random noise will increase the variance of scores and attenuate estimates of the correlation. Corrections for measurement error estimate the degree of attenuation that is expected due to the reliability of a measure, and adjust the correlation to compensate.

According to classical test theory, the true score of an individual represents the expected value or average of scores across an infinite number of parallel measurements. Let $\sigma^2_{T_X}$ represent the variance of these true scores across persons. Measurement errors reflect the deviation of a particular measurement from this true score, and the variance of these errors is represented by σ^2_e. The reliability of a measure, ρ_{XX}, is defined as the proportion of the variance in observed scores that is due to the true score variance. From these definitions, we see that the variance of observed scores, σ^2_X, is a function of the true score variance and the reliability:

$$\sigma^2_X = \frac{\sigma^2_{T_X}}{\rho_{XX}}. \qquad 1.$$

Because reliability is always less than or equal to one, this means that the variance of observed scores will always be greater than or equal to the variance of true scores. Furthermore, the covariance between two measures is assumed to be due entirely to their true scores, and is unaffected by measurement error. Thus, measurement error increases the variance but not the covariance, thereby attenuating the observed correlation relative to the correlation between true scores:

$$\rho_{XY} = \frac{\mathrm{COV}_{XY}}{\sigma_X \sigma_Y} = \frac{\mathrm{COV}_{XY}}{\left(\frac{\sigma_{T_X}}{\sqrt{\rho_{XX}}}\right)\left(\frac{\sigma_{T_Y}}{\sqrt{\rho_{YY}}}\right)} = \sqrt{\rho_{XX}}\sqrt{\rho_{YY}}\,\rho_{T_X T_Y}. \qquad 2.$$

Correction for unreliability is achieved by reversing the attenuation formula. The correction can be applied for measurement error in the predictor, the criterion, or both. The fully corrected correlation is as follows:

$$r_{T_X T_Y} = \frac{r_{XY}}{\sqrt{r_{XX}}\sqrt{r_{YY}}}. \qquad 3.$$

There are a couple of useful ways to interpret corrected correlations. First, they represent the correlation between true scores, that is, the correlation that would be obtained if a hypothetical perfectly reliable measure were used (e.g., a test with an infinite number of items), rather than the operational measure available in a particular study. The true score correlation should not be confused with the correlation between theoretical constructs; it is simply the correlation for a

longer version of the same measure. Any systematic bias or lack of construct validity in the measure becomes part of the true score.

Whether estimates of true score correlations are appropriate depends on the inference being drawn. If the researcher is interested in the relationship among constructs, the particular measure is a nuisance factor, and estimating the relationship for a perfectly reliable measure is a useful piece of information. On the other hand, if the inference involves the measure itself (e.g., Do interviewer ratings of conscientiousness predict attendance?), then the errors inherent in the predictor measure are part of what is being evaluated and correction would be inappropriate. In research on the validity of personnel selection methods, it is common practice to correct for criterion reliability but not predictor reliability (SIOP 2018).

The second way to conceptualize the corrected correlation is that it standardizes the relationship in a different metric. The observed correlation is standardized relative to the SD of observed scores, while the corrected correlation is standardized relative to the SD of true scores. If two studies use different measures with substantially different levels of reliability, measurement error will inflate the variance of one study more than the other, and the resulting observed correlations will not be standardized on a common metric. All else being equal, if the two measures can be considered parallel, the true score SD will provide a consistent basis for standardization across studies. From this perspective, the benefit of the corrected correlation is not that it represents the true relationship but rather that it provides a common metric for combining studies with measures of differing quality.

Reliability corrections can also be applied to mean differences (Schmidt & Hunter 2015, Wiernik & Dahlke 2020). Correcting the observed standardized mean difference (d_O) for unreliability in the outcome variable yields the standardized mean difference on a perfectly reliable measure of the outcome:

$$d_{T_Y} = \frac{d_O}{\sqrt{r_{YY}}}. \qquad 4.$$

Conceptual justification. Reliability corrections are based on classical test theory, which treats measurement errors as random. However, in reality, measurement errors are not simply random noise but rather refer to a myriad of factors other than the construct of interest that influence an individual's responses to items on an assessment (DeShon 2003). Individual items on a measure capture both construct-relevant and construct-irrelevant individual differences, and the particular collection of items included on a measure will have some impact on the total score. Scores may also fluctuate from day to day based on personal contextual factors (e.g., mood, fatigue). Furthermore, idiosyncratic policies by raters may influence the scores they provide. To the extent that such factors are specific to a measurement event, they are expected to vary across repeated measurements and are typically considered sources of error variance in meta-analysis. For a more extensive discussion of sources of measurement error, see Highhouse et al. (2015, pp. 98–102) and Schmidt et al. (2003).

Corrections for reliability should be informed by the relevant sources of measurement error in a particular context (Le et al. 2009, Oswald & McCloy 2003, Schmidt et al. 2003). For example, estimates of internal consistency (e.g., coefficient alpha) represent error associated with sampling items from a content domain, but they do not account for transient error, that is, nonsystematic fluctuations in scores over time (Schmidt et al. 2003). Thus, an important question in selecting a reliability estimate is whether temporal instability is viewed as measurement error in a particular context. If the focal construct is expected to change rapidly (e.g., mood), fluctuations over time might be viewed as construct-relevant variance and could be the primary focus of a study. Conversely, if inferences are made about stable individual differences, variance across time would

be a source of measurement error, and a measure of reliability that reflects transient error would be preferred. Failure to carefully articulate which sources of measurement error are relevant to inferences can easily lead to over- or undercorrection.

When a measure involves subjective judgments by a rater, such as supervisor ratings of job performance, idiosyncratic rater effects are often a major source of measurement error (O'Neill et al. 2015). Interrater reliability corrections require that the rater differences used to estimate reliability be representative of the ratings used to calculate the effect size. Typically, validity coefficients are calculated using ratings from the direct supervisor. A challenge in validation research is that it is rare to have data from two supervisors that can reasonably be considered parallel assessments of job performance. If the two raters used to estimate reliability are not equally knowledgeable about the employee's performance (e.g., one is a direct and the other an indirect supervisor), then the correlation between their ratings might not be an accurate measure of the reliability of ratings by the direct supervisor (Murphy & DeShon 2000).

The most common measure of reliability is coefficient alpha, which is a measure of internal consistency. This is a reasonable estimate of reliability when the construct being measured is unidimensional, and items are treated as interchangeable indicators of that construct. However, if the domain is multidimensional, coefficient alpha will often underestimate and misrepresent the reliability of a measure (Cortina 1993, Cortina et al. 2020). In still other cases, measures are best conceptualized by a formative model (MacKenzie et al. 2005; compare with Edwards 2011) where each item reflects a unique conceptually derived facet of a composite construct. A good example is a multidimensional measure of job performance, where the dimensions are determined from a job analysis with only one item per dimension. In such cases, the degree of correlation among dimensions is not a useful measure of reliability (in fact, it might be an indication of halo error) and internal consistency should not be used to correct effect sizes.

As these examples illustrate, correcting for reliably requires considerable judgment on the part of the researcher. When applying reliability corrections, the write-up of the methodology should be explicit regarding the types of reliability estimates used in a meta-analysis and how these relate to the relevant sources of measurement error in the particular research domain. Mixing of different types of reliability indices should be avoided, although sensitivity analyses comparing the impact of different reliability estimates may be informative.

Assumptions of reliability corrections. Measurement error corrections follow the assumptions of classical test theory, which essentially assumes that measurement errors are random. Particularly important is the assumption that measurement errors on one variable are independent of true scores and measurement errors on other variables. While this is often a reasonable assumption, there may be situations where it is questionable. For example, transient errors in mood might have a shared effect on multiple affectively laden self-report measures (Le et al. 2009). Similarly, inconsistences among raters might reflect unique information known to a particular rater and not just random error (Hoffman & Woehr 2009, Lance et al. 2008, Murphy & DeShon 2000, O'Neill et al. 2015). To the extent that unique rater perspectives relate to personal or contextual characteristics (Tziner et al. 2005), idiosyncratic rater effects may correlate with true scores on other variables. When these assumptions are violated, greater measurement error will not necessarily result in lower correlations among observed variables, and standard formulas may result in an overcorrection (Nimon et al. 2012, Putka et al. 2014, Zimmerman 2007).

Another assumption is that measurement error associated with one person is unrelated to the error for other individuals. However, it is common for data on job performance to be collected from supervisors who each rate multiple subordinates. To the extent that idiosyncratic rater tendencies (e.g., leniency) have a common effect across a rater's subordinates, the measurement errors on observations clustered within raters will not be independent (Ellington et al. 2021).

Appropriateness of reliability estimates. The optimal way to estimate reliability would be to curate an appropriate estimate of reliability for each study. Estimates of reliability are sometimes available from the same studies providing effect size estimates. More often, reliability information must be borrowed from external sources, such as scale development research or test manuals. When using external estimates of reliability, efforts should be made to identify estimates from contexts and populations similar to the studies in the meta-analysis.

Psychometricians have long argued that reliability is a product not only of the test but also of the context and examinee population. Reliability generalization studies often demonstrate considerable variability in reliability across settings (e.g., Salgado & Moscoso 2019). Therefore, identifying a context-appropriate estimate of reliability is critical for accurate artifact correction. For example, the reliability of job performance tends to be higher for less complex jobs (Conway & Huffcutt 1997), perhaps because performance in these contexts is easier to define and observe. Reliabilities also tend to be higher when collected for research rather than administrative purposes (Salgado & Moscoso 2019). The population on which reliability is estimated can also make a difference. For example, observed interrater reliability estimated from samples with restricted variance may underestimate the reliability of ratings collected on broader populations (LeBreton et al. 2003, Murphy & DeShon 2000, Sackett et al. 2002).

Arguably, the most relevant values come directly from the studies included in the meta-analysis. However, like any statistic, small-sample estimates of reliability are subject to considerable sampling error, which will increase uncertainty in corrected estimates (Raju et al. 1991). For this reason, Viswesvaran et al. (2014) recommend using values from reliability generalization research when available. Both perspectives have merit, and the best estimate for a particular situation will involve balancing the competing concerns of relevance and precision.

LeBreton et al. (2014) criticized the overreliance on corrections based on meta-analytic estimates of reliability, such as the widely used value of 0.52 for job performance (Viswesvaran et al. 1996). Such low reliability would cause many psychometricians to question the quality of the measure, raising doubts about whether statistical adjustments can overcome the limitations inherent in such a low-quality measure. LeBreton et al. demonstrate that corrections based on the 0.52 value can readily lead to implausible results, where a few predictors can explain nearly all or even more than 100% of the variance in job performance. There is certainly reason to be skeptical of the large corrections often achieved in validity generalization research. Nevertheless, many of the specific arguments in the LeBreton et al. critique have been challenged by other researchers, who question the data and assumptions used for the demonstration of implausible variance explained and note that single-rater reliabilities of this magnitude are common for ratings in many domains besides job performance (Shen et al. 2014, Viswesvaran et al. 2014).

Range Restriction

Range restriction refers to systematic bias in selection of data points for inclusion in the analysis, which generally leads to underestimation of relationships. Range restriction has been examined largely in the context of correlation coefficients and, in particular, in the validation of assessments used for employee selection. However, even though range restriction effects in selection are relatively obvious, the logic of range restriction can be applied to other effect sizes and other contexts as well (Bobko et al. 2001, Dahlke & Wiernik 2020, Fife et al. 2020).

Two factors that can lead to restriction of range when validating an employment test are applicant pool screening and employee selection. These have different implications for estimating operational validity. Applicant pool screening refers to processes that affect the composition of the applicant pool, such as employees self-selecting into jobs commensurate with their qualifications or employers prescreening applicants through minimum qualification requirements or

assessments administered during earlier stages of the selection process. Consider a study validating a job knowledge test for hiring computer programmers. Individuals who know little about programming are unlikely to apply for such a job; therefore, the applicant pool will be skewed toward higher-knowledge individuals. This, however, is the relevant population of applicants from which the organization will be selecting, so this type of range restriction would not be a source of bias. The resulting correlation might be lower than if the test were evaluated on the general population, but in practice it is this restricted pool of applicants about whom inference will be drawn.

Range restriction can also occur through employee selection. Say 100 applicants are given the job knowledge test and 25 of them with the highest scores are hired. A year later, a measure of job performance on those 25 employees is obtained, and the correlation between test scores and job performance ratings is calculated. The validation study cannot include those with low knowledge because they were not hired and their performance cannot be measured. The sample on which the correlation is computed has a substantially shrunken range of scores on the predictor, relative to the relevant applicant population, which results in underestimation of the correlation that would have been obtained if the full range of the applicant pool were included in the study.

The scenario described above is referred to as direct range restriction, where the test being validated is used to make selection decisions. It is also possible for range restriction to be incidental, where the validation sample is selected on the basis of factors distinct from but correlated with the test being validated.

In most applied settings, range restriction is incidental. First, selection decisions are rarely made using a single assessment, so the focal test is at best only a component of the actual selection process. Second, direct range restriction is relevant only in a predictive validation design, where applicants are assessed and then their performance is measured at a later point in time (Barrett et al. 1981). A more common approach to validation is the concurrent design, where tests are administered to existing employees and a measure of performance is obtained at approximately the same time. In a concurrent design, the employees will often have been selected through varying and unknown procedures. Their selection might have included factors related to the focal test, but it is unlikely that the test being validated was the basis for their selection. Thus, direct range restriction is likely to be rare.

Until fairly recently, most meta-analyses that corrected for range restriction utilized formulas for direct range restriction, despite the fact that range restriction is almost always incidental. Although corrections for incidental range restriction have existed for decades, early versions required information on correlations with the variable used for selection (Sackett & Yang 2000), and these correlations are rarely known in applied settings. A breakthrough came when Hunter et al. (2006) developed a procedure to estimate and correct for incidental range restriction when the selection variable is unknown. This procedure provides more accurate corrections than the direct range restriction formula and has subsequently been refined (Dahlke & Wiernik 2020, Le et al. 2016). More accurate methods exist when raw data or full correlation matrices are available (Fife et al. 2016), but such data will typically not be available for the majority of studies in a meta-analysis. Incidental range restriction corrections for standardized mean differences are also available (Fife et al. 2020, Li 2015).

Note that corrections for multiple artifacts (e.g., both unreliability and range restriction) require careful attention to the sequencing of the corrections and the impact of artifacts on one another, for example, the impact of range restriction on local reliability estimates (Brown et al. 2017). See Schmidt & Hunter (2015) for a detailed discussion of the procedures for correcting multiple artifacts.

Assumptions of range restriction corrections. Range restriction correlations assume that both the observed variables and the criterion used for selection are linearly related and homoscedastic (i.e., that prediction errors have constant variance across all levels of the predictor). Although many predictors used in selection show linear and homoscedastic relationships with job performance (Brown et al. 1988, Coward & Sackett 1990, Robie & Ryan 1999), there is no guarantee this will be true for all variables. The conditions under which violating this assumption will lead to inaccurate estimates are discussed by Gross (1982). Even when the assumptions are violated, corrected estimates will often be more accurate than uncorrected correlations, although the degree of accuracy depends on the form of the relationships in a particular data set (Gross & Fleischman 1983). Oswald & Johnson (1998) found that meta-analytic estimates of the mean corrected correlation were fairly robust to several patterns of nonlinearity and heteroscedasticity. Alternative methods for range restriction correction may be more effective when these assumptions are believed to be violated (Culpepper 2016), but they require access to raw data, limiting their application to meta-analysis.

Additional assumptions are embedded within particular correction formulas. For example, the incidental range restriction correction recommended by Hunter et al. (2006) assumes that the relationship between the unobserved selection variables and the outcome is fully mediated by the predictor, in other words, that selection is not based on any job-relevant characteristics other than what is measured by the focal predictor (Fife et al. 2013). Alternate models have been developed without this assumption (Dahlke & Wiernik 2020, Le et al. 2016), but these require a separate estimate of the degree of restriction on both predictor and outcome variables, limiting the settings where sufficient data will be available to allow their application. Another assumption of incidental range restriction models is that measurement errors on the predictor are independent of the selection variable, which may be unreasonable in situations where selection decisions are made on a composite that includes the focal predictor (Beatty et al. 2014). Although the Hunter et al. method is not optimal when these assumptions have been violated, still it often provides a better estimate of operational validity than other alternatives (i.e., either no correction or assuming direct range restriction) and, therefore, is recommended when the data for more advanced corrections are unavailable.

Estimates of range restriction. Obtaining good data on the degree of range restriction can be challenging, especially when the meta-analyst must rely on data provided in published research reports. Range restriction occurs because criterion data are available only for those candidates who have been hired and, thus, estimates of predictor–criterion correlations are based on a sample that is often unrepresentative of the population about which we want to draw inferences (i.e., job applicants). The degree of restriction is operationalized in terms of the ratio of unrestricted to restricted SD on the predictor, referred to as the *U* ratio.

To estimate the *U* ratio, we require the SD of predictor scores from both the sample used in the validation study (restricted SD_x) and the applicant pool (unrestricted SD_x). A common challenge is that, while the restricted SD_x is typically available from research reports, the unrestricted SD_x is often unknown, making it impossible to apply corrections at the individual-study level.

For example, in a concurrent validation design, only incumbents are administered the test being validated. The data do not include scores on a broader pool that includes those who were not selected. In such cases, several strategies have been developed to provide indirect estimates of *U* (Dahlke & Wiernik 2020, Sackett et al. 2022). None of these methods are perfect, and different methods do not necessarily yield comparable results. Thus, it is important for researchers to carefully consider the most appropriate reference population for a particular context (Carretta & Ree 2022).

Sackett et al. (2022) provide a thorough discussion and critique of strategies for estimating the unrestricted distribution of the application population in validation research. One approach would be to derive the distribution of unrestricted SDs from a subset of studies where applicant data are available. These estimates could be used as the basis for a distribution-based correction or to impute values for studies where this information is missing. This option is attractive because the correction factor is derived from the actual studies included in the meta-analysis and therefore is likely to reflect the nature of applicant pools in the specific context being studied.

This approach, however, has a major limitation (Sackett et al. 2022). It must be assumed that the set of studies providing applicant data are representative of the full collection of studies. This assumption is unlikely to be true because applicant data are most likely to be available in predictive designs, and in such cases, selection will typically be at least partly based on the test being validated. This is likely to produce more extreme range restriction than concurrent validation studies where the test was not used in selection (Sackett & Yang 2000). Indeed, Sackett et al. demonstrate that widely used values for *U* ratios represent levels of range restriction that are quite plausible under direct range restriction but unlikely to occur when range restriction is incidental. Consequently, estimating *U* ratios from predictive studies may overestimate the degree of range restriction in concurrent designs, which are likely to make up many of the studies with missing applicant data. The results will be a substantial overcorrection when applied to the full set of studies.

While it is important to consider the representativeness of the data used for artifact corrections, it is not clear that predictive and concurrent designs will necessarily produce substantially different levels of range restriction. Sackett et al. (2022) correctly note that more-extreme *U* ratios will be obtained under direct than incidental range restriction. However, even in predictive designs, selection decisions are rarely made using a single predictor. Instead, multiple factors will be combined in some way to form a (possibly implicit) measure of applicant suitability, and it is this suitability variable that is used for selection. The focal predictor will be correlated with but not identical to the selection variable, and this is a form of incidental range restriction (Beatty et al. 2014, Sackett et al. 2007). Additional work is needed to fully understand the representativeness of range restriction estimates and optimal correction procedures under typical conditions.

Another strategy for estimating the unrestricted population is to use test norms. When the predictor is a well-established measure, the SD reported in published test norms for the general population might be used to represent the unrestricted applicant population. This value could then be compared with the observed SD from each validation study to obtain a study-specific *U* ratio. This method has the advantage of estimating *U* for each study, rather than generalizing *U* from a subset where complete data were available. A potential limitation of this approach is that, due to self-selection and applicant prescreening, applicant pools tend to be more homogeneous than the general population (Sackett & Ostgaard 1994). Ones & Viswesvaran (2003) found that SDs from job-specific applicant pools were, on average, only around 4% smaller than norm data for personality tests, suggesting that any distortion due to the use of test norms may be minimal. Nevertheless, researchers are advised to test this assumption by examining applicant pool SDs for studies where this information is available.

Several other strategies for estimating *U* ratios exist (Dahlke & Wiernik 2020, Sackett et al. 2022). Of these, the most problematic is the practice of borrowing estimates from published meta-analyses and applying them to an entirely different setting. Because the mechanisms producing range restriction are likely to be context specific, this approach should be avoided.

A common theme in these critiques is that researchers should take care to curate artifact values that are representative of the particular research context. Additionally, a pool of studies may include settings with quite different selection mechanisms, and the artifact correction factor should

be adjusted to reflect the value most appropriate to each study (Berry et al. 2007, Huffcutt et al. 2014). In the absence of a strong justification for the artifact distribution, the researcher is left with a choice between (*a*) no correction, which would underestimate the true effect size, and (*b*) a correction with unknown accuracy, which could produce overestimates. Sackett et al. (2022) advocate for no correction in these settings, because underestimation is likely to be slight in concurrent designs, which typically make up the bulk of available validity studies. Alternatively, the best solution may be to apply multiple corrections based on different assumptions and to explicitly discuss the sensitivity of results to the choices regarding artifact values (e.g., Sackett et al. 2021).

Individual- Versus Distribution-Based Correction

Schmidt & Hunter (2015) describe two approaches for correcting psychometric artifacts in a meta-analysis. First, the researcher can correct each effect size individually for study-specific artifact values, and then conduct the meta-analysis on the collection of corrected values. This approach provides the most relevant correction for each study but requires that artifact information be available for all studies. Missing artifact values can be imputed using the average of the available information (Raju et al. 1991).

Second, distribution-based correction estimates the mean and variance of the artifacts and applies the correction at the aggregate level. The meta-analysis is conducted on raw effect sizes, and the mean and SD of effect sizes are adjusted according to the artifact distribution. The distribution-based approach enables correction for artifacts in contexts where individual-based correction is not feasible due to lack of reported information. Importantly, artifact distributions can be estimated using data from beyond the set of studies included in the meta-analysis. For example, reliability generalization studies have provided estimates of the distribution of reliabilities for many established measures, and the results of these studies may provide a reasonable basis for an artifact distribution for meta-analyses examining those variables.

When corrections are based on an artifact distribution, it is particularly important that the meta-analysis present a strong justification that the artifact distribution is representative of the context and sample of studies included in the meta-analysis. As noted in the above discussions of measurement error and range restriction corrections, meta-analyses have been criticized for uncritically adopting questionable values for psychometric artifacts (LeBreton et al. 2014, Sackett et al. 2022).

Distribution-based artifact corrections assume that artifacts are independent of other artifacts and the effect size. James et al. (1992) suggest that this assumption could be violated because contextual factors that impact the variability of scores (e.g., climate strength) could affect both the effect size and the reliability of measures. Although there is limited research on the topic, the available empirical evidence suggests the possibility of nonindependent artifacts. Schmidt & Hunter (2015) note that incidental range restriction and unreliability effects are not independent, but this can be accounted for through use of their interactive approach to distribution-based meta-analysis. Other research has documented correlations across studies between predictor and criterion reliability (Köhler et al. 2015) and between effect size and reliability (Yuan et al. 2020). Studies have demonstrated that the presence of correlated artifacts is particularly problematic for estimates of the variance of effect sizes (Raju et al. 1998), especially when artifacts are correlated with the true effect size.

One way to deal with correlated artifacts is to apply individual-study corrections, as long as partial study-level artifact information is available, and substitute the mean for unknown artifact values when necessary (Raju et al. 1991). Alternatively, James et al. (1992) describe a distribution-based artifact corrections model that accounts for correlations among artifacts (compare with Thomas & Raju 2004), but this approach has not been widely adopted.

META-ANALYSIS MODELS

The primary goal of meta-analysis is to estimate the mean and variance of a collection of effect sizes. The model for a meta-analysis, in its general form, is as follows:

$$\mathrm{ES}_j = \theta_j + e_j. \qquad 5.$$

In this equation, the observed effect size computed in a particular sample, ES_j, is a function of a study-specific population effect size, θ_j, and sampling error, e_j. In this model, there are two potential sources of variance in effect size. First, a particular study may differ from others because of study-specific choices in how study j was conducted (e.g., the specific manipulation, context, and measures) that cause study j to differ from other study designs. These differences are reflected by the study-specific effect, θ_j. The second source of variance, σ_e^2, results from differences due to the sample of individuals included in a particular study, that is, differences that would occur if the study design were replicated exactly on a new sample of participants. As in classical statistical analysis, the variance due to sampling of participants is heavily dependent on sample size, with sampling errors becoming smaller as sample size increases. In contrast, the variance of study-specific effects is independent of sample size.

A meta-analysis will typically include an estimate of the mean effect size, M_θ, and may also include a measure of the variability of the study-specific effects (i.e., the variance of θ_j across studies), operationalized in terms of the between-study variance (e.g., τ^2) or SD (e.g., τ or SD_θ). Several statistics for summarizing and interpreting the between-study variance are discussed in the section titled Quantifying Heterogeneity, below.

A variety of analytic models have been developed to estimate the mean and variance of effect sizes (Schulze 2004, Viechtbauer 2005). Specifying a model involves several choices that should be clearly documented. Too often, research reports use brief descriptors (e.g., "random effects" or "Schmidt–Hunter") that leave some ambiguity about the nuances of the approach. Increasingly, software used for meta-analysis allows considerable flexibility in specifying the model structure and estimation method. Transparency and replicability demand that research reports provide clear and specific explanation and justification for these choices. Some modeling options are described next.

Transformations

Some meta-analysis methods involve a transformation of effect sizes before the statistical analysis is conducted. For example, when analyzing dichotomous outcomes, it is common to take a natural log transformation of the OR (Haddock et al. 1998). When analyzing correlations, some meta-analysis methods apply the Fisher r-to-z transformation before conducing the analysis, while others conduct the meta-analysis directly on the correlation coefficient. Research in the context of validity generalization has consistently shown that the r-to-z transformation leads to an upward bias, especially when there is considerable variability in correlations across studies (Schmidt & Hunter 2015, Schulze 2004). Therefore, for correlations, analyzing the effect size in the original metric is preferred.

Fixed- Versus Random-Effects Model

In much of the meta-analysis literature, the decision to use a fixed- versus random-effects model is presented as a dichotomous choice. However, there are actually several interrelated choices involved. Consequently, simply labeling an analysis as fixed or random is often insufficient to fully specify the model.

The fundamental difference between a fixed- and a random-effects model refers to the study-specific effects (θ_j in Equation 5) (Borenstein et al. 2010, Field 2001, Hedges & Vevea 1998). In a

random-effects model, the study-specific effects observed in a particular meta-analysis are viewed as a random sample from a population of possible studies. In a fixed-effects model, the study-specific effects take on a limited set of values that define the population to which inferences will be generalized.

Most often, the fixed-effects model is equated with a common effect model, which assumes that the study-specific effects are nonexistent (making the subscript in θ_j unnecessary). In the common effect model, all studies estimate a common population effect size and all variability among effect sizes is due to sampling error. Because most meta-analyses produce evidence of nontrivial variance beyond sampling error (Schmidt et al. 2009), random-effects models are generally recommended in organizational research. A fixed-effects model can also account for differences among several specified populations by including study-level covariates (e.g., participant subgroups or study design features) in a meta-regression. Random-effects models, in contrast, allow for unknown and unmodeled differences among studies. Rather than assigning effect sizes to known subgroups, the meta-analysis estimates the variance of study-specific effects (τ^2).

A mixed-effects model (Hedges 1992) combines these two approaches, including separate estimates for distinct subgroups of studies (fixed effects) while also estimating the between-study variance within each of these subgroups (random effects). Meta-regression (Gonzalez-Mulé & Aguinis 2018, Tipton et al. 2019) is an example of a mixed-effects model, where the regression coefficients for study-level moderators capture the fixed effects and the residual variance reflects additional study-specific random effects that are not explained by the predictors in the model.

Because random-effects models account for two sources of variance (sampling of studies and sampling of participants), confidence intervals on the mean effect size tend to be wider in random-effects than in fixed-effects models (Erez et al. 1996). When substantial between-study variance exists and the number of studies is small, the random-effects confidence interval can be substantially wider than the fixed-effects version. Given that sizable between-study variance is common organizational research (Schmidt et al. 2009), using a fixed-effects model will often overstate the degree of certainty in the mean estimate.

The random-effects model assumes that the pool of studies included in a meta-analysis represents a random sample from a larger population of possible studies. This is an improvement over the common effect model, which makes the unrealistic assumption that there are no real differences across studies (i.e., that all observed differences are due to sampling error). Still, some researchers have criticized the random-effects assumption as equally unrealistic (Bonett 2008, 2009). The idea that there exists a well-defined superpopulation of potential studies and that individual studies are randomly sampled from this population does not match well with the incremental processes through which research is conducted and published or the processes through which studies are identified for inclusion in a meta-analysis. In defense of the random-effects model, Borenstein (2019) argues that the pool of studies in a meta-analysis can be conceptualized as a sample from some population, even if we are not able to fully specify the nature of that population. As long as we remain mindful of this ambiguity, the random-effects model provides a useful way to understand the inconsistencies in a pool of studies.

Another criticism of the random-effects model involves the assumed form of the superpopulation of studies. Most models for random effects have been developed under the assumption that the population of potential effect sizes has a normal distribution, which allows the distribution of effect size to be represented parsimoniously through the mean and SD of this distribution. Technically, other distributions could also be used as the basis for a random-effects model (Burr & Doss 2005, Lee & Thompson 2008), but most available random-effects methods assume a normal distribution. The assumption that distribution of effect sizes is normal is rarely tested and may

be difficult to justify. Fortunately, inferences about the mean effect size are fairly robust to typical levels of nonnormality in the distribution of effect sizes (Rubio-Aparicio et al. 2018).

Several different methods have been developed to estimate the variance of random effects. A major distinction is whether the between-study random variance component is estimated along with other model parameters or in a secondary step. Two-stage random-effects estimation is exemplified by the Schmidt–Hunter approach, which starts by estimating a weighted mean and observed variance of effect sizes, and then uses the observed variance minus artifact variance as the estimate of the between-study variance. In this approach, the between-study variance does not inform the mean effect size estimate, although it is used in calculating confidence and credibility intervals.

Other random-effects models [e.g., restricted maximum likelihood (REML), DerSimonian–Laird] estimate all parameters concurrently. This is most apparent in the assignment of study weights, where the random variance component influences the study weights and therefore can alter the mean effect size or coefficients in a meta-regression. Choices in defining study weights are discussed in more detail below. Within the concurrent estimation approach, a number of alternate estimators have been developed (Langan et al. 2019, Schulze 2004, Veroniki et al. 2016, Viechtbauer 2005). Although all of the methods tend to provide similar results, the Schmidt & Hunter (2015) approach has a tendency to underestimate the between-study variance (Viechtbauer 2005). One method that has been found to consistently perform well is REML.

In early meta-analytic work, an important advantage of the Schmidt & Hunter (2015) approach was its ability to apply corrections for psychometric artifacts, an issue that has been ignored in much of the research comparing alternate estimation methods. The small gains in accuracy from using alternate estimators tend to be dwarfed by the substantial differences resulting from psychometric corrections (Hall & Brannick 2002). Consequently, the Schmidt–Hunter method remains the dominant approach in contexts where artifact corrections are desired. However, this situation may be changing, as new models increasingly allow the incorporation of artifact corrections with other estimators of the mean and variance of effect sizes. For example, Brannick et al. (2019a) demonstrate the effectiveness of random-effects REML estimation with correlations corrected individually for criterion unreliability. Additionally, Raju & Drasgow (2003) developed maximum likelihood estimators for correlations with both individual- and distribution-based artifact correction. For distribution-based artifact corrections, the Schmidt–Hunter interactive model has been well supported (Law et al. 1994) and remains the most common approach.

Study Weights

Meta-analysis generally applies study weights when computing the mean and variance of effect sizes. By giving more weight to studies with more precise effect sizes, one can estimate the overall mean effect size with greater precision. Study weights have been operationalized in a variety of ways.

Inverse sampling error weights. A common approach in fixed-effects meta-analysis is to define weights based on the sampling error variance of the effect size estimate. If the studies all estimate a common population effect size, weighting by the inverse of the sampling error ($w_j = 1/\sigma^2_{e_j}$) provides the most efficient estimate of the mean effect size (Hedges 1982). Studies with smaller sampling error variance will, on average, fall closer to the population value, and by giving more weight to these studies, the weighted average becomes more precise.

One complication in the use of inverse sampling error weights is that the sampling error variance is itself a function of the population effect size. For example, the sampling error variance for

the correlation coefficient is typically estimated as

$$\sigma_e^2 = \frac{\left(1-\rho^2\right)^2}{N-1}, \quad 6.$$

where ρ is the population correlation and N is the sample size. Similarly, the sampling error variance for the standardized mean difference is approximately

$$\sigma_e^2 = \frac{n_1+n_2}{n_1 n_2} + \frac{\delta^2}{2\left(n_1+n_2\right)}, \quad 7.$$

where δ is the population standardized mean difference, and n_1 and n_2 are the sample sizes in the two groups being compared. This dependence of sampling error variance on the effect size is problematic because the population effect size is unknown, and values from an individual study are often estimated with considerable uncertainty. Using the sample effect size from each study in these formulas introduces another source of error into the analysis. That is, when computing a weighted average, sampling error affects not only the sample effect size estimates but also the weight that is given to each sample. Sampling error in the weights decreases the accuracy of the meta-analysis, and because the weights are correlated with the effect size estimate, this introduces bias into the weighted average.

Several solutions have been adopted to address this issue. Schmidt & Hunter (2015) avoid the issue by using only sample size in the weights (this method is described more fully in the section titled Sample Size Weights, below). Hedges & Olkin (1985) recommended a Fisher *r*-to-*z* transformation for analysis of correlations, which removes the correlation parameter from the sampling error variance estimate. However, the transformation also introduces other biases that tend to outweigh this benefit (Schulze 2004). A third strategy is to replace the sample-specific estimate of the population effect size with the average value calculated across studies. Because the same average value is used for all studies, sampling error in the effect size estimate does not influence differential weights of studies, and this method generally improves the accuracy and precision of meta-analytic results (Aguinis 2001, Brannick et al. 2019a, Van Den Noortgate & Onghena 2003). Such analyses are conducted in two stages. First, an initial estimate of the effect size is obtained using a sample size–weighted average. Next, this mean effect size is used to calculate the sampling error variance, which can be used to compute inverse variance weights or random-effects weights.

Sample size weights. An inspection of the sampling error variance formulas in the preceding section reveals that, if the effect size is treated as a constant, the only study-specific feature of inverse variance weights is the sample size. Therefore, for constant population effect size, sample size weights will be equivalent to inverse sampling error weights for correlations, and will provide a close approximation of the standardized mean difference (Schmidt & Hunter 2015). Thus, the common practice of weighting effect size by sample size is equivalent to using inverse sampling error weights.

While sample size weights provide a simple and effective option for hand calculations, especially in the analysis of correlations, there are several advantages of using the sampling error variance rather than sample size. First, for effect sizes other than correlations, the overall sample size will often only approximate the inverse sampling error weights. For example, in the analysis of the standardized mean difference, Equation 7 accounts for unequal sample size between groups, which will be more accurate than weights based on the overall *N*. The difference between the two approaches is often small, but because the calculations can be readily coded into a spreadsheet or analysis software, there is no reason not to use the most accurate estimate available. Second, using the sampling error variance provides a straightforward approach to account to research

design characteristics. For example, effect sizes computed from repeated-measures designs often have substantially smaller sampling error variance than those estimated from a between-groups design with the same sample size (Morris 2008, Morris & DeShon 2002). A final advantage of using sampling error variance rather than sample size is that these values can be incorporated into random-effects weights, as I discuss in the next section.

Random-effects weights. Random-effects models (Borenstein et al. 2010, Hedges & Olkin 1985, Hedges & Vevea 1998) recognize two types of sampling involved in a meta-analysis. First, each study design is to some extent unique, because the researcher makes a variety of idiosyncratic choices in the study design. Thus, each study can be viewed as a random sample from a population of potential study designs. Second, participants in each study are randomly sampled from a population of potential participants. Estimates of variance due to participant sampling are described in the section titled Inverse Sampling Error Weights, above.

If we allow for true heterogeneity in population effect size, studies with larger N do not necessarily fall closer to the mean effect size. Consequently, in a random-effects model, weighting by inverse sampling error variance or sample size is no longer optimal. Consider a meta-analysis where there is large between-study variability (i.e., where some designs produce a large population effect size while others are near zero). A particular large-sample study might represent a study design that produces an effect size considerably larger or smaller than the average. Even though the large sample size produces a precise estimate of effect sizes for that study design, it is not necessarily more representative of the collection of studies. Giving more weight to large-sample studies in this situation is not justified, and the researcher would be better off assigning equal weight to all studies.

On the other hand, if the between-study variability is small relative to sampling error variance, then effect sizes from large-sample studies will be close to the population average. As the between-study variance approaches zero, inverse sampling error weights become more optimal.

This interplay of between-study variance (τ^2) and study-specific sampling error variance ($\sigma^2_{e_j}$) is reflected in random-effects weights:

$$w_j = \frac{1}{\tau^2 + \sigma^2_{e_j}}. \qquad 8.$$

When τ^2 is small, the weights are similar to inverse variance weights, whereas when τ^2 is large relative to $\sigma^2_{e_j}$, the weights will be fairly constant across studies.

A potential issue with random-effects weights is that the between-study variance component is poorly estimated when the number of studies is small, which can lead to inaccuracy in confidence intervals when random-effects weights are used. An adjustment proposed by Hartung & Knapp (2001) and Sidik & Jonkman (2002) produces more accurate results (Langan et al. 2019) and should generally be used for random-effects meta-analysis.

Unit weights. Bonett (2008, 2009) challenged the assumptions that underlie all of these weighting schemes. As noted above, inverse sampling error weights assume a common effect model, which is unrealistic in many research domains. Conversely, treating a pool of studies as randomly sampled from a broader population does not match well with the deliberative process through which studies are conducted or curated for the purpose of a meta-analysis. Given these concerns, Bonnet recommends using unit-weighted effect sizes. While the unit-weighted average will be unbiased and requires fewer assumptions than other weighting methods, this approach tends to be less precise than other methods (producing wider confidence intervals) when the assumptions of those approaches hold (Brannick et al. 2019a, Schmidt & Hunter 2015). Additionally, even when assumptions regarding the distribution of effect sizes have been violated, random-effects

RECOMMENDATIONS FOR STUDY WEIGHTS

1. Use random-effects weights whenever estimating a random-effects model (which is almost always preferred).
2. When calculating the sampling error variance, use the mean effect size rather than the study-specific effect size.
3. For random-effects models, the Hartung–Knapp adjustment should be applied to confidence intervals.

weights remain quite accurate (Rubio-Aparicio et al. 2018). Therefore, the use of differential weights is generally preferred over unit weights.

Summary. If the distribution of effect sizes is symmetric, any of these weighting methods will provide an unbiased estimate of the mean effect size (Schmidt & Hunter 2015). The methods can, however, differ in precision. Brannick et al. (2019a) compared several weighting schemes for meta-analysis of correlations, finding that only random-effects weights performed well across varying levels of heterogeneity. Specifically, sample size weights perform well when the data resemble a common effect model but produce confidence intervals that are too narrow under conditions of large between-study variance (compare with Field 2005). In contrast, unit weights work well when there is substantial between-study variance but become relatively less precise when the between-study variance is small. Random-effects weights produce results similar to unit weights when between-study variance is large and similar to sample size weights when between-study variance is small; therefore, they are generally among the most accurate methods across varying levels of heterogeneity. When adopting random-effects weights, confidence intervals should be computed using the Hartung–Knapp approach (Hartung & Knapp 2001, Sidik & Jonkman 2002) (see the sidebar titled Recommendations for Study Weights).

Quantifying Heterogeneity

A key benefit of the random-effects model is that it provides a way to evaluate the degree of consistency or inconsistency in study results. There is a lamentable tendency for conclusions from a meta-analysis to focus primarily on the average effect, ignoring the considerable variability that is often present (Carlson & Ji 2011, DeSimone et al. 2019, Higgins et al. 2009, LeBreton et al. 2017, Oswald & McCloy 2003, Tett et al. 2017). When substantial variability exists, the mean is a poor representation of the expected findings in individual studies. In such cases, the conclusions from a meta-analysis should acknowledge that the effect size expected in any particular setting is uncertain and should address the practical and theoretical implications of the full range of findings.

Several heterogeneity statistics have been developed to assist the researcher in quantifying and interpreting the degree of heterogeneity. Borenstein et al. (2010) recommend reporting multiple measures of heterogeneity, including statistical significance (e.g., the Q test), absolute magnitude (e.g., τ, credibility interval), and relative magnitude (e.g., I^2) (see the sidebar titled Interpreting Effect Size Heterogeneity).

INTERPRETING EFFECT SIZE HETEROGENEITY

1. Conclusions from a meta-analysis should reflect the range of effect sizes, not just the mean.
2. Report multiple heterogeneity statistics, including credibility or prediction intervals.
3. Report confidence intervals on heterogeneity statistics.

Q test. One of the earliest indices of heterogeneity was the Q test (Hedges & Olkin 1985), which compares the observed variance of effect sizes with the variance expected due to sampling error. This is a test of statistical significance that evaluates the null hypothesis of no between-study variance (i.e., that all studies estimate a common effect size). Like any statistical significance test, it is sensitive to the sample size—both the N of the original studies and the number of studies included in the meta-analysis. Research has shown that the test has low power when the number of studies is low to moderate (Hedges & Pigott 2001, Jackson 2006, Viechtbauer 2007b). In other words, the Q test will often fail to detect heterogeneity when true differences across studies exist.

The Q test can be computed in fixed-effects as well as random-effects models. Meta-analysts might be tempted to use this test to determine whether a fixed- or random-effects model should be applied. However, because the test has low power for sample sizes common to meta-analyses, relying on the results of a significance test to choose between the fixed- and random-effects models is not advised (Borenstein 2019).

Between-study variance. When a random- or mixed-effects model is estimated, the analysis produces an estimate of the between-study variance component. This statistic, whether in the form of a variance (τ^2) or a standard deviation (SD_θ), provides a direct index of the degree of effect size heterogeneity in a pool of studies, and should always be reported along with the mean effect size in random-effects analyses. The SD in particular provides a concise summary statistic representing approximately how far on average a randomly selected study differs from the mean. To facilitate interpretability, it is useful to supplement τ^2 or SD_θ with a credibility or prediction interval.

Credibility and prediction intervals. A credibility interval provides a range of values representing the spread of effect sizes due to between-study heterogeneity (Schmidt & Hunter 2015, Whitener 1990). This interval differs from the observed distribution of effect sizes because the observed distribution is influenced by both true between-study variance and sampling error, while the credibility reflects the degree of between-study variance beyond that expected due to sampling error. It also differs from the confidence interval, which applies to the mean effect size. The confidence interval provides the expected range of results if the mean effect size were to be calculated on a new sample of studies with the same characteristics (e.g., the same number of studies and sample sizes) as the current meta-analysis. The credibility interval, on the other hand, provides the expected range of results for an individual study randomly sampled from the population of potential studies, where the effect size is estimated without sampling error. Confidence and credibility intervals provide distinct and useful information, and both should be reported in a meta-analysis.

In early validity generalization research, it was common to report only the lower bound of the credibility interval, because the most relevant question was whether the entire range of plausible values represented useful levels of validity. However, it has become common to report both upper and lower bounds, providing a more complete picture of the range of effect sizes in the research base.

A closely related statistic is the prediction or tolerance interval (Brannick et al. 2021). Like the credibility interval, the prediction interval provides a range of values where the population effect size from a single randomly selected study is expected to be found. The difference is that the prediction interval takes into account uncertainty in the estimate of the mean and variance of effect sizes. Credibility intervals treat these estimates as if they were known parameters, whereas probability intervals account for the fact that they are only estimates. Consequently, prediction intervals tend to be wider than credibility intervals. Brannick et al. illustrate several methods to construct prediction and tolerance intervals. Because summary statistics in many meta-analyses are based on a small to moderate number of studies, prediction intervals provide a more realistic picture of the expected range of effect sizes. To illustrate, Morris et al. (2015) report an analysis of

nine validity coefficients with a mean of 0.17 (standard error = 0.06) and an SD_ρ of 0.11, resulting in an 80% credibility interval of [0.08, 0.31]. Using the same data, Brannick et al. computed an 80% bootstrap tolerance interval of [−0.03, 0.46], which indicates considerably greater heterogeneity than the range suggested by the credibility interval.

When interpreting the degree of heterogeneity, it can be useful to consider the substantive implications of the endpoints of the credibility or prediction interval. Do the upper and lower bounds lead to substantially different conclusions about the magnitude of effect or its practical implications? Does the interval include or approach zero? It can also be informative to compare the endpoints with empirical benchmarks for effect sizes in the relevant research domain (Wiernik et al. 2017).

Percent variance statistics. Another way to conceptualize heterogeneity is through partitioning the observed variance of effect sizes into variance associated with differences between studies (τ^2) and variance attributable to statistical artifacts (σ_e^2). Early research using psychometric meta-analysis often reported the percent of variance due to artifacts. An early rule of thumb treated the population as having a common effect size if at least 75% of the variance was due to artifacts (Schmidt & Hunter 2015). Subsequent research was critical of the 75% rule (Cornwell & Ladd 1993), but this criticism does not undermine the value of using percent variance as a measure of the degree of homogeneity (Oh & Roth 2017).

In recent years, researchers have mostly switched to reporting the proportion of variance between studies due to true heterogeneity of effect size, reflected in the I^2 statistic (Higgins & Thompson 2002). For a meta-analysis where the only artifact is sampling error, I^2 is equal to one minus the proportion of variance due to artifacts. By highlighting the between-study variance rather than the artifact variance, I^2 is more consistent with other heterogeneity statistics, and it is better aligned with the goal of quantifying heterogeneity, rather than ruling out artifacts as a source of observed differences.

A major concern with percent of variance statistics is that they reflect the relative rather than absolute magnitude of heterogeneity (Borenstein et al. 2017). If a pool of studies has large sample sizes, sampling error will be small and a large I^2 can represent fairly trivial differences in effect size. Conversely, with small sample sizes, much of the variance will be attributed to sampling error and meaningful moderators might be present despite a fairly small I^2. Additionally, when the number of studies is small, I^2 can be quite inaccurate, so researchers should interpret percent variance statistics with caution (von Hippel 2015).

Estimating uncertainty in between-study variance statistics. Like any statistic, estimates of the between-study variance component (τ^2, SD_ρ) are subject to sampling error, and they may differ substantially from population values when the sample size is small. Importantly, the precision of τ^2 is driven largely by the number of studies, not the number of participants. In many meta-analyses, the number of studies is modest, especially when examining moderator subgroups. Therefore, estimates of the between-study variance should be interpreted with caution in these cases.

Methods exist to build confidence intervals on τ^2 (Biggerstaff & Tweedie 1997, Borenstein et al. 2021, Brannick et al. 2019b, Veroniki et al. 2016, Viechtbauer 2007a) and I^2 (Higgins & Thompson 2002), and these intervals are available in contemporary meta-analytic software packages such as `metafor` (Viechtbauer 2010). Although currently not standard practice, adding standard errors to between-study variance estimates would provide additional clarity regarding the precision of meta-analytic conclusions.

Morris et al. (2017) illustrate an approach to conducting a sensitivity analysis with respect to sampling error in the between-study variance component. The first step is to build a confidence

interval around the between-study variance estimate. Next, the researcher uses each endpoint of this confidence interval to construct a separate credibility interval around the mean effect size. For example, Morris et al. (2015) reported a mean uncorrected validity of 0.27 for individual psychological assessments predicting job performance, with $SD_\rho = 0.12$ resulting in an 80% credibility interval of [0.12, 0.42]. Taking into account the confidence interval on SD_ρ of [0.08, 0.19], this credibility interval might be as small as [0.17, 0.37] or as large as [0.02, 0.52]. In this case, there is evidence of substantial variability in validity across all plausible values of SD_ρ; however, the degree of heterogeneity is quite uncertain. When writing up the results, the researcher should note that the 37 studies included in this meta-analysis are not sufficient to provide a precise estimate of the expected range of validities.

This type of sensitivity analysis is closely connected to the prediction interval discussed above, but they approach the problem from different angles. The sensitivity analysis depicts uncertainty about the degree of heterogeneity through the range of plausible solutions (showing both narrower and wider credibility intervals). The prediction interval, on the other hand, reflects an expected value taken across the distribution of plausible between-study variance estimates. It seeks to define a single, more conservative interval that takes into account uncertainty in the range of heterogeneity. The two approaches provide complementary information: The prediction interval focuses on what can be known given error in our estimate of heterogeneity, while the sensitivity analysis seeks to convey in concrete terms the degree of uncertainty about heterogeneity.

MODERATOR ANALYSES

Most meta-analysts want to go beyond estimating the degree of heterogeneity to test hypotheses about its determinants. Thus, it is common for researchers to examine study characteristics that moderate the magnitude of effect size. Such analyses can be accomplished using either a subgrouping or a meta-regression approach. Each of these approaches has strengths and weaknesses, as discussed below.

Often, researchers will use a test of effect size heterogeneity as a preliminary step to determine whether the search for moderators is justified. Unfortunately, the test for heterogeneity often has low power (Hedges & Pigott 2004). Therefore, if moderators have been hypothesized a priori, these tests should be conducted regardless of the size of the between-study variance estimate or the significance of a heterogeneity test.

Subgroup Analysis

The simplest and most easily interpretable way to evaluate moderators is to sort the effect sizes into subgroups and conduct a separate meta-analysis for each subgroup (Schmidt & Hunter 2015). Moderators can be evaluated by comparing the mean effect across subgroups. Additionally, the residual between-study variance within each subgroup will indicate additional unmodeled differences across studies, although these heterogeneity estimates can be imprecise due to the smaller number of studies within subgroups.

When comparing effect sizes between moderator groups, a common approach is to build a confidence interval separately for each group and then conclude that effect sizes are different if the confidence intervals do not overlap. If it is reasonable to assume that the between-study variance is the same for all subgroups, a more powerful test will use the data from all subgroups to obtain a pooled residual variance estimate (Borenstein et al. 2017). The use of the pooled residual variance will be particularly important when there are many subgroups and where some have a small number of studies. Flexibility in adopting either pooled or separate estimates of τ^2 is a strength of the subgroup approach.

Subgroup analysis can become challenging when multiple moderator variables are examined. As with any statistical analysis, collinearity among predictors creates ambiguity in attributing shared prediction to individual moderators (Steel & Kammeyer-Mueller 2002, Viswesvaran & Sanchez 1998). Separate analyses of each moderator can be misleading because it ignores the confounding of correlated predictors. One way to deal with this issue is to further divide each subgroup by additional moderators, creating a nested structure of subgroups where the lowest level represents unique combinations of all moderators. An advantage of this approach is that it may reveal interactions among moderators (e.g., differences among subgroups on moderator A are found for high levels of moderator B, but not at low levels of B). However, nested subgrouping can become unwieldy as the number of moderators increases, and in practice it often produces some subgroups with a small number of studies. Additionally, the nested structure requires an ordering of the moderators (e.g., first split on moderator A, then on moderator B, and so on), and the order can influence what patterns are revealed. If there is no strong conceptual preference among moderators, it can be informative to repeat the subgrouping analysis using different sequences. However, doing so further increases the number of analyses that must be conducted and summarized.

Meta-Regression

In meta-regression, the effect size is regressed onto indicators for study characteristics that are hypothesized to moderate the effect size. Study characteristics can be represented as continuous predictors (e.g., date of publication) or categorical indicators (e.g., lab versus field studies). Meta-regression is particularly valuable when the researcher wants to simultaneously evaluate multiple moderators or when moderators are continuous (Steel & Kammeyer-Mueller 2002). Additionally, if the pool of studies is sufficiently large, meta-regression is capable of modeling curvilinear relationships and interactions among moderators. A useful summary of recommendations for conducting meta-regression is presented by Gonzalez-Mulé & Aguinis (2018).

Considerable research has evaluated and refined the methodology for estimating meta-regression models (Tipton et al. 2019). As with other meta-analysis methods, analyses are typically conducted using weighted least squares, where studies with greater precision are assigned greater weight, using one of the weighting methods described above in the section titled Study Weights. Standard tests for the significance of regression coefficients have been found to perform poorly in meta-regression, especially when the number of studies is not large. A method proposed by Knapp & Hartung (2003) provides the most accurate confidence intervals and significance tests on regression coefficients (Viechtbauer et al. 2015).

Although meta-regression is a flexible and useful method, researchers should be attentive to the fact that it is susceptible to all of the problems that can occur in any regression analysis (Schmidt 2017). Furthermore, many of these problems are exacerbated by the small number of data points (i.e., studies) and missing data on moderator variables, which are common in meta-analyses. While meta-regression accounts for correlations among moderators, high levels of collinearity can create ambiguity in the interpretation of regression coefficients. Estimates of R^2 from meta-regression can be inaccurate when the number of studies is below 40 (López-López et al. 2014) and will likely overestimate the strength of moderator effects unless adjustments for shrinkage are applied (Schmidt 2017, Schmidt & Hunter 2015). Additionally, it is important to evaluate the sensitivity of results to outliers (Viechtbauer & Cheung 2010).

Meta-regression models typically assume homoscedasticity of the between-study variance. The meta-regression model allows the mean effect sizes to vary across study-level moderators but provides a single pooled estimate of the residual between-study variance. Effectively, the meta-regression model assumes that the between-study variance is the same within all study subgroups

or across levels of continuous moderators. Recent research has highlighted the importance of testing this assumption and of the potential inaccuracy that can occur when this assumption is violated and the number of studies differs across subgroups (Rubio-Aparicio et al. 2017). Methods that relax this assumption to estimate subgroup-specific residual variance have recently become available (Rodriguez et al. 2021).

New Methods for Moderator Analysis

The above discussion has focused on the two most common approaches to moderator analysis. In addition to these methods, recent methodological advances have introduced models for exploratory moderator analysis, where moderator groups are derived from the data rather than from a priori hypotheses. One of these new approaches uses latent mixture models to identify multiple subpopulations in a heterogeneous collection of effect sizes (Zhang et al. 2022). This idea is not new to validity generalization research (Thomas 1990), but methodological advances have made latent mixture analyses more feasible. Another innovative strategy is to use classification and regression trees to identify interactions among moderator variables (Li et al. 2017, 2020). Both of these methods will likely function best when the number of effect sizes is large, and additional work is needed to explore their statistical power and utility when compared with traditional methods used in organizational research.

Nonindependent Effect Sizes

It is common for a study to yield multiple effect sizes that are relevant to a meta-analysis. Multiple effect sizes might be observed for a variety of reasons, including separate results for multiple outcome variables, multiple comparison groups, or multiple endpoints in a longitudinal study. If a publication reports data from multiple studies with different participants, these are often reasonably treated as independent effect sizes. In other cases, multiple effect sizes from the same study violate the independence assumption of the meta-analysis, which can bias confidence intervals on the mean effect size and increase type I error rates for moderator analyses (López-López et al. 2018).

A common approach is to obtain a single effect size estimate from each study, either by averaging (Rosenthal & Rubin 1986, Schmidt & Hunter 2015) or by selecting one of the multiple effect sizes according to some decision rule (e.g., selecting the most reliable outcome or selecting at random). The drawback of this approach is that it ignores information about potentially important differences in effect sizes across types of outcome variables. This analysis could be supplemented by separate analyses of each outcome; however, such an analysis ignores information about the correlations among random effects for different outcomes (Cheung 2019). Modern methods allow researchers to include all relevant effect sizes in a meta-analysis while properly accounting for the statistical dependency among them. These methods allow for greater flexibility in testing moderators, since they can account for both within-study and between-study characteristics that affect the effect size.

When the goal of the analysis is to estimate a well-defined collection of related effect sizes, the optimal approach will be multivariate meta-analysis (Becker 2000, Cheung 2013). A good example of this approach is the estimation of the correlation matrix as a preliminary step in meta-analytic structural equation modeling (Cheung 2008). Multivariate meta-analysis takes full advantage of the data available on multiple variables but is feasible only when many studies provide fairly complete data on the full collection of effect sizes.

In most meta-analyses, multiple effect sizes reflect less systematic processes, such as idiosyncratic choices by individual studies to include multiple operationalizations of predictor or outcome

DEALING WITH MULTIPLE EFFECT SIZES FROM THE SAME STUDY

1. Include all effect sizes in the analysis.
2. Account for nonindependence through multilevel or robust meta-analysis.
3. Apply network meta-analysis for multiple treatment conditions.

variables. The goal in these situations is not to obtain a full matrix of all effect sizes but rather to model the variability among effect sizes while accounting for the nesting of effect sizes within studies. Two useful methodologies have been developed for such data (López-López et al. 2017, 2018). Multilevel meta-analysis uses hierarchical linear modeling to partition the variance of effect sizes into within-study and between-study components (Cheung 2014, Gooty et al. 2021, Van Den Noortgate et al. 2015). Another approach is robust meta-analysis (Hedges et al. 2010, Tanner-Smith et al. 2016, Tipton 2015). Rather than specifying the dependence structure, as in multilevel meta-analysis, robust meta-analysis applies standard meta-regression methods, with adjustments for the overall level of dependency among effect sizes. Both methods perform well in accounting for nonindependent effect sizes (López-López et al. 2017). With either method, adjusting for dependencies among effect sizes tends to produce wider confidence intervals and lower power for moderator tests. This is not a limitation of the statistical analysis, but rather a reflection of the fact that the available pool of studies is often too small to obtain precise results.

Another approach to dealing with dependent effect sizes is network meta-analysis, which was developed to address the unique issues that arise in experimental research with multiple treatment comparisons. While traditional meta-analysis focuses on pairwise comparisons, network meta-analysis (Efthimiou et al. 2016, Salanti 2012) has emerged as the best way to incorporate multiple comparisons among subgroups. Network models allow for different combinations of subgroups from each study and provide a combination of direct evidence (treatment A versus B) and indirect evidence (learning about A versus B by combining evidence from A versus C with B versus C). Network models can be estimated though multivariate (Mavridis et al. 2015) or multilevel (Lu & Ades 2004) models. A useful aspect of network models consists of the methods for examining and addressing inconsistencies between direct and indirect evidence and from different research designs (Higgins et al. 2012, Law et al. 2016). To date, network meta-analysis has been applied primarily in medical research, but it is likely to become more common in future meta-analyses in organizational research (see the sidebar titled Dealing with Multiple Effect Sizes from the Same Study).

CONCLUSION

This article has reviewed a few of the many important choices involved in conducting a meta-analysis (for a broader review, see Rudolph et al. 2020, Steel et al. 2021). Notably, the article has not addressed the critical issues involved in conducting a comprehensive literature search (Gusenbauer & Haddaway 2020, Siddaway et al. 2019), including strategies for identifying unpublished studies (Adams et al. 2017, Rothstein & Hopewell 2009) and for screening the studies obtained (Polanin et al. 2019). Additionally, the article has not discussed methods for assessing publication bias (Carter et al. 2019, Kepes et al. 2012). These topics are just as critical for effective meta-analysis as the methodological choices explored in this review.

Transparency and open science practices are critical to the integrity of science, and meta-analysis is no exception. It has long been recognized that choices and judgment calls when conducting a meta-analysis can influence the results (Park et al. 2020, Wanous et al. 1989; compare with Aguinis et al. 2011). It is therefore critical that reports of meta-analytic research include a full

and detailed description of analytic decisions and methodological choices (Aytug et al. 2012). Several sources provide excellent guidelines for reporting meta-analytic work (DeSimone et al. 2021, Kepes et al. 2013, Siddaway et al. 2019), including the American Psychological Association's Quantitative Meta-Analysis Reporting Standards (APA 2018). Researchers conducting meta-analysis and journals that publish this work would be well advised to familiarize themselves with these standards.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I thank Amanda Neuman for assisting with the preparation of this review.

LITERATURE CITED

Adams RJ, Smart P, Huff AS. 2017. Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies. *Int. J. Manag. Rev.* 19(4):432–54

Aguinis H. 2001. Estimation of sampling variance of correlations in meta-analysis. *Pers. Psychol.* 54(3):569–90

Aguinis H, Dalton DR, Bosco FA, Pierce CA, Dalton CM. 2011. Meta-analytic choices and judgment calls: implications for theory building and testing, obtained effect sizes, and scholarly impact. *J. Manag.* 37(1):5–38

Aguinis H, Pierce CA, Culpepper SA. 2009. Scale coarseness as a methodological artifact: correcting correlation coefficients attenuated from using coarse scales. *Organ. Res. Methods* 12(4):623–52

APA (Am. Psychol. Assoc.). 2018. *JARS Quant table 9: quantitative meta-analysis article reporting standards*. J. Art. Report. Stand., APA, Washington, DC. **https://apastyle.apa.org/jars/quant-table-9.pdf**

Aytug ZG, Rothstein HR, Zhou W, Kern MC. 2012. Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organ. Res. Methods* 15(1):103–33

Baguley T. 2009. Standardized or simple effect size: What should be reported? *Br. J. Psychol.* 100(3):603–17

Banks GC, Rogelberg SG, Woznyj HM, Landis RS, Rupp DE. 2016. Evidence on questionable research practices: the good, the bad, and the ugly. *J. Bus. Psychol.* 31(3):323–38

Barrett GV, Phillips JS, Alexander RA. 1981. Concurrent and predictive validity designs: a critical reanalysis. *J. Appl. Psychol.* 66(1):1–6

Beatty AS, Barratt CL, Berry CM, Sackett PR. 2014. Testing the generalizability of indirect range restriction corrections. *J. Appl. Psychol.* 99(4):587–98

Becker BJ. 1988. Synthesizing standardized mean-change measures. *Br. J. Math. Stat. Psychol.* 41(2):257–78

Becker BJ. 2000. Multivariate meta-analysis. In *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, ed. HEA Tinsley, SD Brown, pp. 499–525. San Diego, CA: Academic

Becker BJ, Wu M-J. 2007. The synthesis of regression slopes in meta-analysis. *Stat. Sci.* 22(3):414–29

Berry CM, Sackett PR, Landers RN. 2007. Revisiting interview–cognitive ability relationships: attending to specific range restriction mechanisms in meta-analysis. *Pers. Psychol.* 60(4):837–74

Biggerstaff BJ, Tweedie RL. 1997. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Stat. Med.* 16(7):753–68

Bobko P. 1983. An analysis of correlations corrected for attenuation and range restriction. *J. Appl. Psychol.* 68(4):584–89

Bobko P, Roth PL, Bobko C. 2001. Correcting the effect size of *d* for range restriction and unreliability. *Organ. Res. Methods* 4(1):46–61

Bond CF, Wiitala WL, Richard FD. 2003. Meta-analysis of raw mean differences. *Psychol. Methods* 8(4):406–18

Bonett D. 2008. Meta-analytic interval estimation for bivariate correlations. *Psychol. Methods* 13:173–81

Bonett D. 2009. Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychol. Methods* 14:225–38

Borenstein M. 2019. *Common Mistakes in Meta-Analysis and How to Avoid Them.* Englewood, NJ: Biostat
Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods* 1(2):97–111
Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. 2021. *Introduction to Meta-Analysis*. New York: Wiley
Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. 2017. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Res. Synth. Methods* 8(1):5–18
Brannick MT, French KA, Rothstein HR, Kiselica AM, Apostoloski N. 2021. Capturing the underlying distribution in meta-analysis: credibility and tolerance intervals. *Res. Synth. Methods* 12(3):264–90
Brannick MT, Potter SM, Benitez B, Morris SB. 2019a. Bias and precision of alternate estimators in meta-analysis: benefits of blending Schmidt–Hunter and Hedges approaches. *Organ. Res. Methods* 22(2):490–514
Brannick MT, Potter S, Teng Y. 2019b. Quantifying uncertainty in the meta-analytic lower bound estimate. *Psychol. Methods* 24(6):754–73
Brown RD, Oswald FL, Converse PD. 2017. Estimating operational validity under incidental range restriction: some important but neglected issues. *Pract. Assess. Res. Eval.* 22(6):1–6
Brown SH, Stout JD, Dalessio AT, Crosby MM. 1988. Stability of validity indices through test score ranges. *J. Appl. Psychol.* 73(4):736–42
Burke MJ, Landis RS, Burke MI. 2014. .80 and beyond: recommendations for disattenuating correlations. *Ind. Organ. Psychol.* 7(4):531–35
Burr D, Doss H. 2005. A Bayesian semiparametric model for random-effects meta-analysis. *J. Am. Stat. Assoc.* 100(469):242–51
Carlson KD, Ji FX. 2011. Citing and building on meta-analytic findings: a review and recommendations. *Organ. Res. Methods* 14(4):696–717
Carretta TR, Ree MJ. 2022. Correction for range restriction: lessons from 20 research scenarios. *Mil. Psychol.* 34(5):551–69
Carter EC, Schönbrodt FD, Gervais WM, Hilgard J. 2019. Correcting for bias in psychology: a comparison of meta-analytic methods. *Adv. Methods Pract. Psychol. Sci.* 2(2):115–44
Cheung MW-L. 2008. A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling. *Psychol. Methods* 13(3):182–202
Cheung MW-L. 2013. Multivariate meta-analysis as structural equation models. *Struct. Equ. Model. Multidiscip. J.* 20(3):429–54
Cheung MW-L. 2014. Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychol. Methods* 19(2):211–29
Cheung MW-L. 2019. A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychol. Rev.* 29(4):387–96
Conway JM, Huffcutt AI. 1997. Psychometric properties of multisource performance ratings: a meta-analysis of subordinate, supervisor, peer, and self-ratings. *Hum. Perform.* 10(4):331–60
Cornwell JM, Ladd RT. 1993. Power and accuracy of the Schmidt and Hunter meta-analytic procedures. *Educ. Psychol. Meas.* 53(4):877–95
Cortina JM. 1993. What is coefficient alpha? An examination of theory and applications. *J. Appl. Psychol.* 78(1):98–104
Cortina JM, Nouri H. 2000. *Effect Size for ANOVA Designs*. Thousand Oaks, CA: SAGE
Cortina JM, Sheng Z, Keener SK, Keeler KR, Grubb LK, et al. 2020. From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the *Journal of Applied Psychology*. *J. Appl. Psychol.* 105(12):1351–81
Coward WM, Sackett PR. 1990. Linearity of ability–performance relationships: a reconfirmation. *J. Appl. Psychol.* 75(3):297–300
Culpepper SA. 2016. An improved correction for range restricted correlations under extreme, monotonic quadratic nonlinearity and heteroscedasticity. *Psychometrika* 81(2):550–64
Cumming G. 2012. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge/Taylor & Francis
Dahlke JA, Sackett PR. 2018. Refinements to effect sizes for tests of categorical moderation and differential prediction. *Organ. Res. Methods* 21(1):226–34

Dahlke JA, Wiernik BM. 2020. Not restricted to selection research: accounting for indirect range restriction in organizational research. *Organ. Res. Methods* 23(4):717–49
DeShon RP. 2003. A generalizability theory perspective on measurement error corrections in validity generalization. In *Validity Generalization: A Critical Review*, ed. KR Murphy, pp. 365–402. Mahwah, NJ: Erlbaum
DeSimone JA, Brannick MT, O'Boyle EH, Ryu JW. 2021. Recommendations for reviewing meta-analyses in organizational research. *Organ. Res. Methods* 24(4):694–717
DeSimone JA, Köhler T, Schoen JL. 2019. If it were only that easy: the use of meta-analytic research by organizational scholars. *Organ. Res. Methods* 22(4):867–91
Edwards JR. 2011. The fallacy of formative measurement. *Organ. Res. Methods* 14(2):370–88
Efthimiou O, Debray TPA, van Valkenhoef G, Trelle S, Panayidou K, et al. 2016. GetReal in network meta-analysis: a review of the methodology. *Res. Synth. Methods* 7(3):236–63
Ellington JK, McAbee ST, Landis RS, Mead AD. 2021. I only have one rater per ratee, so what? The impact of clustered performance rating data on operational validity estimates. *J. Bus. Psychol.* 36(1):33–54
Erez A, Bloom MC, Wells MT. 1996. Using random rather than fixed effects models in meta-analysis: implications for situational specificity and validity generalization. *Pers. Psychol.* 49(2):275–306
Feingold A. 2009. Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychol. Methods* 14(1):43–53
Field AP. 2001. Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. *Psychol. Methods* 6(2):161–80
Field AP. 2005. Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychol. Methods* 10(4):444–67
Fife DA, Hunter MD, Mendoza JL. 2016. Estimating unattenuated correlations with limited information about selection variables: alternatives to Case IV. *Organ. Res. Methods* 19(4):593–615
Fife DA, Mendoza J, Day E, Terry R. 2020. Estimating subgroup differences in staffing research when the selection mechanism is unknown: a response to Li's Case IV correction. *Organ. Res. Methods* 23(2):367–84
Fife DA, Mendoza JL, Terry R. 2013. Revisiting Case IV: a reassessment of bias and standard errors of Case IV under range restriction. *Br. J. Math. Stat. Psychol.* 66(3):521–42
Finkelstein LM, Burke MJ, Raju MS. 1995. Age discrimination in simulated employment contexts: an integrative analysis. *J. Appl. Psychol.* 80(6):652–63
Glass GV, McGaw B, Smith ML. 1981. *Meta-Analysis in Social Research*. Thousand Oaks, CA: SAGE
Gonzalez-Mulé E, Aguinis H. 2018. Advancing theory by assessing boundary conditions with metaregression: a critical review and best-practice recommendations. *J. Manag.* 44(6):2246–73
Gooty J, Banks GC, Loignon AC, Tonidandel S, Williams CE. 2021. Meta-analyses as a multi-level model. *Organ. Res. Methods* 24(2):389–411
Gross AL. 1982. Relaxing the assumptions underlying corrections for restriction of range. *Educ. Psychol. Meas.* 42(3):795–801
Gross AL, Fleischman L. 1983. Restriction of range corrections when both distribution and selection assumptions are violated. *Appl. Psychol. Meas.* 7(2):227–37
Gusenbauer M, Haddaway NR. 2020. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Res. Synth. Methods* 11(2):181–217
Haddock CK, Rindskopf D, Shadish WR. 1998. Using odds ratios as effect sizes for meta-analysis of dichotomous data: a primer on methods and issues. *Psychol. Methods* 3(3):339–53
Hall SM, Brannick MT. 2002. Comparison of two random-effects methods of meta-analysis. *J. Appl. Psychol.* 87(2):377–89
Hartung J, Knapp G. 2001. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat. Med.* 20(24):3875–89
Hedges LV. 1982. Estimation of effect size from a series of independent experiments. *Psychol. Bull.* 92(2):490–99
Hedges LV. 1992. Meta-analysis. *J. Educ. Stat.* 17(4):279–96
Hedges LV. 2009. Effect sizes in nested designs. In *The Handbook of Research Synthesis and Meta-Analysis*, ed. H Cooper, LV Hedges, JC Valentine, pp. 337–55. New York: Russell Sage Found. 2nd ed.

Hedges LV, Olkin I. 1985. *Statistical Methods for Meta-Analysis*. San Diego, CA: Academic
Hedges LV, Pigott TD. 2001. The power of statistical tests in meta-analysis. *Psychol. Methods* 6(3):203–17
Hedges LV, Pigott TD. 2004. The power of statistical tests for moderators in meta-analysis. *Psychol. Methods* 9(4):426–45
Hedges LV, Tipton E, Johnson MC. 2010. Robust variance estimation in meta-regression with dependent effect size estimates. *Res. Synth. Methods* 1(1):39–65
Hedges LV, Vevea JL. 1998. Fixed- and random-effects models in meta-analysis. *Psychol. Methods* 3(4):486–504
Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. 2012. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res. Synth. Methods* 3(2):98–110
Higgins JPT, Thompson SG. 2002. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* 21(11):1539–58
Higgins JPT, Thompson SG, Spiegelhalter DJ. 2009. A re-evaluation of random-effects meta-analysis. *J. R. Stat. Soc. A* 172(1):137–59
Highhouse S, Doverspike D, Guion RM. 2015. *Essentials of Personnel Assessment and Selection*. New York: Routledge. 2nd ed.
Hoffman BJ, Woehr DJ. 2009. Disentangling the meaning of multisource performance rating source and dimension factors. *Pers. Psychol.* 62(4):735–65
Huffcutt AI, Culbertson SS, Weyhrauch WS. 2014. Moving forward indirectly: reanalyzing the validity of employment interviews with indirect range restriction methodology. *Int. J. Sel. Assess.* 22(3):297–309
Hunter JE, Schmidt FL, Le H. 2006. Implications of direct and indirect range restriction for meta-analysis methods and findings. *J. Appl. Psychol.* 91(3):594–612
Jackson D. 2006. The power of the standard test for the presence of heterogeneity in meta-analysis. *Stat. Med.* 25(15):2688–99
James LR, Demaree RG, Mulaik SA, Ladd RT. 1992. Validity generalization in the context of situational models. *J. Appl. Psychol.* 77(1):3–14
Kelley K, Preacher KJ. 2012. On effect size. *Psychol. Methods* 17(2):137–52
Kepes S, Banks GC, McDaniel M, Whetzel DL. 2012. Publication bias in the organizational sciences. *Organ. Res. Methods* 15(4):624–62
Kepes S, Keener SK, McDaniel MA, Hartman NS. 2022. Questionable research practices among researchers in the most research-productive management programs. *J. Organ. Behav.* 43(7):1190–208
Kepes S, McDaniel MA, Brannick MT, Banks GC. 2013. Meta-analytic reviews in the organizational sciences: two meta-analytic schools on the way to MARS (the Meta-Analytic Reporting Standards). *J. Bus. Psychol.* 28(2):123–43
Knapp G, Hartung J. 2003. Improved tests for a random effects meta-regression with a single covariate. *Stat. Med.* 22(17):2693–710
Köhler T, Cortina JM, Kurtessis JN, Gölz M. 2015. Are we correcting correctly? Interdependence of reliabilities in meta-analysis. *Organ. Res. Methods* 18(3):355–428
Lance CE, Hoffman BJ, Gentry WA, Baranik LE. 2008. Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Hum. Resour. Manag. Rev.* 18(4):223–32
Langan D, Higgins JPT, Jackson D, Bowden J, Veroniki AA, et al. 2019. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res. Synth. Methods* 10(1):83–98
Law KS, Schmidt FL, Hunter JE. 1994. A test of two refinements in procedures for meta-analysis. *J. Appl. Psychol.* 79(6):978–86
Law M, Jackson D, Turner R, Rhodes K, Viechtbauer W. 2016. Two new methods to fit models for network meta-analysis with random inconsistency effects. *BMC Med. Res. Methodol.* 16:87
Le H, Oh I-S, Schmidt FL, Wooldridge CD. 2016. Correction for range restriction in meta-analysis revisited: improvements and implications for organizational research. *Pers. Psychol.* 69(4):975–1008
Le H, Schmidt FL, Putka DJ. 2009. The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organ. Res. Methods* 12(1):165–200
LeBreton JM, Burgess JRD, Kaiser RB, Atchley EK, James LR. 2003. The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organ. Res. Methods* 6(1):80–128
LeBreton JM, Scherer KT, James LR. 2014. Corrections for criterion reliability in validity generalization: a false prophet in a land of suspended judgment. *Ind. Organ. Psychol.* 7(4):478–500

LeBreton JM, Schoen JL, James LR. 2017. Situational specificity, validity generalization, and the future of psychometric meta-analysis. In *Handbook of Employee Selection*, ed. JL Farr, NT Tippins, pp. 93–114. London: Taylor & Francis. 2nd ed.

Lee KJ, Thompson SG. 2008. Flexible parametric models for random-effects distributions. *Stat. Med.* 27(3):418–34

Li JC-H. 2015. Cohen's *d* corrected for Case IV range restriction: a more accurate procedure for evaluating subgroup differences in organizational research. *Pers. Psychol.* 68(4):899–927

Li JC-H, Cui Y, Chan W. 2013. Bootstrap confidence intervals for the mean correlation corrected for Case IV range restriction: a more adequate procedure for meta-analysis. *J. Appl. Psychol.* 98(1):183–93

Li X, Dusseldorp E, Meulman JJ. 2017. Meta-CART: a tool to identify interactions between moderators in meta-analysis. *Br. J. Math. Stat. Psychol.* 70(1):118–36

Li X, Dusseldorp E, Su X, Meulman JJ. 2020. Multiple moderator meta-analysis using the R-package Meta-CART. *Behav. Res. Methods* 52(6):2657–73

Lipsey MW, Wilson DB. 1993. The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *Am. Psychol.* 48(12):1181–209

López-López JA, Marín-Martínez F, Sánchez-Meca J, Van Den Noortgate W, Viechtbauer W. 2014. Estimation of the predictive power of the model in mixed-effects meta-regression: a simulation study. *Br. J. Math. Stat. Psychol.* 67(1):30–48

López-López JA, Page MJ, Lipsey MW, Higgins JPT. 2018. Dealing with effect size multiplicity in systematic reviews and meta-analyses. *Res. Synth. Methods* 9(3):336–51

López-López JA, Van Den Noortgate W, Tanner-Smith EE, Wilson SJ, Lipsey MW. 2017. Assessing meta-regression methods for examining moderator relationships with dependent effect sizes: a Monte Carlo simulation. *Res. Synth. Methods* 8(4):435–50

Lu G, Ades AE. 2004. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat. Med.* 23(20):3105–24

MacKenzie SB, Podsakoff PM, Jarvis CB. 2005. The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *J. Appl. Psychol.* 90(4):710–30

Mavridis D, Giannatsi M, Cipriani A, Salanti G. 2015. A primer on network meta-analysis with emphasis on mental health. *Evid. Based Ment. Health* 18(2):40–46

Morris SB. 2008. Estimating effect sizes from pretest-posttest-control group designs. *Organ. Res. Methods* 11(2):364–86

Morris SB, Daisley RL, Wheeler M, Boyer P. 2015. A meta-analysis of the relationship between individual assessments and job performance. *J. Appl. Psychol.* 100(1):5–20

Morris SB, DeShon RP. 1997. Correcting effect sizes computed from factor analysis of variance for use in meta-analysis. *Psychol. Methods* 2(2):192

Morris SB, DeShon RP. 2002. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol. Methods* 7(1):105–25

Morris SB, McAbee ST, Landis RS, Bauer KN. 2017. Don't get too confident: uncertainty in SD_ρ. *Ind. Organ. Psychol.* 10(3):467–72

Morris SB, Shokri A. 2021. Effect size and effect uncertainty in organizational research methods. In *Oxford Research Encyclopedia of Business and Management*, ed. MA Hitt. Oxford, UK: Oxford Univ. Press. **https://doi.org/10.1093/acrefore/9780190224851.013.238**

Murphy KR. 2000. Impact of assessments of validity generalization and situational specificity on the science and practice of personnel selection. *Int. J. Sel. Assess.* 8(4):194–206

Murphy KR, DeShon RP. 2000. Interrater correlations do not estimate the reliability of job performance ratings. *Pers. Psychol.* 53(4):873–900

Murphy KR, Myors B, Wolach A. 2009. *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*. New York: Routledge/Taylor & Francis. 3rd ed.

Nimon K, Zientek L, Henson R. 2012. The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Front. Psychol.* 3:102

Nye CD, Sackett PR. 2017. New effect sizes for tests of categorical moderation and differential prediction. *Organ. Res. Methods* 20(4):639–64

Oh I-S, Roth PL. 2017. On the mystery (or myth) of challenging principles and methods of validity generalization (VG) based on fragmentary knowledge and improper or outdated practices of VG. *Ind. Organ. Psychol.* 10(3):479–85

Olejnik S, Algina J. 2000. Measures of effect size for comparative studies: applications, interpretations, and limitations. *Contemp. Educ. Psychol.* 25(3):241–86

Olian JD, Schwab DP, Haberfeld Y. 1988. The impact of applicant gender compared to qualifications on hiring recommendations: a meta-analysis of experimental studies. *Organ. Behav. Hum. Decis. Process.* 41(2):180–95

O'Neill TA, McLarnon MJW, Carswell JJ. 2015. Variance components of job performance ratings. *Hum. Perform.* 28(1):66–91

Ones DS, Viswesvaran C. 2003. Job-specific applicant pools and national norms for personality scales: implications for range-restriction corrections in validation research. *J. Appl. Psychol.* 88(3):570–77

Oswald FL, Ercan S, McAbee ST, Ock J, Shaw A. 2015. Imperfect corrections or correct imperfections? Psychometric corrections in meta-analysis. *Ind. Organ. Psychol.* 8(2):e1–4

Oswald FL, Johnson JW. 1998. On the robustness, bias, and stability of statistics from meta-analysis of correlation coefficients: some initial Monte Carlo findings. *J. Appl. Psychol.* 83(2):164–78

Oswald FL, McCloy RA. 2003. Meta-analysis and the art of the average. In *Validity Generalization: A Critical Review*, ed. KR Murphy, pp. 311–38. Mahwah, NJ: Erlbaum

Park HSH, Wiernik BM, Oh I-S, Gonzalez-Mulé E, Ones DS, Lee Y. 2020. Meta-analytic five-factor model personality intercorrelations: eeny, meeny, miney, moe, how, which, why, and where to go. *J. Appl. Psychol.* 105(12):1490–529

Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. 2019. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Res. Synth. Methods* 10(3):330–42

Putka DJ, Hoffman BJ, Carter NT. 2014. Correcting the correction: when individual raters offer distinct but valid perspectives. *Ind. Organ. Psychol.* 7(4):543–48

Raju NS, Anselmi TV, Goodman JS, Thomas A. 1998. The effect of correlated artifacts and true validity on the accuracy of parameter estimation in validity generalization. *Pers. Psychol.* 51(2):453–65

Raju NS, Brand PA. 2003. Determining the significance of correlations corrected for unreliability and range restriction. *Appl. Psychol. Meas.* 27(1):52–71

Raju NS, Burke MJ, Normand J, Langlois GM. 1991. A new meta-analytic approach. *J. Appl. Psychol.* 76(3):432–46

Raju NS, Drasgow F. 2003. Maximum likelihood estimation in validity generalization. In *Validity Generalization: A Critical Review*, ed. KR Murphy, pp. 263–85. Mahwah, NJ: Erlbaum

Raju NS, Fralicx R, Steinhaus SD. 1986. Covariance and regression slope models for studying validity generalization. *Appl. Psychol. Meas.* 10(2):195–211

Robie C, Ryan AM. 1999. Effects of nonlinearity and heteroscedasticity on the validity of conscientiousness in predicting overall job performance. *Int. J. Sel. Assess.* 7(3):157–69

Rodriguez JE, Williams DR, Bürkner P-C. 2021. Heterogeneous heterogeneity by default: testing categorical moderators in random-effects meta-analysis. PsyArXiv tqcka. **https://doi.org/10.31234/osf.io/tqcka**

Rosenthal R, Rubin DB. 1986. Meta-analytic procedures for combining studies with multiple effect sizes. *Psychol. Bull.* 99(3):400–6

Roth PL, Le H, Oh I-S, Van Iddekinge CH, Buster MA, et al. 2014. Differential validity for cognitive ability tests in employment and educational settings: not much more than range restriction? *J. Appl. Psychol.* 99(1):1–20

Rothstein HR, Hopewell S. 2009. Grey literature. In *The Handbook of Research Synthesis and Meta-Analysis*, ed. H Cooper, LV Hedges, JC Valentine, pp. 103–25. New York: Russell Sage Found. 2nd ed.

Rubio-Aparicio M, López-López JA, Sánchez-Meca J, Marín-Martínez F, Viechtbauer W, Van Den Noortgate W. 2018. Estimation of an overall standardized mean difference in random-effects meta-analysis if the distribution of random effects departs from normal. *Res. Synth. Methods* 9(3):489–503

Rubio-Aparicio M, Sánchez-Meca J, López-López JA, Botella J, Marín-Martínez F. 2017. Analysis of categorical moderators in mixed-effects meta-analysis: consequences of using pooled versus separate estimates of the residual between-studies variances. *Br. J. Math. Stat. Psychol.* 70(3):439–56

Rudolph CW, Chang CK, Rauvola RS, Zacher H. 2020. Meta-analysis in vocational behavior: a systematic review and recommendations for best practices. *J. Vocat. Behav.* 118:103397
Sackett PR. 2014. When and why correcting validity coefficients for interrater reliability makes sense. *Ind. Organ. Psychol.* 7(4):501–6
Sackett PR, Laczo RM, Arvey RD. 2002. The effects of range restriction on estimates of criterion interrater reliability: implications for validation research. *Pers. Psychol.* 55(4):807–25
Sackett PR, Lievens F, Berry CM, Landers RN. 2007. A cautionary note on the effects of range restriction on predictor intercorrelations. *J. Appl. Psychol.* 92(2):538–44
Sackett PR, Ostgaard DJ. 1994. Job-specific applicant pools and national norms for cognitive ability tests: implications for range restriction corrections in validation research. *J. Appl. Psychol.* 79(5):680–84
Sackett PR, Yang H. 2000. Correction for range restriction: an expanded typology. *J. Appl. Psychol.* 85(1):112–18
Sackett PR, Zhang C, Berry CM. 2021. Challenging conclusions about predictive bias against Hispanic test takers in personnel selection. *J. Appl. Psychol.* **https://doi.org/10.1037/apl0000978**
Sackett PR, Zhang C, Berry CM, Lievens F. 2022. Revisiting meta-analytic estimates of validity in personnel selection: addressing systematic overcorrection for restriction of range. *J. Appl. Psychol.* 107(11):2040–68
Salanti G. 2012. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Res. Synth. Methods* 3(2):80–97
Salgado JF, Moscoso S. 2019. Meta-analysis of interrater reliability of supervisory performance ratings: effects of appraisal purpose, scale type, and range restriction. *Front. Psychol.* 10:2281
Schmidt FL. 2017. Statistical and measurement pitfalls in the use of meta-regression in meta-analysis. *Career Dev. Int.* 22(5):469–76
Schmidt FL, Hunter JE. 1996. Measurement error in psychological research: lessons from 26 research scenarios. *Psychol. Methods* 1(2):199–223
Schmidt FL, Hunter JE. 2015. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, CA: SAGE. 3rd ed.
Schmidt FL, Hunter JE, Pearlman K, Hirsh HR, Sackett PR, et al. 1985. Forty questions about validity generalization and meta-analysis. *Pers. Psychol.* 38(4):697–798
Schmidt FL, Le H, Ilies R. 2003. Beyond alpha: an empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychol. Methods* 8(2):206–24
Schmidt FL, Oh I-S, Hayes TL. 2009. Fixed- versus random-effects models in meta-analysis: model properties and an empirical comparison of differences in results. *Br. J. Math. Stat. Psychol.* 62(1):97–128
Schulze R. 2004. *Meta-Analysis: A Comparison of Approaches*. Göttingen, Ger.: Hogrefe
Senior AM, Viechtbauer W, Nakagawa S. 2020. Revisiting and expanding the meta-analysis of variation: the log coefficient of variation ratio. *Res. Synth. Methods* 11(4):553–67
Shen W, Cucina JM, Walmsley PT, Seltzer BK. 2014. When correcting for unreliability of job performance ratings, the best estimate is still .52. *Ind. Organ. Psychol.* 7(4):519–24
Siddaway AP, Wood AM, Hedges LV. 2019. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annu. Rev. Psychol.* 70:747–70
Sidik K, Jonkman JN. 2002. A simple confidence interval for meta-analysis. *Stat. Med.* 21(21):3153–59
SIOP (Soc. Ind. Organ. Psychol.). 2018. Principles for the validation and use of personnel selection procedures. *Ind. Organ. Psychol. Perspect. Sci. Pract.* 11(Suppl. 1):2–97
Steel PD, Beugelsdijk S, Aguinis H. 2021. The anatomy of an award-winning meta-analysis: recommendations for authors, reviewers, and readers of meta-analytic reviews. *J. Int. Bus. Stud.* 52(1):23–44
Steel PD, Kammeyer-Mueller JD. 2002. Comparing meta-analytic moderator estimation techniques under realistic conditions. *J. Appl. Psychol.* 87(1):96–111
Tanner-Smith EE, Tipton E, Polanin JR. 2016. Handling complex meta-analytic data structures using robust variance estimates: a tutorial in R. *J. Dev. Life-Course Criminol.* 2(1):85–112
Tett RP, Hundley NA, Christiansen ND. 2017. Meta-analysis and the myth of generalizability. *Ind. Organ. Psychol.* 10(3):421–56
Thomas A, Raju NS. 2004. An evaluation of James et al.'s 1992 VG estimation procedure when artifacts and true validity are correlated. *Int. J. Sel. Assess.* 12(4):299–311

Thomas H. 1990. A likelihood-based model for validity generalization. *J. Appl. Psychol.* 75(1):13–20
Tipton E. 2015. Small sample adjustments for robust variance estimation with meta-regression. *Psychol. Methods* 20(3):375–93
Tipton E, Pustejovsky JE, Ahmadi H. 2019. A history of meta-regression: technical, conceptual, and practical developments between 1974 and 2018. *Res. Synth. Methods* 10(2):161–79
Tziner A, Murphy KR, Cleveland JN. 2005. Contextual and rater factors affecting rating behavior. *Group Organ. Manag.* 30(1):89–98
Van Den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J. 2015. Meta-analysis of multiple outcomes: a multilevel approach. *Behav. Res. Methods* 47(4):1274–94
Van Den Noortgate W, Onghena P. 2003. Estimating the mean effect size in meta-analysis: bias, precision, and mean squared error of different weighting methods. *Behav. Res. Methods Instrum. Comput.* 35(4):504–11
van Zundert CHJ, Miočević M. 2020. A comparison of meta-methods for synthesizing indirect effects. *Res. Synth. Methods* 11(6):849–65
Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, et al. 2016. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res. Synth. Methods* 7(1):55–79
Viechtbauer W. 2005. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J. Educ. Behav. Stat.* 30(3):261–93
Viechtbauer W. 2007a. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat. Med.* 26(1):37–52
Viechtbauer W. 2007b. Hypothesis tests for population heterogeneity in meta-analysis. *Br. J. Math. Stat. Psychol.* 60(1):29–60
Viechtbauer W. 2010. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* 36:1–48
Viechtbauer W, Cheung MW-L. 2010. Outlier and influence diagnostics for meta-analysis. *Res. Synth. Methods* 1(2):112–25
Viechtbauer W, López-López JA, Sánchez-Meca J, Marín-Martínez F. 2015. A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychol. Methods* 20(3):360–74
Viswesvaran C, Ones DS, Schmidt FL. 1996. Comparative analysis of the reliability of job performance ratings. *J. Appl. Psychol.* 81(5):557–74
Viswesvaran C, Ones DS, Schmidt FL, Le H, Oh I-S. 2014. Measurement error obfuscates scientific knowledge: Path to cumulative knowledge requires corrections for unreliability and psychometric meta-analyses. *Ind. Organ. Psychol.* 7(4):507–18
Viswesvaran C, Sanchez JI. 1998. Moderator search in meta-analysis: a review and cautionary note on existing approaches. *Educ. Psychol. Meas.* 58(1):77–87
von Hippel PT. 2015. The heterogeneity statistic I^2 can be biased in small meta-analyses. *BMC Med. Res. Methodol.* 15:35
Wanous JP, Sullivan SE, Malinak J. 1989. The role of judgment calls in meta-analysis. *J. Appl. Psychol.* 74(2):259–64
Whitener EM. 1990. Confusion of confidence intervals and credibility intervals in meta-analysis. *J. Appl. Psychol.* 75(3):315–21
Wiernik BM, Dahlke JA. 2020. Obtaining unbiased results in meta-analysis: the importance of correcting for statistical artifacts. *Adv. Methods Pract. Psychol. Sci.* 3(1):94–123
Wiernik BM, Kostal JW, Wilmot MP, Dilchert S, Ones DS. 2017. Empirical benchmarks for interpreting effect size variability in meta-analysis. *Ind. Organ. Psychol.* 10(3):472–79
Yuan Z, Morgeson FP, LeBreton JM. 2020. Maybe not so independent after all: the possibility, prevalence, and consequences of violating the independence assumptions in psychometric meta-analysis. *Pers. Psychol.* 73(3):491–516
Zhang N, Wang M, Xu H. 2022. Disentangling effect size heterogeneity in meta-analysis: a latent mixture approach. *Psychol. Methods* 27(3):373–99
Zimmerman DW. 2007. Correction for attenuation with biased reliability estimates and correlated errors in populations and samples. *Educ. Psychol. Meas.* 67(6):920–39