Check for updates

Invited Forum: Meta-Analysis
General Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and
Practices in Psychological Science
2020, Vol. 3(1) 94–123
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245919885611
www.psychologicalscience.org/AMPPS
SAGE

Obtaining Unbiased Results in Meta-Analysis: The Importance of Correcting for Statistical Artifacts

Brenton M. Wiernik[1] **and Jeffrey A. Dahlke**[2]
[1]Department of Psychology, University of South Florida, and [2]Human Resources Research Organization, Alexandria, Virginia

Abstract
Most published meta-analyses address only artifactual variance due to sampling error and ignore the role of other statistical and psychometric artifacts, such as measurement error variance (due to factors including unreliability of measurements, group misclassification, and variable treatment strength) and selection effects (including range restriction or enhancement and collider biases). These artifacts can have severe biasing effects on the results of individual studies and meta-analyses. Failing to account for these artifacts can lead to inaccurate conclusions about the mean effect size and between-studies effect-size heterogeneity, and can influence the results of meta-regression, publication-bias, and sensitivity analyses. In this article, we provide a brief introduction to the biasing effects of measurement error variance and selection effects and their relevance to a variety of research designs. We describe how to estimate the effects of these artifacts in different research designs and correct for their impacts in primary studies and meta-analyses. We consider meta-analyses of correlations, observational group differences, and experimental effects. We provide R code to implement the corrections described.

Keywords
psychometric meta-analysis, measurement error, reliability, range restriction, range variation, selection bias, collider bias, open materials

Received 11/6/18; Revision accepted 10/8/19

Meta-analysis is a critical tool for increasing the rigor of research syntheses by increasing confidence that apparent differences in findings across samples are not merely attributable to statistical artifacts (Schmidt, 2010). However, most published meta-analyses are concerned only with artifactual variance due to sampling error and ignore the role of other statistical and psychometric artifacts, such as measurement error variance (stemming from factors including unreliability of measurements, group misclassification, and variable treatment strength) and selection effects (including range restriction or enhancement and collider biases). In this article, we provide a brief introduction to the biasing effects of measurement error variance and selection effects and their relevance to a variety of research designs. We describe how to estimate the effects of these artifacts in different research designs and how to correct for their impacts in meta-analyses. We consider meta-analyses of correlations, observational group differences, and experimental effects.

As we just noted, most published meta-analyses are concerned only with sampling error variance and do not correct for other statistical artifacts. For example, we reviewed the 71 meta-analyses published in *Psychological Bulletin* during 2016 through 2018 and found that only 6 made corrections for measurement error variance, and only 1 corrected for selection biases (a similar review by Schmidt, 2010, found similarly low rates of corrections for statistical artifacts). Corrections for measurement error variance are commonly applied in industrial-organizational (I-O) psychology meta-analyses,

Corresponding Author:
Brenton M. Wiernik, Department of Psychology, University of South Florida, Tampa, FL 33620
E-mail: brenton@wiernik.org

but rarely applied in meta-analyses in other psychology subfields or other disciplines (Schmidt & Hunter, 2015; Schmidt, Le, & Oh, 2009). Corrections for selection effects are even rarer. They are typically performed only in meta-analyses of personnel selection and educational-admissions research (Dahlke & Wiernik, 2019a; Sackett & Yang, 2000). Measurement error variance and selection artifacts have severe biasing effects on the results of individual studies and meta-analyses. Failing to account for these artifacts can lead to inaccurate estimates and conclusions about the mean effect size and between-studies effect-size heterogeneity, and they can also influence the results of meta-regression, publication-bias analyses, and sensitivity analyses. When measurement error variance and selection effects are considered in meta-analyses outside of I-O psychology, they are often treated simply as indicators of general "study quality" and used as exclusion criteria or as moderators (Schmidt & Hunter, 2015, p. 485); both approaches are suboptimal, as measurement error variance and selection effects have predictable formulaic impacts on observed study results. Therefore, the best way to handle measurement error variance and selection effects is to apply statistical corrections that account for the known impacts of these artifacts. In this article, we describe the impacts of measurement error variance and selection effects on primary studies and meta-analyses, as well as methods to correct for these artifacts.

Disclosures

R code to reproduce the simulation results and corresponding figures presented in this article is available at https://osf.io/cp6rt/. R code to implement the correction and meta-analytic methods described in this article is available in the *psychmeta* package (Dahlke & Wiernik, 2019b).

Measurement Error Variance

Measurement error variance is an artifact that causes observed (i.e., measured) values to deviate from the "true" values of underlying latent variables (Schmidt & Hunter, 1996).[1] For example, consider a political psychologist assessing political orientation using a 10-item measure with items rated on a 7-point scale. Some respondents might obtain a mean score of 5 (somewhat conservative) or 3 (somewhat liberal) across the 10 items, when their true score is in fact 4 (moderate, centrist). Measurement error variance is also called unreliability (Rogers, Schmitt, & Mullins, 2002), observational error (Walter, 1983), and information bias (L. K. Alexander, Lopes, Ricchetti-Masterson, & Yeatts, 2014a). In the case of continuous variables, measurement error variance is also called low precision (Stallings & Gillmore, 1971); in the case of dichotomous or grouping variables, measurement error variance is also called misclassification (L. K. Alexander et al., 2014a).

Measurement error variance can come from a variety of sources, including truly random response errors (e.g., momentary distractions, arbitrary choices between adjacent scale points), transcription errors, transient or temporal effects (e.g., poor performance on a cognitive assessment due to fatigue on a particular day), environmental effects, content sampling effects (i.e., the specific items or content in a measure not functioning the same way as possible alternative items or content), rater effects (e.g., raters' differences in knowledge, motivation, or beliefs), and low sensitivity or specificity of measurement instruments, among others. These diverse sources of error can be grouped into four "classical" categories (Cronbach, 1947; Schmidt, Le, & Ilies, 2003; see Table 1 for descriptions).

It is important to note that measurement error variance concerns more than just whether participants' responses were correctly recorded; rather, measurement error variance is more about the process that generated the responses and whether the responses would be the same if the process were repeated at different times, by different raters, or with a different instrument or item set (Schmidt & Hunter, 1996). For psychological measures, random response error and transient error are typically the largest sources of measurement error (Ones, Wiernik, Wilmot, & Kostal, 2016).

The amount of measurement error variance in a sample of scores is quantified using a reliability coefficient, $r_{xx'}$, defined as the proportion of the observed-score variance, σ^2_{obs}, that is consistent (i.e., believed to be true), or 1 minus the proportion of the observed-score variance attributable to measurement error[2]:

$$r_{xx'} = \frac{\sigma^2_{\text{true}}}{\sigma^2_{\text{obs}}} = 1 - \frac{\sigma^2_{\text{error}}}{\sigma^2_{\text{obs}}} \tag{1}$$

Conceptually, the reliability coefficient is the correlation between two parallel measures of a construct. The square root of the reliability coefficient ($q_x = \sqrt{r_{xx'}}$) is the correlation between observed scores from the measured variable and true scores from its underlying latent variable (Schmidt & Hunter, 2015). This relationship is illustrated in Figure 1.

At the level of individual scores, random measurement error cannot be corrected, as the effect on a specific score is unknown; the best that can be done is to place a confidence interval around a score by using the reliability coefficient and the standard deviation of observed scores to estimate the standard error of

Table 1. Four Classical Sources of Measurement Error and Reliability Estimators Sensitive to Each

Source of error	Description, examples, and appropriate reliability estimators
Random response error	Truly random error specific to each item or response (e.g., random fluctuations in response time, momentary lapses in attention, transcription errors); unique to each measurement *Reliability estimators*: All reliability estimators
Transient error	Error due to the specific time or environment in which data are gathered (e.g., participant's mood, environmental distractions); shared by measures completed within a meaningfully short time span *Reliability estimators*: Test-retest reliability (coefficient of stability) Delayed parallel-forms reliability (coefficient of equivalence and stability) Delayed internal consistency (e.g., test-retest α; Green, 2003)
Content sampling error (item or test specific factor error)	Error due to the specific content used on a measure (e.g., participant's idiosyncratic interpretations of specific test items rather than their standing on the latent construct); shared by measures with the same or highly similar content *Reliability estimators*: Parallel-forms reliability (coefficient of equivalence) Delayed parallel-forms reliability (coefficient of equivalence and stability) Internal consistency (e.g., α, ω, split-half reliability, item response theory–based reliability estimates) Delayed internal consistency (e.g., test-retest α; Green, 2003)
Rater sampling error (rater specific factor error)	Error due to the specific raters used to gather data, including both rater main effects (e.g., effects of idiosyncratic beliefs, mood, and leniency) and Rater × Ratee interaction effects (e.g., due to differences in opportunities to observe); shared by measures with the same rater *Reliability estimators*: Interrater reliability (see Putka, Le, McCloy, & Diaz, 2008, for a discussion of estimators) Delayed interrater reliability

measurement (Dudek, 1979; Revelle, 2009). However, when scores are aggregated in data sets and are used to compute correlations between two variables or to compare scores across groups, measurement error variance has a known, predictable impact of biasing the effect size toward the null (i.e., toward zero). For example, consider the case shown in Figure 2. The scatterplots show hypothetical data for a researcher studying the relationship between intentions to recycle and actual recycling behavior, both measured using self-reports. The true correlation between these latent constructs, ρ, is .50, but the reliability of each measure, $r_{xx'}$, is .80. Consequently, the expected observed correlation, r_{xy}, is only .40. In other words, the expected observed correlation underestimates the true correlation between the latent constructs by .10, or 20%.[3]

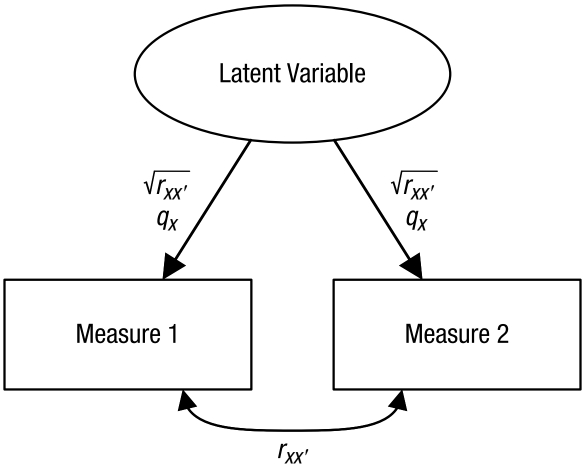

Fig. 1. Structural model illustrating the relationship among two parallel measures and their underlying latent variable. The correlation between the measures ($r_{xx'}$) is their reliability coefficient, and the correlation between each measure and the underlying latent variable is the square root of the reliability coefficient ($\sqrt{r_{xx'}}$), also referred to as q_x.

Impacts of measurement error variance in meta-analysis

Measurement error variance will impact the results of meta-analyses in three ways: by (a) biasing the mean effect size toward zero, (b) inflating effect-size heterogeneity and confounding moderator effects, and (c) confounding publication-bias and sensitivity analyses.

Mean effect size. When constructs are measured with random error, the mean effect size in a meta-analysis of relations between measures of these constructs will be biased toward the null (i.e., toward zero for *r* or *d* values). The amount of this null bias in the mean effect size is a function of mean reliability across studies. For example, if

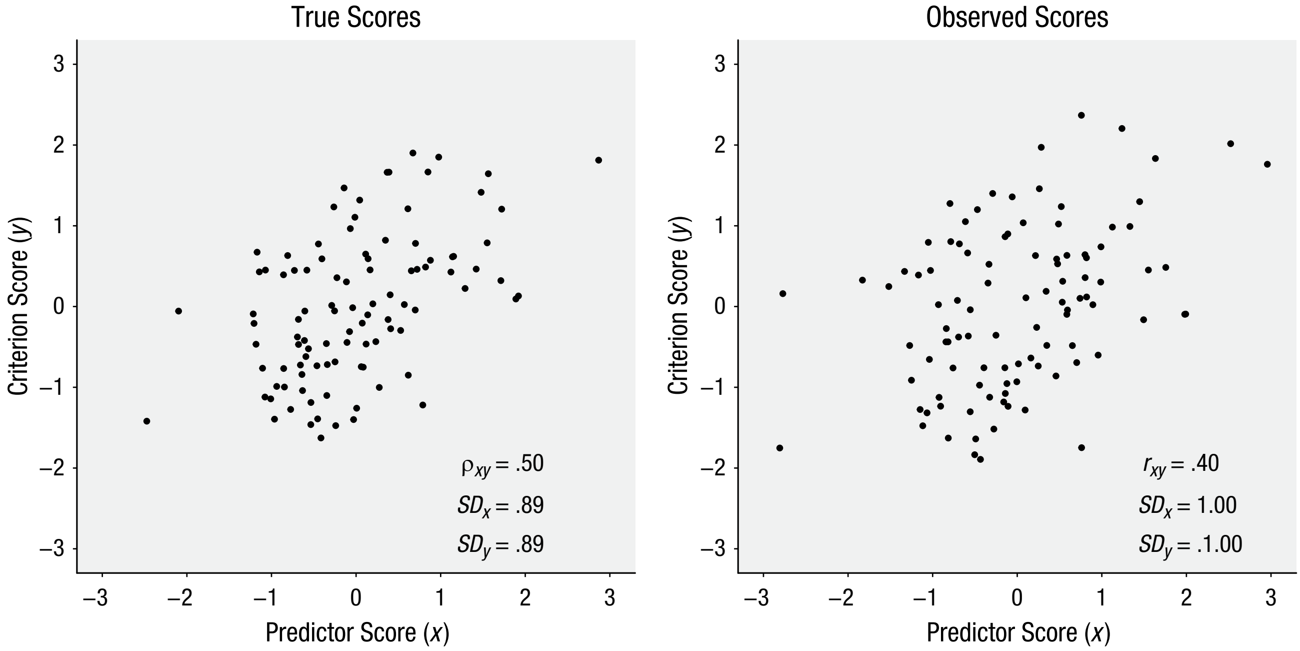

Fig. 2. Illustration of measurement error variance's impact on correlation coefficients. Both the predictor (*x*) and the criterion (*y*) measures have reliabilities, $r_{xx'}$, of .80 (q_x = .89). Consequently, although the true correlation between the constructs, ρ_{xy}, is .50 (left panel), the expected observed correlation, r_{xy}, is only .40 (right panel).

the true mean correlation between neuroticism and life satisfaction across 30 studies, $\bar{\rho}$, is −.32; the average reliability for neuroticism, $\bar{r}_{xx'}$, is .80 ($\sqrt{\bar{r}_{xx'}}$ = .89); and the average reliability for life satisfaction ($\bar{r}_{yy'}$) is .70 ($\sqrt{\bar{r}_{yy}}$ = .84), then the expected mean correlation between measures of neuroticism and life satisfaction will be only −.24 ($\bar{r}_{xy} = \bar{\rho} \times \sqrt{\bar{r}_{xx'}} \times \sqrt{\bar{r}_{yy'}}$ = −.32 × .89 × .84 = −.24); that is, the expected mean correlation will be .08, or 25%, smaller in magnitude than the true mean correlation between the latent constructs.

Effect-size heterogeneity and moderator effects. Measurement error variance can also bias estimates of the between-studies heterogeneity random-effects variance component (i.e., τ^2 in Hedges-Vevea notation; SD^2_{res}, SD^2_ρ, SD^2_δ, in Hunter-Schmidt notation; Hedges & Vevea, 1998; Schmidt & Hunter, 2015). If the studies included in a meta-analysis differ in their measures' reliabilities, heterogeneity estimates will be artifactually inflated, erroneously suggesting larger potential moderator effects. This is most serious if a moderator variable is correlated with measure reliability across studies. For example, a meta-analysis might compare the efficacy of exposure therapy and mindfulness practice for treating posttraumatic stress disorder (PTSD). If, in reality, these therapies are equally effective (e.g., $\bar{\delta}$ = .40), but PTSD symptoms are measured more reliably in studies of exposure therapy ($\bar{r}_{yy'}$ = .90) than in studies of mindfulness ($\bar{r}_{yy'}$ = .60), the observed effect sizes would falsely suggest that exposure therapy is more effective (observed $\bar{d}$ = .38) than mindfulness (observed $\bar{d}$ = .31).

Conversely, measurement error variance can also obscure true moderator effects that do exist. Measurement error variance will increase the standard errors of mean effect sizes in subgroup moderator analyses and moderator regression coefficients in meta-regression. These larger standard errors might lead meta-analysts to retain the null hypothesis of no moderator effects.

Publication-bias and sensitivity analyses. Measurement error variance can also cause publication-bias analyses to suggest the presence of bias when none exists. For example, consider a case in which meta-analysts compare published with unpublished studies and find that published studies have larger effect sizes (e.g., $\bar{r}_{xy}$ = .45 vs. .30). They might regard this finding as suggesting publication bias and conclude that the published studies substantially overestimate the true relationship between constructs. However, if the published studies also have better reliability than the unpublished studies (e.g., mean reliability = .90 vs. .60), an alternative explanation is that the unpublished studies were rejected because of the poor quality of their measures, rather than because of bias against null findings.

More seriously, cumulative meta-analysis, the precision-effect test and precision-effect-estimate-with-standard-errors

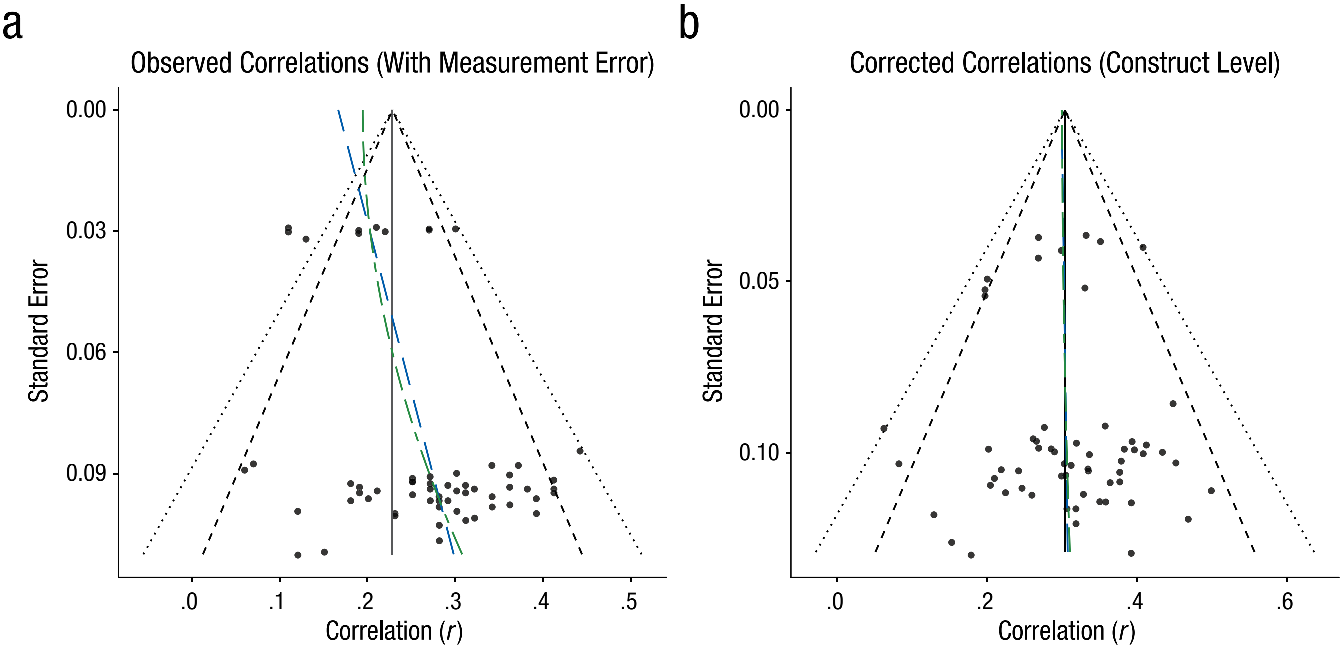

Fig. 3. Illustration of biasing effects of differential reliability on publication-bias analyses: (a) funnel plot of the relation between observed correlation effect sizes and standard error in published reports and (b) funnel plot of the same data after the correlations have been corrected for measurement error variance. Black vertical lines indicate mean effect sizes; black dashed lines indicate 95% confidence regions, and black dotted lines indicate 99% confidence regions. Blue long-dashed lines are precision-effect-test (PET) regression lines. Green dashed lines are precision-effect-estimate-with-standard-errors (PEESE) regression lines. The true correlations between the constructs have a mean, $\bar{\rho}$, of .50 (SD_ρ = .05). In reality, no publication bias is present. The large studies at the top of each funnel plot ($\bar{N}$ = 1,000, *SD* = 50) have low reliability ($\bar{r}_{xx'}$ = .50, *SD* = .10). The smaller studies below ($\bar{N}$ = 100, *SD* = 10) have high reliability ($\bar{r}_{xx'}$ = .80, *SD* = .10). The plot for the corrected correlations shows a high degree of symmetry. In contrast, because of differential reliability, the plot for the observed correlations (with measurement error variance) is asymmetric, erroneously suggesting publication bias.

estimator (PET-PEESE), selection models, and other procedures for correcting for publication bias will provide accurate results only if included studies have equal reliability or are corrected for measurement error variance (Carter, Schönbrodt, Gervais, & Hilgard, 2017; Schmidt & Hunter, 2015). If measure reliability is correlated with sample size, then small studies may have larger effect sizes for reasons other than publication bias. For example, a meta-analysis might find that smaller-*N* studies show a larger relationship between personality traits and mental-health outcomes. However, if these smaller studies used longer, more reliable personality scales, but the larger studies used less reliable, ultrashort scales (cf. Credé, Harms, Niehorster, & Gaye-Valentine, 2012), then this small-sample effect is the result of differences in measurement quality, not due to publication bias. This effect is illustrated in Figure 3. The funnel plot for observed correlations (Fig. 3a) is asymmetric. Fitting PET and PEESE regression models (see Carter et al., 2017, for details on computation and interpretation) shows a substantial relation of effect size with either standard error (PET: b_{SEi} = 1.19, 95% confidence interval = [0.53, 1.86]) or sampling error variance (PEESE: b_{Vi} = 9.35, 95% CI = [4.04, 14.66]). These results would typically be interpreted as indicating at least some effect-size inflation due to publication bias or questionable research practices. Conversely, the funnel plot for corrected correlations (Fig. 3b) is more symmetric, and the PET and PEESE models show negligible relations of effect size with either standard error (PET: b_{SEi} = 0.06, 95% CI = [−0.60, 0.73]) or sampling variance (PEESE: b_{Vi} = 0.62, 95% CI = [−3.78, 5.01]).

In short, differential reliability can distort inferences about publication bias by weakening and distorting observed relations of effect sizes with standard errors or other metrics.

Correcting for measurement error variance

As noted earlier, measurement error poses a serious threat to the accuracy of conclusions drawn from primary studies and meta-analyses. Researchers are well advised to reduce these biases by using more reliable scales. However, measurement error variance can also be statistically corrected post hoc to estimate the true relation between latent constructs, without measurement error variance. In this section, we discuss impacts of measurement error variance on correlations and group comparisons, and present methods for correcting for measurement error variance for different research designs.

Applying corrections for measurement error variance will generally yield less biased estimates of relationships between constructs, particularly in meta-analyses, but corrections for measurement error do not come without a cost. Applying statistical corrections increases the sampling error in the corrected effect sizes, leading to larger standard errors and wider confidence intervals. It is always better to reduce measurement error by using more reliable measurement procedures during data collection, rather than to rely on statistical corrections after the fact (Oswald, Ercan, McAbee, Ock, & Shaw, 2015).

Correcting for measurement error variance in correlations. When a meta-analysis cumulates correlation coefficients, the correlation will be attenuated by measurement error variance in both variables. This is the case for both psychological scales (e.g., personality measures) and other variables (e.g., course grades, self-reported eating behavior, objectively measured exercise behavior, clinician-rated symptoms). Even demographic variables are subject to measurement error variance, such as incorrectly rounding ages, mismarking responses, or adjusting one's responses depending on whether or not one is allowed to endorse multiple races or ethnicities or to endorse a nonbinary gender. The amount of attenuation in a correlation is a multiplicative function of the square roots of the reliabilities of the two measures:

$$r_{\text{obs}} = r_{\text{true}}\sqrt{r_{xx'}}\sqrt{r_{yy'}} \tag{2}$$

To correct for this attenuation due to measurement error variance, divide the observed correlation by the product of the square root of the reliabilities (Spearman, 1904)[4]:

$$r_c = \frac{r_{\text{obs}}}{\sqrt{r_{xx'}}\sqrt{r_{yy'}}} \tag{3}$$

This formula is analogous to estimating the correlation between latent variables in a structural equation model when the specified model has simple structure (Westfall & Yarkoni, 2016).

Correcting for measurement error variance in group-comparison research.

Measurement error in the dependent variable. Measurement error variance is also an issue in group-comparison research (e.g., studies of gender differences or studies comparing people with and without a psychological diagnosis). In these studies, measurement error variance in the dependent variable has an effect on the standardized mean difference (Cohen's *d*) that is similar to the effect of measurement error variance on correlations. Cohen's *d* is calculated as

$$d_{\text{obs}} = \left(Mean_1 - Mean_2\right)/\sqrt{\frac{\left(N_1-1\right)Var_1 + \left(N_2-1\right)Var_2}{N_1+N_2-2}}, \tag{4}$$

where Var_1 and Var_2 are the variances for Groups 1 and 2, respectively. The value under the square-root operator in Equation 4 is the weighted average within-group variance (Var_{pooled}), so the equation can be written more simply as

$$d_{\text{obs}} = \left(Mean_1 - Mean_2\right)/\sqrt{Var_{\text{pooled}}} \tag{4a}$$

Random measurement error in the dependent variable does not affect the difference between group means, but it does increase the magnitude of the within-group variances:

$$Var_1 = Var_{\text{true}_1}/r_{yy'_1}$$

$$Var_2 = Var_{\text{true}_2}/r_{yy'_2} \tag{5}$$

Note that dividing by $r_{yy'}$, a number less than 1, will increase the value of Var_{true}. These increased group variances cause the denominator of Equations 4 and 4a to be too large:

$$d_{\text{obs}} = (Mean_1 - Mean_2)/\sqrt{\frac{(N_1-1)(Var_{\text{true}_1}/r_{yy'_1}) + (N_2-1)(Var_{\text{true}_2}/r_{yy'_2})}{N_1+N_2-2}} \tag{6}$$

$$d_{\text{obs}} = (Mean_1 - Mean_2)/\sqrt{Var_{\text{true}_{\text{pooled}}}/r_{yy'_{\text{pooled}}}}, \tag{6a}$$

where $r_{yy'_{\text{pooled}}}$ is the sample-size-weighted average within-group reliability coefficient for the dependent variable. By increasing the size of the denominator used to standardize the *d* value, measurement error variance in the dependent variable systematically biases d_{obs} toward zero as a function of the square root of the dependent variable's reliability:

$$d_{\text{obs}} = d_{\text{true}}\sqrt{r_{yy'_{\text{pooled}}}} \tag{7}$$

Measurement error variance in the dependent variable can be corrected using a correction similar to that for correlations[5]:

$$d_c = d_{\text{obs}}/\sqrt{r_{yy'_{\text{pooled}}}} \tag{8}$$

If estimates of $r_{yy'_{\text{pooled}}}$ or the within-group reliability coefficients are unavailable, measurement error variance in the dependent variable can also be corrected using $r_{yy'}$, the reliability coefficient for the dependent variable computed in the combined total sample. The total-sample reliability coefficient includes between-group variance, so it will generally be larger than the pooled within-group reliability coefficient (i.e., $r_{yy'} > r_{yy'_{pooled}}$; Waller, 2008). A three-step procedure is used to correct d using the total-sample $r_{yy'}$ (cf. Schmidt & Hunter, 2015). First, convert d to a point-biserial correlation coefficient:

$$r_{\text{obs}} = d_{\text{obs}} / \sqrt{\frac{1}{p(1-p)} + d_{\text{obs}}^2}, \qquad (9)$$

where p is the proportion of the sample in one group. Second, correct r_{obs} using Equation 3: $r_c = r_{\text{obs}} / \sqrt{r_{yy'}}$. Finally, convert r_c back to a d value:

$$d_c = \frac{r_c}{\sqrt{p(1-p)(1-r_c^2)}} \qquad (10)$$

This three-step approach is also used to correct d values for other statistical artifacts, as we discuss later in this article.

Group misclassification. Measurement error variance can also be present in the independent variable used in group comparisons. This occurs when individuals are misclassified as being a member of one group when they are actually members of another group (Wacholder, Hartge, Lubin, & Dosemeci, 1995). For example, in a study comparing coping behaviors of people with and without a specific disorder, some participants in the no-disorder group may actually have the disorder but be undiagnosed. Similarly, in a study comparing substance users with nonusers, some participants may be unwilling to respond honestly when asked about their substance use. The impact of misclassification on observed d values is illustrated in the simulated data in Figure 4. In this case, a misclassification rate of 20% reduced the magnitude of the group difference from a true d value of 1.01 to an observed d value of only 0.72, a reduction by .29, or 28.7%. Expected magnitudes of attenuation in d values caused by differing degrees of misclassification and varying levels of dependent-variable reliability are shown in Table 2.

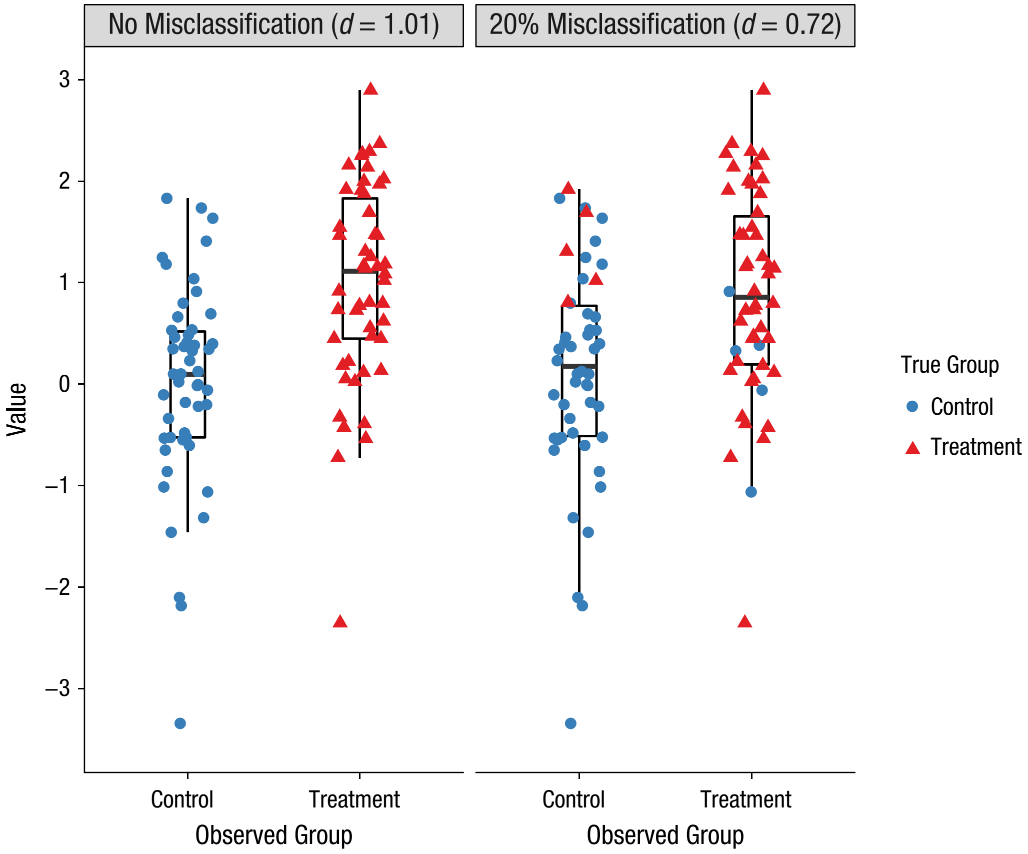

Fig. 4. Illustration of the impact of misclassification on d values. Even with no measurement error variance in the dependent variable, the misclassification rate of 20% (10% of participants in each group misclassified) leads to nearly a 30% reduction in the observed d value.

Table 2. Attenuation of *d* Values for Varying Proportions of Group Misclassification and Varying Levels of Dependent-Variable Reliability, Given a δ Parameter of 1.00

Dependent-variable reliability	Proportion misclassified										
	.00	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
1.00	1.00	0.77	0.56	0.36	0.18	0.00	−0.18	−0.36	−0.56	−0.77	−1.00
.90	0.94	0.72	0.53	0.34	0.17	0.00	−0.17	−0.34	−0.53	−0.72	−0.94
.80	0.87	0.68	0.49	0.32	0.16	0.00	−0.16	−0.32	−0.49	−0.68	−0.87
.70	0.81	0.63	0.46	0.30	0.15	0.00	−0.15	−0.30	−0.46	−0.63	−0.81
.60	0.74	0.58	0.42	0.28	0.14	0.00	−0.14	−0.28	−0.42	−0.58	−0.74
.50	0.67	0.52	0.39	0.26	0.13	0.00	−0.13	−0.26	−0.39	−0.52	−0.67
.40	0.59	0.46	0.34	0.23	0.11	0.00	−0.11	−0.23	−0.34	−0.46	−0.59
.30	0.51	0.40	0.30	0.20	0.10	0.00	−0.10	−0.20	−0.30	−0.40	−0.51
.20	0.41	0.32	0.24	0.16	0.08	0.00	−0.08	−0.16	−0.24	−0.32	−0.41

Note: The values presented are the observed *d* values. The sizes of the treatment and control groups are assumed to be equal, as are the misclassification rates for the two groups.

The impact of group misclassification on *d* values is a function of r_{gG}, the correlation between observed group membership (g) and actual group membership (G)[6]:

$$d_{\text{obs}} = r_{gG} r_{\text{obs}} / \sqrt{p(1-p)(1-r_{gG}^2 r_{\text{obs}}^2)}, \tag{11}$$

where r_{obs} is the point-biserial correlation calculated using Equation 9.

To correct a *d* value for group misclassification, use the three-step procedure described earlier (cf. Schmidt & Hunter, p. 262). First, convert d_{obs} to r_{obs} using Equation 9. Second, correct r_{obs} using the same disattenuation formula described in Equation 3:

$$r_c = r_{\text{obs}} / r_{gG} \tag{12}$$

Note that r_{obs} is divided by r_{gG} itself, not the square root of r_{gG}, because r_{gG} is analogous to the square root of the reliability (the correlation between the measured variable and the latent construct). Third, convert r_c back to d_c using a variant of Equation 10[7]:

$$d_c = \frac{r_c}{\sqrt{p^{\text{true}}(1-p^{\text{true}})(1-r_c^2)}}, \tag{13}$$

where p^{true} is an estimate of the true group proportion (without misclassification). If p^{true} is unknown, the observed group proportion, p, can be used with the assumption that misclassification is equal across groups.

You can estimate r_{gG} by conducting a study to quantify the accuracy of observed group classifications. For example, in the case of diagnosis for a disorder, epidemiological data could be consulted to estimate rates of overdiagnosis and underdiagnosis of the disorder. For the example of substance users, self-report accuracy could be assessed using chemical drug tests in a subset of the sample (or in a similar external example). With this information, r_{gG} is computed as the phi coefficient between observed and actual group membership:

$$r_{gG} = \sqrt{\chi_{gG}^2 / N}, \tag{14}$$

where χ_{gG}^2 is the chi-squared statistic for the 2 × 2 contingency table for group membership and *N* is the total sample size. The value of r_{gG} can also be computed directly:

$$r_{gG} = \frac{TP - BR_{\text{obs}} BR_{\text{true}}}{\sqrt{BR_{\text{obs}}(1 - BR_{\text{obs}}) BR_{\text{true}}(1 - BR_{\text{true}})}}, \tag{15}$$

where *TP* is the proportion of true positives (individuals correctly assigned to one of the two groups) in the total sample, BR_{obs} is the base-rate proportion of the sample observed to be part of that group, and BR_{true} is the base-rate proportion of the sample actually part of that group.

If the group sizes and misclassification rates are similar for the two groups, this correlation can be approximated using the proportion of correctly classified individuals in the full sample:

$$\begin{aligned} r_{gG} &\approx 1 - 2 \times (1 - \text{accuracy rate}) = \\ &1 - 2 \times \text{proportion misclassified} \end{aligned} \tag{16}$$

When the group sizes or misclassification rates for treatment and control groups are uneven, using Equation 16 will overestimate r_{gG} and thus result in undercorrection of the *d* value for misclassification. This undercorrection can be severe when the group sizes differ greatly (Chicco, 2017). In these cases, Equation 14 or 15 will be more accurate.

Correcting for measurement error variance in experimental research. The same principles described for observational research apply to group differences calculated in experimental and treatment research. Measurement error variance in both the manipulation or intervention and the dependent variable will attenuate observed effect sizes.

Measurement error variance in the dependent variable. In experimental research, dependent-variable constructs can be measured in a variety of ways, including responses to scales or questionnaires, observers' ratings of participants' behaviors, or performance on trials of behavioral tasks. Each of these methods can produce multiple types of random measurement error that can be accounted for by specific types of reliability coefficients.

Measurement error variance in the dependent variable: scales or questionnaires. If the dependent variable is participants' responses to a scale or questionnaire, reliability can be estimated using internal consistency (α, ω, test-information-based reliability from item response theory, etc.), parallel-forms reliability, or test-retest reliability for the scale, depending on which sources of measurement error variance are most important (see Choosing an Appropriate Reliability Estimate for Corrections). To correct for unreliability, calculate the reliability coefficient within each group, then compute the sample-size-weighted average within-group reliability, $r_{yy'_{\text{pooled}}}$. Finally, correct the d value for the experimental effect using $r_{yy'_{\text{pooled}}}$ in Equation 8. Alternatively, correct for unreliability by using the total-sample reliability, $r_{yy'}$: Convert the d value to r, apply the correction, and convert r_c back to d, as shown in Equations 9 and 10.

Measurement error variance in the dependent variable: multiple trials. In experiments in which participants complete multiple trials of a task (e.g., several trials in a Stroop task) and group comparisons are made using participants' average performance across trials, the key source of measurement error is random response error (Cooper, Gonthier, Barch, & Braver, 2017). For example, a participant may randomly have a somewhat longer reaction time on one trial than on the next. For these designs, reliability can be estimated using coefficient α (treating each trial as an item) or the Spearman-Brown-corrected split-half reliability, r_{SBSH}:

$$r_{SBSH} = (2 \times r_{h1h2})/(1 + r_{h1h2}), \tag{17}$$

where r_{h1h2} is the correlation of a participant's mean on one half of the trials with their mean on the other half of the trials. The Spearman-Brown-corrected split-half reliability is useful because it can be calculated if participants complete varying numbers of trials (e.g., if only trials with correct responses are retained). It is also appropriate if different participants complete trials for the stimuli in different orders and order or fatigue effects are a concern. Values of r_{SBSH} can vary depending on how trials are split (e.g., odd vs. even, first half vs. last half), so r_{SBSH} should be calculated by creating a large number of random splits (e.g., 5,000), calculating r_{h1h2} for each split, and then using the average r_{h1h2} across splits to compute r_{SBSH}. The *splithalf* package in R (Parsons, Kruijt, & Fox, 2019) provides functions to facilitate calculating r_{SBSH} for multitrial data.

If the dependent variable is the difference between participants' average performance in two conditions (e.g., congruent vs. incongruent trials), then reliability must be estimated for these difference scores, not for the individual condition averages. Generally, difference scores are much less reliable than averages of single scores (Williams & Zimmerman, 1977). To calculate r_{SBSH} for difference scores, divide the trials for each condition in half and calculate difference scores using these half-sets of trials (i.e., $D_{h1} = A_{h1} - B_{h1}$, where A_{h1} is a participant's mean performance on half the trials in Condition A, B_{h1} is the participant's mean performance on half the trials in Condition B, and D_{h1} is the participant's difference score for half of the trials). Calculate r_{h1h2} as the correlation between the difference scores for the two halves of the trials and enter this value in Equation 17.

To correct for measurement error variance in the experimental d value, calculate the reliability coefficient (α or r_{SBSH}) within each experimental group, then compute the sample-size-weighted average within-group reliability, $r_{yy'_{\text{pooled}}}$. Finally, correct the d value for the experimental effect using $r_{yy'_{\text{pooled}}}$ in Equation 8.

Measurement error variance in the dependent variable: single outcomes. If the dependent variable in an experiment is performance on a single-trial behavioral task (e.g., the length of time a participant persists in attempting to solve unsolvable anagrams, whether a participant donates to a charity, the amount of material recycled by an office during a month), a key question is how consistent performance on this task would be even if the treatment condition did not change. If performance would be inconsistent (i.e., there is a high degree of random response and transient error), the effect of the experimental manipulation on the construct underlying task performance will be underestimated. In these cases, researchers can estimate test-retest reliability by having members of the control group complete the measure again at a later time and using this within-group reliability value ($r_{yy'_{\text{control}}}$) to correct the experimental d value using Equation 8.

Measurement error variance in the dependent variable: observer ratings. If the dependent variable is observers' ratings of participants' behaviors (e.g., ratings of participants' aggressive behaviors), the key source of measurement error variance is rater-specific error. In these cases, interrater reliability is the most appropriate method to correct for measurement error. Estimate interrater reliability using an appropriate intraclass correlation, ICC(1), based on the number of raters and the measurement design (see Putka, Le, McCloy, & Diaz, 2008, for a discussion of considerations for estimating interrater reliability, including the use of reliability estimates based on generalizability theory to account for ill-structured measurement designs). Calculate the reliability coefficient within each group and the sample-size-weighted average within-group reliability, $r_{yy'_{\text{pooled}}}$. Then, use $r_{yy'_{\text{pooled}}}$ and Equation 8 to correct the experimental d value. Note that interrater reliability will not capture inconsistency in participants' actual responses over time. To correct for this form of measurement error, an approach similar to that for single outcomes can be employed. Have the control group complete the task a second time, then calculate the ICC(1) for different raters' rating at different times. Use this estimate of $r_{yy'_{\text{control}}}$ in Equation 8, to correct the experimental d value.

Measurement error in experimental manipulations. Measurement error in an experimental manipulation can occur if participants differ in their attentiveness or responsiveness to the manipulation. For example, in a study using vignettes to study attributions of responsibility for accidents, some participants may not read the vignettes carefully and thus not be affected by differences in situational features across conditions. Similarly, in a study on responses to a confederate's aggressive behaviors (as compared with responses to a confederate's nonaggressive behaviors in a control condition), participants may differ in how they evaluate the behaviors. In such cases, members of the experimental group might effectively function as though they were actually members of the control group, or vice versa. In intervention or treatment studies, there may be differences in treatment strength across participants due to treatment noncompliance or logistical challenges. For example, in a study evaluating a mindfulness intervention in which members of the treatment group are supposed to complete a daily mindfulness exercise for 14 days, they may complete the exercise only on some of the intended days, so that received dosage varies within the treatment group. Treatment measurement error variance might also occur if members of the control group are inadvertently exposed to the treatment. For example, in the mindfulness study, some control-group members might engage in their own mindfulness practice outside of the experiment.

Experimental researchers are often interested in the relationship of the dependent variable with the construct targeted by a manipulation or intervention (e.g., perceived aggression or mindfulness), rather than merely the impact of assigning participants to a treatment intended to influence this target construct. This distinction is analogous to the distinction in clinical research between treatment "as assigned" and treatment "as received" (Ten Have et al., 2008). If there is substantial within-group heterogeneity in attentiveness to, interpretation of, or compliance with a manipulation, this can strongly bias the experimental d value as an index of the relationship between the target construct and the dependent variable. If a valid manipulation check or compliance check is available, this bias can be estimated or corrected using several methods for experimental causal mediation analysis (Imai, Keele, & Tingley, 2010). For example, in the aggression study, participants could be asked to rate how aggressive they perceived the confederate's behavior to be. In the mindfulness study, participants could be asked to report how often they engaged in any sort of mindfulness or meditation practice during the study period.

When a manipulation or compliance check is available, the most straightforward correction method for correction is the instrumental-variable estimator (Angrist, Imbens, & Rubin, 1996). The reliability of treatment is estimated as r_{gM}, the correlation between assigned group membership (g) and the measured manipulation or compliance check (M). To correct an experimental d value for measurement error variance due to differential attentiveness to or compliance with treatment, use the three-step procedure described earlier. First, convert d_{obs} to r_{obs} using Equation 9. Second, correct r_{obs} using the same type of disattenuation formula described in Equation 12:

$$r_c = r_{\text{obs}} / r_{gM} \tag{18}$$

Note that r_{obs} is divided by r_{gM} itself, not its square root. Third, convert r_c back to d_c using Equation 10.

The instrumental-variable estimator relies on three key assumptions (Imai et al., 2010; cf. McNamee, 2009; Mealli & Rubin, 2002):

1. The manipulation or compliance check fully accounts for the effect of the treatment on the dependent variable (i.e., that treatment effect is fully mediated through the manipulation check; this is the *exclusion restriction assumption*);
2. The treatment affects the manipulation or compliance check in the same direction for all participants (e.g., no one misinterprets the aggressive confederate as behaving *less* aggressively than the nonaggressive confederate; no participants assigned to the mindfulness intervention practices *less* mindfulness than they would have otherwise; this is the *monotonicity assumption*);

3. There is no interaction between the treatment and the manipulation check (e.g., the manipulation-check procedure does not enhance the treatment effect by clueing participants in on a concealed purpose of the study; Hauser, Ellsworth, & Gonzalez, 2018; this is the *noninteraction assumption*).

If these assumptions are unreasonable, other estimators with different assumptions can be used (Imai et al., 2010), but these are less readily applied in meta-analysis.[8]

Choosing an appropriate reliability estimate for corrections.

What are the major sources of measurement error variance? As shown in Table 1, there are a variety of different types of reliability coefficients, sensitive to different sources of measurement error variance. When choosing a reliability coefficient to correct for measurement error, it is important to carefully consider which sources of error are likely to have had major impacts on the measures and to select a method that is sensitive to these sources (for overviews, see Revelle, 2009; Schmidt & Hunter, 1996; Schmidt et al., 2003). For example, in a study using a personality scale to predict supervisor-rated job performance, a major source of error will be the supervisors' idiosyncratic beliefs and ability to observe; interrater reliability is thus an appropriate method to capture this source of error (Connelly & Ones, 2010; Viswesvaran, Ones, & Schmidt, 1996). In a study predicting life satisfaction, transient effects, such as participants' mood or temporary circumstances, will be a major source of error; test-retest reliability is an appropriate method to capture this source of error (Green, 2003; Le, Schmidt, & Putka, 2009; Schmidt et al., 2003). Although internal-consistency statistics such as coefficient α are the most commonly reported reliability estimates, in many cases these will not capture critical sources of measurement error; when correlations are corrected for internal consistency alone, the true correlations between constructs can be substantially misestimated.

Internal-consistency methods can also be inappropriate even if content sampling error is a major concern. As an estimate of reliability, internal consistency assumes *item homogeneity*—that all items are indicators of the same underlying construct and that the reliability of items as indicators of this construct is reflected solely by the covariance among items, not the items' unique variance. Some measures are heterogeneous composites of items with nonredundant content. For example, a biodata inventory designed to predict borderline personality disorder (BPD) might assess a diverse set of life experiences that are each related to BPD but not highly correlated with one another. Such a measure would have low internal consistency but could still be reliable as a measure of overall BPD risk due to life experiences. For these types of measures, parallel-forms reliability and test-retest reliability are more appropriate reliability estimators, provided the lag between measurements is long enough to account for transient effects yet short enough that participants do not change systematically (e.g., BPD symptoms could be mitigated by therapy between temporally distant testing occasions, which would undermine a reliability estimate). Short-form measures of personality traits and other constructs are related examples. Many short-form measures of constructs include items chosen to tap a construct broadly and with limited redundancy (Rammstedt & John, 2007; Yarkoni, 2010). By removing redundancy, these measures will show low internal consistency, but they can still show high convergent correlations with longer measures. For short-form measures, parallel-forms and test-retest reliability are also more appropriate methods to estimate reliability.[9]

What if reliability estimates are not available for a sample? Researchers conducting a meta-analysis are likely to find that for many included studies, authors do not report reliability estimates for measures or experimental manipulations or do not report the most appropriate form of reliability (e.g., the authors report coefficient α rather than test-retest or interrater reliability; Flake, Pek, & Hehman, 2017; Fried & Flake, 2018). Even in primary studies, it is often not possible to estimate a relevant form of reliability with a specific sample. To correct for measurement error variance in these cases, it is necessary to draw an estimate from some other source. One approach is to draw reliability estimates from external sources, such as test manuals, previous meta-analyses, or previous studies using the same measure or manipulation. This approach is reasonable only if it can be assumed that the sample (or samples) in the present study and the external source (or sources) were drawn from the same underlying population (or at least comparable populations). Another approach is to conduct one or more new studies specifically for the purpose of estimating relevant reliability coefficients. For example, Beatty, Walmsley, Sackett, Kuncel, and Koch (2015) conducted a large-scale study to estimate reliability coefficients for university grade point averages that are specifically based on the number of courses taken. In meta-analyses, if reliability information is available for some studies but not others, artifact-distribution and reliability-generalization methods can also be used to address missing reliability data in individual studies (see Correcting for Artifacts in Meta-Analysis).

When should you correct for measurement error variance? Whether and how to correct for measurement error variance is the subject of ongoing debate in many areas of psychology and other fields (Muchinsky,

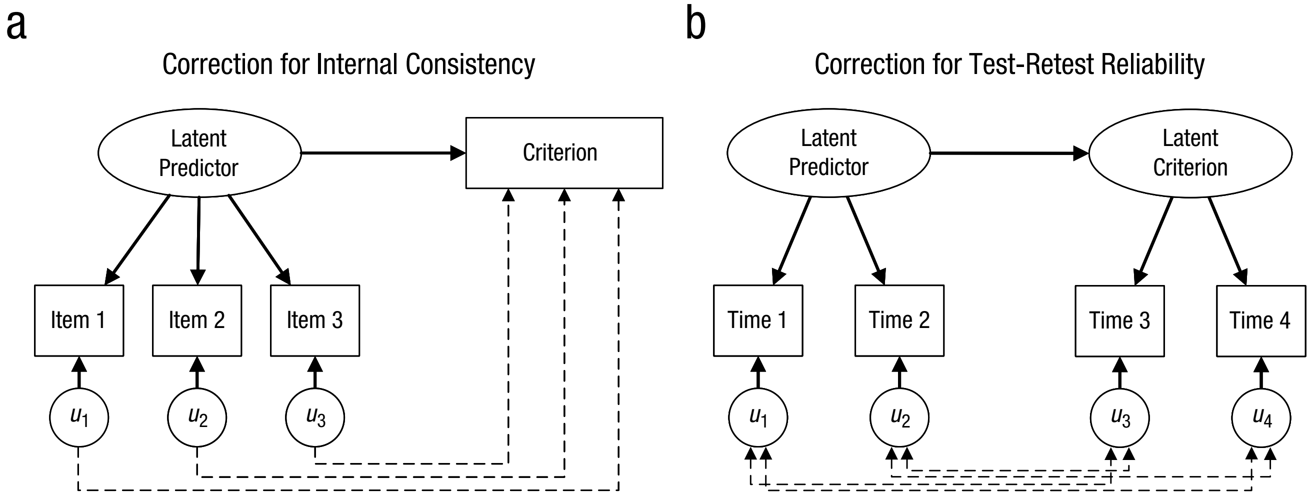

Fig. 5. Structural models illustrating the assumptions of Spearman's disattenuation formula for two types of reliability coefficients: (a) internal-consistency coefficients (e.g., coefficient α, coefficient ω) and (b) the coefficient of stability (e.g., test-retest reliability). Variables labeled *u* are the unique variances associated with each item (internal consistency) or each occasion (test-retest reliability). Dashed lines indicate key assumed-zero paths.

1996; Schmidt & Hunter, 1996; cf. Schennach, 2016, for more favorable attitudes in econometrics). The appropriateness of correcting for measurement error variance depends on the nature of the research question. Are you interested in understanding measures or their underlying constructs? For example, if the goal of a study is to evaluate the diagnostic accuracy of a tool for practical use, correcting for measurement error variance in the criterion variable would be appropriate, but correcting for measurement error variance in the diagnostic tool would overestimate its practical utility because practitioners have access only to patients' observed scores, not their standing on the underlying construct (Binning & Barrett, 1989). However, if the scientific goal of a study is to identify the nature of underlying constructs, relationships, and principles, then *not* correcting for measurement error variance can lead to highly inaccurate theoretical conclusions (Westfall & Yarkoni, 2016). Advancing scientific knowledge requires estimating measurement error variance and correcting for its impacts on research findings, especially in meta-analysis. For meta-analyses, even if the research question concerns observed measures, controlling for *differences* in reliability across studies is important to ensure accuracy of moderator and sensitivity analyses and to avoid erroneously interpreting differences across studies as substantive in origin (see Fig. 3).

The accuracy of corrections for measurement error variance depends on whether the assumptions of the model for measurement error variance used in the corrections are reasonably met. One important assumption of the correction procedures we have described is that error scores for the independent variable are unrelated to the true scores for the dependent variable (and vice versa). A structural model illustrating this assumption for internal-consistency corrections is shown in Figure 5a. Correcting for internal consistency assumes that the dependent variable is predicted only by the interitem covariance of the independent variable (i.e., the variance the items have in common). If individual items in the independent variable have unique predictive power for the dependent variable, then the corrected correlation will either overestimate or underestimate the true correlation (Rhemtulla, van Bork, & Borsboom, 2019; cf. Putka, Hoffman, & Carter, 2014, for similar issues regarding interrater reliability). In these cases, correcting for test-retest reliability instead may yield a more accurate estimate of the true correlation between latent constructs, as the transient errors in observed measures are less likely to have unique predictive power.

This assumption can also be a particular concern for corrections for group misclassification. For example, in a study comparing impulsivity in substance-use patients who report resuming use with impulsivity among patients who report remaining sober, it may be that more impulsive patients are more likely to lie about resuming use. This will result in an underestimate of the correlation between relapse and impulsivity. In this case, the correction in Equation 12 will undercorrect, and the corrected correlation will still be a negatively biased estimate of the true relationship between these constructs.

A second important assumption of the procedures we have described is that error scores for the two variables are unrelated. If errors for the two variables are correlated, the corrections will overestimate the true relationship between the two constructs. Figure 5b illustrates this assumption for test-retest reliability corrections. If the predictor variable (e.g., a personality trait) and the criterion variable (e.g., subsequent life satisfaction) are measured at different times, then the two measures will not share transient errors, and correcting for test-retest reliability is appropriate. If the two variables are measured at roughly the same time, however, then the measures might share transient errors. In this case, the observed correlation between the two measures will reflect not only covariance between the latent constructs, but also covariance between the shared transient errors. For example, stressful life events might cause an employee to be more likely to endorse items measuring both anxiety and job burnout. If anxiety and job burnout are measured at the same time, such that responses to both sets of items are affected by the same transient events, this would inflate the observed correlation between these two variables. As a result, correcting using test-retest reliability will overestimate the true relationship between constructs. In this case, parallel-forms reliability or internal consistency would be a more appropriate reliability method to use in correcting for measurement error variance.[10]

When deciding whether to correct for measurement error variance and which reliability estimate to use, meta-analysts and primary researchers must carefully consider whether these assumptions are reasonable for the study design and the chosen reliability method. Alternative methods for estimating true-score correlations without these assumptions are also possible depending on the specific research design (see, e.g., Charles, 2005; Schennach, 2016; Zimmerman, 2007).

Finally, as noted previously, although corrections for measurement error variance can reduce bias in estimates of relationships between constructs, these corrections come with the cost of increased sampling error in the corrected effect sizes. The increased uncertainty in corrected effect sizes can be indexed by adjusting standard errors as shown in Table 3 or by applying the same correction to the endpoints of the effect-size confidence interval as is applied to the point estimate, for example,

$$\frac{r_{\text{LowerBound}}}{\sqrt{r_{xx'}}\sqrt{r_{yy'}}} \leq \frac{r_{xy}}{\sqrt{r_{xx'}}\sqrt{r_{yy'}}} \leq \frac{r_{\text{UpperBound}}}{\sqrt{r_{xx'}}\sqrt{r_{yy'}}} \tag{19}$$

When an effect size is corrected for measurement error variance, corresponding corrections must always be applied to the standard error and confidence interval for the effect size.

Selection Effects

Most psychologists are familiar with the concept of measurement error variance, even if corrections for measurement error are uncommon in primary studies and meta-analyses. Awareness of the impacts of diverse types of selection effects on study results, along with available methods to correct those study results, is much less widespread.[11] A selection effect is present when the relationship observed between two variables in a sample is not representative of the relationship in the target or reference population as a result of selection or conditioning on one or more variables (Sackett & Yang, 2000). In psychology, selection effects are most frequently discussed in terms of *range restriction*, that is, reduced variances of the variables in the sample relative to the target population. Range restriction biases effect sizes toward null or zero. However, this is only one of many possible effects of selection. Depending on the exact selection mechanism, selection effects can attenuate sample effect sizes, inflate them, or even reverse their direction (R. A. Alexander, 1990; Ree, Carretta, Earles, & Albert, 1994; Sackett, Lievens, Berry, & Landers, 2007). These diverse types of selection effects are variously called range restriction, range enhancement, range variation (Schmidt & Hunter, 2015), selection bias (L. K. Alexander, Lopes, Ricchetti-Masterson, & Yeatts, 2014b), and collider bias (Greenland, 2003; Rohrer, 2018).

The impacts of selection effects on correlations are illustrated in Figure 6. Selection effects can be direct (i.e., when one or both of the variables in a relationship is directly selected on or truncated). For example, only severely ill patients might be referred to an in-patient treatment center. In such a setting, any relations between potential antecedents and symptoms will likely be reduced because of limited symptom variability. This scenario is illustrated in Figure 6a. The correlation, r, between Z and X is .50 in the population, but is reduced to only .33 if Z is directly restricted to only the top 50% of scores. Direct selection can also increase observed correlations through range enhancement. For example, if a researcher uses an extreme-groups design and compares outcomes for the top 25% and bottom 25% of scorers on an attitudes measure, the variance on attitudes is artificially inflated, and the relationship between attitudes and outcomes will be overestimated relative to the relationship in the full population (Preacher, Rucker, MacCallum, & Nicewander, 2005; Wherry, 2014, p. 50).

Selection effects can also be indirect, or incidental (i.e., when there is selection on a third variable related to one or both of the two focal variables). This process is sometimes called *conditioning on a confounder* (Rohrer, 2018). For example, the relationship between overall health and life satisfaction may be reduced in a sample of professionals, as socioeconomic status is related to both health and life satisfaction (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). In Figure 6a, indirect selection on Z reduces the correlation, r,

Table 3. Formulas for Corrected Effect Sizes and Sampling Error Variances for Different Artifact-Correction Models

Artifact model and effect size	Corrected effect size	Corrected sampling error variance
Measurement error alone		
r	$r_c = r_{\text{obs}} / \left(\sqrt{r_{xx'}}\sqrt{r_{yy'}}\right)$	$SE^2_{r_c} = SE^2_{r_{\text{obs}}} \times \left(r_c / r_{\text{obs}}\right)^2$
d	With no misclassification: $d_c = d_{\text{obs}} / \sqrt{r_{yy'_{\text{pooled}}}}$	With no misclassification: $SE^2_{d_c} = SE^2_{d_{\text{obs}}} \times$ $(d_c / d_{\text{obs}})^2$
	With misclassification:	With misclassification:
	A. Convert *d* to *r* using $r = d / \sqrt{\frac{1}{p(1-p)} + d^2}$	A. Convert $SE^2_{d_{\text{obs}}}$ to $SE^2_{r_{\text{obs}}}$ using $SE^2_{r_{\text{obs}}} = \frac{SE^2_{d_{obs}}}{\left(1 + d^2_{\text{obs}} p[1-p]\right)^2 \left(d^2_{\text{obs}} + \frac{1}{p(1-p)}\right)}$
	B. Apply the correction for *r*, using the total-sample $\sqrt{r_{yy'}}$ and with $\sqrt{r_{xx'}} = r_{gG}$ or $\sqrt{r_{xx'}} = \sqrt{r_{tt'}}$	B. Calculate $SE^2_{r_c} = SE^2_{r_{\text{obs}}} \times \left(r_c / r_{\text{obs}}\right)^2$
	C. Convert r_c to d_c using $d_c = r_c / \sqrt{p^{\text{true}}\left(1 - p^{\text{true}}\right)\left(1 - r_c^2\right)}$	C. Convert $SE^2_{r_c}$ to $SE^2_{d_c}$ using $SE^2_{d_c} = \frac{SE^2_{r_c}}{p^{\text{true}}\left(1 - p^{\text{true}}\right)\left(1 - r_c^2\right)^3}$
	If p^{true} is unknown, use the observed *p* with the assumption of equal misclassification across groups.	
Univariate direct selection		
r	$r_c = \frac{r_{\text{obs}}}{u_x\sqrt{1 - u_x^2\left(1 - r_{xx'}\right)}\sqrt{\left(\frac{1}{u_x^2} - 1\right)r^2_{\text{obs}} + r_{yy'}}}$	$SE^2_{r_c} = SE^2_{r_{\text{obs}}} \times \left(r_c / r_{\text{obs}}\right)^2$
d	A. Convert *d* to *r* using $r = d / \sqrt{\frac{1}{p(1-p)} + d^2}$	A. Convert $SE^2_{d_{\text{obs}}}$ to $SE^2_{r_{\text{obs}}}$ using $SE^2_{r_{\text{obs}}} = \frac{SE^2_{d_{\text{obs}}}}{\left(1 + d^2_{\text{obs}} p[1-p]\right)^2 \left(d^2_{\text{obs}} + \frac{1}{p(1-p)}\right)}$
	B. Apply the correction for *r*, using the total-sample $\sqrt{r_{yy'}}$ and with $\sqrt{r_{xx'}} = r_{gG}$ or $\sqrt{r_{xx'}} = \sqrt{r_{tt'}}$	B. Calculate $SE^2_{r_c} = SE^2_{r_{\text{obs}}} \times (r_c / r_{\text{obs}})^2$
	C. Convert r_c to d_c using $d_c = r_c / \sqrt{p^*(1 - p^*)(1 - r_c^2)}$	C. Convert $SE^2_{r_c}$ to $SE^2_{d_c}$ using $SE^2_{d_c} = \frac{SE^2_{r_c}}{p^*(1 - p^*)(1 - r_c^2)^3}$
Univariate indirect selection		
r	$r_c = \frac{r_{\text{obs}}}{\sqrt{r^2_{\text{obs}} + \frac{u_x^2 r_{xx'}(r_{xx'} r_{yy'} - r^2_{\text{obs}})}{1 - u_x^2(1 - r_{xx'})}}}$	$SE^2_{r_c} = SE^2_{r_{\text{obs}}} \times (r_c / r_{\text{obs}})^2$

(continued)

Table 3. (Continued)

Artifact model and effect size	Corrected effect size	Corrected sampling error variance
d	A. Convert d to r using $r = d/\sqrt{\frac{1}{p(1-p)} + d^2}$ B. Apply the correction for r, using the total-sample $\sqrt{r_{yy'}}$ and with $\sqrt{r_{xx'}} = r_{gG}$ or $\sqrt{r_{xx'}} = \sqrt{r_{tt'}}$ C. Convert r_c to d_c using $d_c = r_c/\sqrt{p^*(1-p^*)(1-r_c^2)}$	A. Convert $SE^2_{d_{\text{obs}}}$ to $SE^2_{r_{\text{obs}}}$ using $SE^2_{r_{\text{obs}}} = \frac{SE^2_{d_{\text{obs}}}}{\left(1 + d^2_{\text{obs}}p[1-p]\right)^2 \left(d^2_{\text{obs}} + \frac{1}{p(1-p)}\right)}$ B. Calculate $SE^2_{r_c} = SE^2_{r_{\text{obs}}} \times (r_c/r_{\text{obs}})^2$ C. Convert $SE^2_{r_c}$ to $SE^2_{d_c}$ using $SE^2_{d_c} = \frac{SE^2_{r_c}}{p^*(1-p^*)(1-r_c^2)^3}$
Bivariate direct selection		
r	$r_c = \frac{\frac{r^2_{\text{obs}} - 1}{2r_{\text{obs}}} u_x u_y + \text{sign}(r_{\text{obs}})\sqrt{\frac{(1-r^2_{\text{obs}})^2}{4r^2_{\text{obs}}} u_x^2 u_y^2 + 1}}{\sqrt{1 - u_x^2(1-r_{xx'})}\sqrt{1-u_y^2(1-r_{yy'})}}$	$SE^2_{r_c} = SE^2_{r_{\text{obs}}} \times (r_c/r_{\text{obs}})^2$
Bivariate indirect selection		
r	$r_c = \frac{r_{\text{obs}} u_x u_y + \lambda\sqrt{\lvert 1-u_x^2 \rvert \lvert 1-u_y^2 \rvert}}{\sqrt{1-u_x^2(1-r_{xx'})}\sqrt{1-u_y^2(1-r_{yy'})}}$	See Dahlke and Wiernik (2019a) for details

Note: All reliability values in the formulas presented in this table are assumed to be estimated in the restricted sample (subject to selection effects). If the reliability values are estimated in samples from the unrestricted reference population, they should be adjusted using the following formula: $r^*_{xx'} = 1 - (1 - r_{xx'})/u_x^2$. For bivariate indirect selection, $\lambda = \text{sign}(\rho_{ZX}\rho_{ZY}[1-u_x][1-u_y]) \frac{\text{sign}(1-u_x)\min\left(u_x, \frac{1}{u_x}\right) + \text{sign}(1-u_y)\min\left(u_y, \frac{1}{u_y}\right)}{\min\left(u_x, \frac{1}{u_x}\right) + \min\left(u_y, \frac{1}{u_y}\right)}$, where ρ_{ZX} and ρ_{ZY} are the true correlations of the selection construct, Z, with X and Y, corrected for measurement error variance and selection effects. Note that only the −1 or +1 signs of ρ_{ZX} and ρ_{ZY} are needed to compute λ. Other symbols are defined in the text. Performance of bivariate-selection-correction methods for d values in meta-analysis has not yet been evaluated.

between X and Y from .50 to only .41. Indirect selection can even reverse the sign of a correlation if the two focal variables are strongly related to the selection mechanism (Ree et al., 1994). For example, imagine that success in a class is equally determined by ability and effort, and that these two predictors are uncorrelated (see Fig. 6b). If a sample of highly successful students is selected, ability and effort will become artifactually strongly negatively correlated because of indirect range restriction, or what is sometimes referred to as *conditioning on a collider* (Rohrer, 2018). This effect occurs because high effort can compensate for low ability (or vice versa) as a path to success.

Impacts of selection effects in meta-analysis

Selection effects influence meta-analysis results in the same three ways as measurement error variance does. The mean effect size in the observed samples will be biased relative to the effect size in the desired reference population; the magnitude of this bias is based on the average selection process across studies. Similarly, if selection processes or the size of selection effects differ across studies (e.g., if some studies were sampled from a general population, but others were restricted to university students; cf. Murray, Johnson, McGue, & Iacono, 2014), this will inflate the random-effects variance component and bias the results of moderator and publication-bias analyses.

Quantifying and correcting for selection effects

The impact of selection effects is quantified using a u ratio—a ratio of the standard deviation of a variable in the sample to the standard deviation of that variable in the reference population:

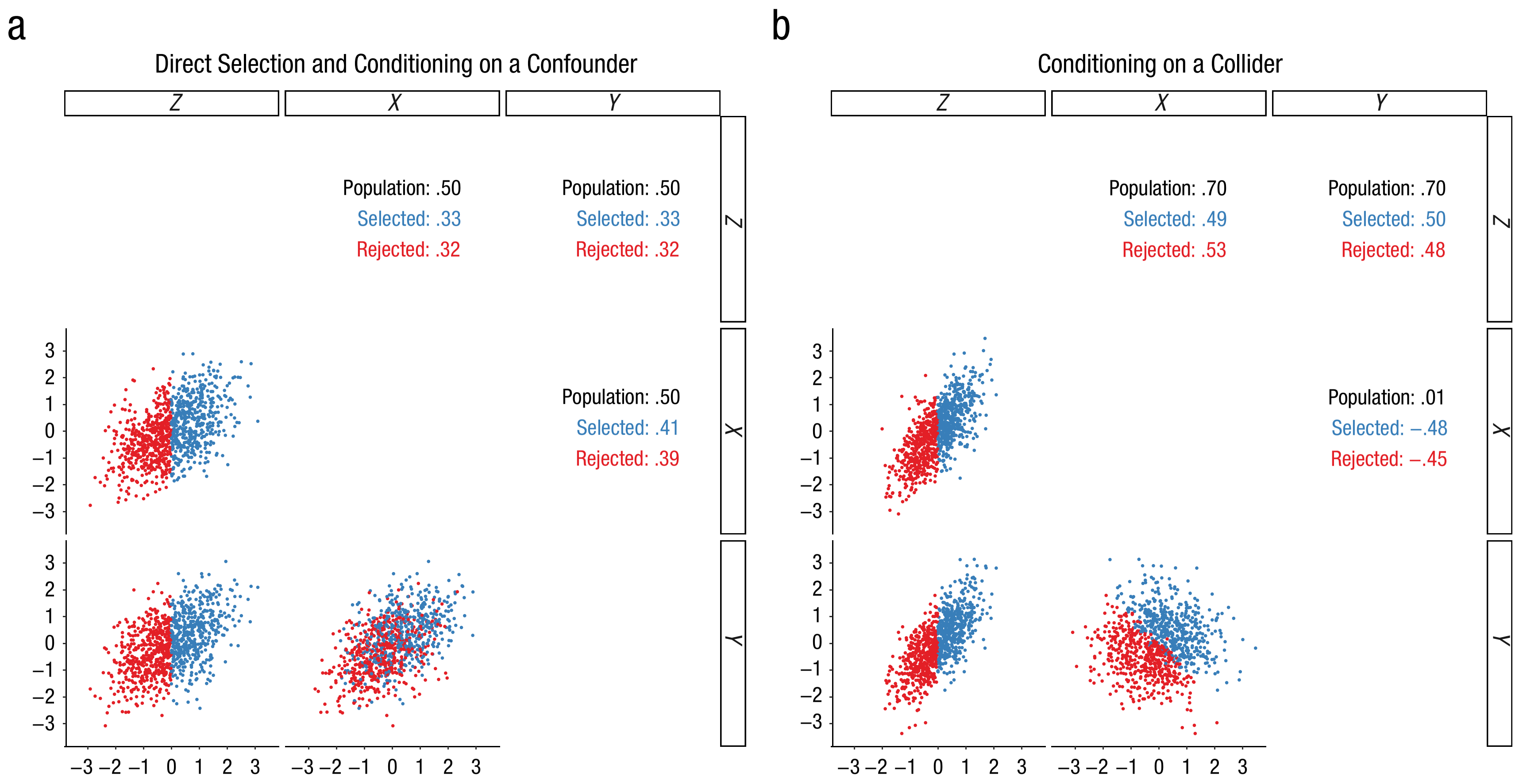

Fig. 6. Impact of direct range and indirect range restriction on correlations. In (a), *X*, *Y*, and *Z* are three distinct variables, and all have intercorrelations of .50. In (b), *X* and *Y* are uncorrelated, and *Z* is a composite of *X* and *Y* plus a small amount of error variance. In each panel, black correlations reflect the total population, blue correlations and data points reflect the selected group (the top 50% of scorers on *Z*), and red correlations and data points reflect the rejected group. Selected r_{XZ} and r_{YZ} are directly range restricted; selected r_{XY} is indirectly range restricted.

$$u_x = SD_{x_{\text{sample}}} / SD_{x_{\text{reference}}} \tag{20}$$

If a reference standard deviation is unavailable, the u ratio for a continuous variable can be estimated from the sample and reference reliability coefficients:

$$u_x = \sqrt{(1 - r_{xx'_{\text{reference}}})/(1 - r_{xx'_{\text{sample}}})} \tag{21}$$

If the exact selection mechanism is known, very accurate estimates of the population relationship can be calculated using a multivariate selection correction (Held & Foley, 1994; derived by Aitken, 1935; Lawley, 1944; based on Pearson, 1903).[12] Such complete information is rarely available in psychology research. However, there are several accurate approximations that rely on information more commonly available to primary researchers and meta-analysts; these are listed in Table 3. An appropriate correction model can be selected according to whether (a) direct selection occurs on one variable (univariate direct range restriction, UVDRR), (b) direct selection occurs on both variables (bivariate direct range restriction, BVDRR), or (c) indirect selection occurs and u ratios are available for one variable (univariate indirect range restriction, UVIRR) or both variables (bivariate indirect range restriction, BVIRR). For example, for the indirect selection in Figure 6b, the u ratio is .82 for X and Y. Using these values and the observed correlation, r_{obs}, of −.48 in the formula for the BVIRR correction, one obtains an estimated population correlation, r_c, of .005, a value that reverses sign and differs in magnitude by .475, or 99%.

An important nuance to consider is that selection never affects only one variable. When one variable is restricted, other variables with which it is correlated are also restricted. For example, grade point average will be less variable in a sample of university students selected on high school grade point average than in an unselected group. The correction formulas in Table 3 correct for these indirect selection effects on the basis of assumptions about the selection process (see Dahlke & Wiernik, 2019a, and Hunter, Schmidt, & Le, 2006, for discussions).

Selection effects in correlations

Selection effects are most commonly discussed in terms of correlations, and methods for correcting for selection effects have been developed directly for this metric. To correct for selection effects in correlations, identify an appropriate reference sample and the most likely selection mechanism. Then, apply the appropriate correction formula from Table 3.

Selection effects in group comparison and experimental research

Selection effects impact group comparisons in the same way as they do correlations (Bobko, Roth, & Bobko, 2001; Li, 2015). For example, a study of gender differences in political orientation may show attenuated mean differences in a sample of urban university students as a result of their reduced variability in political orientation compared with the general population (e.g., given students who are more liberal overall). In experimental research, d values for a manipulation or treatment may differ across studies simply because of differences in response variability across the studies.

To correct a d value for these selection and range-variation effects, use the three-step procedure described earlier. First, convert d_{obs} to r_{obs} using Equation 9:

$$r_{\text{obs}} = d_{\text{obs}} / \sqrt{\frac{1}{p(1-p)} + d_{\text{obs}}^2}$$

Second, correct r_{obs} using the appropriate formula from Table 3. For these corrections, u_y and $\sqrt{r_{yy'}}$ should be the combined total-sample (not pooled within-sample) values. Third, convert r_c back to the d metric, d_c, using a variant of Equation 10:

$$d_c = \frac{r_c}{\sqrt{p^*(1-p^*)(1-r_c^2)}}, \tag{22}$$

where p^* is an adjusted group proportion. As discussed earlier, direct selection on one variable also indirectly induces selection on other correlated variables. When applied to correlations, the equations in Table 3 implicitly correct for this indirect selection. However, this adjustment must be manually applied when converting back from r_c to the d_c metric. If the observed group proportions were used to convert back from r_c to d_c, this would imply no change in the grouping variable's variance after the correction, which is generally possible only if the d value equals 0 (i.e., if there is no correlation between group membership and score on the dependent variable).

The effect of a correction for selection effects on the variance of the dichotomous grouping variable can be estimated as

$$var_{g_c} = p(1-p)\left(1 + r^2\left[\frac{1}{u_y^2} - 1\right]\right), \tag{23}$$

which can be converted back to a proportion:[13]

$$p^* = .5 \times (1 - \sqrt{1 - 4 \times var_{g_c}}) \tag{24}$$

However, if var_{g_c} is greater than .25 (i.e., the maximum variance of a binary variable), then convert it using

$$p^* = .5 \times (1 - \sqrt{1 - 4 \times [.5 - var_{g_c}]}) \qquad (25)$$

Choosing a reference sample

Appropriate reference samples can be chosen in several ways. If there is a clear population about which meta-analysts wish to make inferences (e.g., if the goal is to estimate the relationship between attitudes and grades among the university's entire student body, rather than only scholarship recipients; DesJardins, McCall, Ott, & Kim, 2010) and local standard deviation or reliability estimates for this population are available, these can be used directly. Reference-group population values can also be drawn from external sources (e.g., standard deviations from test manuals' norms or national statistical databases). Critically, researchers must ensure that these external reference samples actually represent the population of interest (e.g., a standard deviation from test norms for the general population of American adults may not accurately estimate the standard deviation for a target population of South African job applicants).

In a meta-analysis, even if there is not a clear reference population to which the meta-analyst wishes to generalize results, differences in an independent or dependent variable's variability across samples will produce artifactual between-studies heterogeneity. In such cases, this range variation should be accounted for by adjusting each effect size to reflect a common pooled or total-sample standard deviation across the included samples that have the same measure (see Dahlke & Wiernik, 2019a, for details). Corrections based on this reference standard deviation remove artifactual effect-size heterogeneity without changing the mean effect size.

When should you correct for selection effects?

Selection effects are pervasive in psychological research (Rohrer, 2018). Studies in all subfields can be affected by nonrepresentative sampling, attrition, nonresponse, and other selection processes. Failure to consider and correct for selection effects can lead to highly inaccurate substantive conclusions. For example, many studies have shown a negative correlation between cognitive ability and conscientiousness, suggesting support for an intelligence-compensation hypothesis (i.e., that conscientiousness develops in part as a mechanism to overcome limitations of ability). Murray et al. (2014) showed that these findings were the result of selection effects in studies conducted among university students; students are admitted to universities primarily on the basis of high-school grade point average. A high grade point average, and the knowledge it reflects, can be attained through some combination of effective learning (high ability) and hard work (conscientiousness). By selecting on an outcome of these two variables, the university admissions process induces an artifactual negative correlation between ability and conscientiousness (this is the classic conditioning-on-a-collider effect; Rohrer, 2018). Similarly, studies relating admissions tests to success among graduate students underestimate the predictive validity of these tests because applicants with very low test scores are rarely admitted to graduate programs (Kuncel, Wee, Serafin, & Hezlett, 2010). These examples illustrate that it is critical to consider potential selection effects and apply appropriate corrections in order to draw accurate theoretical inferences and make sound data-based decisions.

The formulas used to correct for selection effects assume that (a) variables are linearly related and (b) conditional residual variances are equal in the selected sample and the target population (Sackett & Yang, 2000).[14] The linearity assumption is absolutely essential, as correlations quantify linear relations, and corrections for selection effects therefore cannot extrapolate information about nonlinear associations. The equal-residual-variances assumption is satisfied when one's chosen correction formula reflects the direct or indirect nature of the selection effects and includes all of the selection variables that gave rise to the selection effects (or, in the case of the UVIRR correction, when the formula includes a correction variable that fully mediates the effect of the actual selection variable; see Hunter et al., 2006). Incomplete satisfaction of this assumption (e.g., using only a subset of the selection variables to make a correction) is suboptimal, but as long as the right correction formula is used, the corrected statistic will still provide a better estimate of the effect size in the target population than will the observed effect size (e.g., Beatty, Barratt, Berry, & Sackett, 2014; cf. Rohrer, 2018).

In both primary studies and meta-analyses, it is valuable to consider potential selection processes that may be at play and to apply a correction model to account for the most likely or most reasonable processes (e.g., Mendoza, Bard, Mumford, & Ang, 2004). Doing so can yield less-biased estimates of the true relationships between constructs and can also be regarded as a sensitivity analysis gauging the potential impact of selection effects on study results. It is even possible to correct for multiple potential selection processes separately and compare the results to consider potential impacts of different selection mechanisms.

As is the case with corrections for measurement error variance, statistical adjustments for selection effects come with the cost of increased sampling error in the corrected effect sizes. Confidence intervals and standard errors must be adjusted to account for this increased uncertainty. It is always better to improve the sampling design before conducting a study than to rely on statistical adjustments during data analysis.

Correcting for Artifacts in Meta-Analysis

Both measurement error variance and selection effects should be considered and corrected for in meta-analyses. Artifact corrections can be approached in two ways. First, each effect size can be individually corrected for artifacts before the meta-analysis is conducted. Second, uncorrected effect sizes can be meta-analyzed, simply weighted by effective sample size or inverse sampling error variance, and the results of this "barebones" meta-analysis (Schmidt & Hunter, 2015) can be corrected post hoc using the mean and variance of the distribution for each artifact. In this section, we describe these approaches, as well as their advantages and disadvantages.

Individual-corrections meta-analysis

In the individual-corrections approach, each effect size is corrected using artifact values based on the same sample that provided the effect size. For each effect size, the artifacts of concern should be identified, and then the relevant correction formulas should be applied. Table 3 presents a variety of correction formulas for correlations, *d* values, and their standard errors that can be used to correct for measurement error and/or various types of selection effects. For example, if you want to correct an experimental study's *d* value for measurement error variance in the dependent variable, measurement error variance in the treatment, and indirect range restriction for the dependent variable, apply the *d*-value procedure for UVIRR in Table 3 using the d_{obs} value and the sample values for r_{gM}, $\sqrt{r_{yy'}}$, and u_y. In individual-corrections meta-analysis, each effect size can potentially be corrected for a different set of artifacts (e.g., if range restriction is a concern in some studies but not others).[15]

Each effect size's sampling error must also be adjusted to account for increased uncertainty stemming from the corrections. For correlations, the sampling-error adjustment for most corrections takes the following form:

$$SE^2_{r_c} = SE^2_{r_{\text{obs}}} \times (r_c / r_{\text{obs}})^2 \tag{26}$$

That is, the sampling error variance is adjusted by the square of the adjustment applied to the effect size. For bivariate indirect selection, the adjustment to sampling error variance is more complex, to account for the unique additive term in this correction (see Dahlke & Wiernik, 2019a, for details).

When *d* values are corrected only for measurement error in the dependent variable and $\sqrt{r_{yy'_{pooled}}}$ is used, the sampling-error adjustment follows the same process as Equation 26:

$$SE^2_{d_c} = SE^2_{d_{\text{obs}}} \times (d_c / d_{\text{obs}})^2 \tag{27}$$

For corrections involving converting *d* to *r* and back, the sampling error is adjusted by converting $SE^2_{d_{\text{obs}}}$ to $SE^2_{r_{\text{obs}}}$, applying Equation 26, and then converting $SE^2_{r_c}$ to $SE^2_{d_c}$:

$$SE^2_{d_c} = \frac{SE^2_{d_{\text{obs}}}\left(\dfrac{r_c}{r_{\text{obs}}}\right)^2}{(1 + d^2_{\text{obs}} p[1-p])^2 \left(d^2_{\text{obs}} + \dfrac{1}{p(1-p)}\right) p^*(1-p^*)(1-r_c^2)^3}, \tag{28}$$

where p^* is the true group proportion in the reference population.[16]

If meta-analyses are conducted using inverse-sampling-error-variance weights (or weights based on these, such as Hedges-Vevea random-effects weights; Hedges & Vevea, 1998), the adjusted sampling error variances for the corrected effect sizes are used in the weights in place of the original sampling error variances. The meta-analytic computations for the mean effect size, random-effects variance component (in Hedges-Vevea notation, τ^2_c; in Hunter-Schmidt notation, SD^2_ρ or SD^2_δ),[17] and other results then proceed normally.

A potential problem with weighting effect sizes by inverse sampling error variance estimated in each sample is that the sampling-error-variance formulas for correlations and *d* values include the population effect sizes. As a result, weighting effect sizes by inverse sampling error variance introduces dependency between study weights and study effect sizes, which can bias meta-analysis results (Hedges & Olkin, 1985, p. 110; Kulinskaya & Bakbergenuly, 2018; Schmidt & Hunter, 2015; Shuster, 2010). An alternative method that avoids this bias is to weight by sample size (or effective sample size in the case of *d* values with unequal group sizes; Kulinskaya & Bakbergenuly, 2018). If sample-size weights are used, then adjusted weights are computed as follows[18]:

$$w^*_i = N_i \times (r_{\text{obs}i} / r_{ci})^2$$

$$w^*_i = N_i \times (d_{\text{obs}i} / d_{ci})^2 \tag{29}$$

In the Hunter-Schmidt meta-analytic procedure, these adjusted weights are then used to compute the mean effect size, the observed and expected effect-size variances, and the random-effects variance component (in the Hunter-Schmidt notation, SD_ρ^2 or SD_δ^2; in the Hedges-Vevea notation, τ_c^2):

$$\bar{\rho} = \sum w_i^* r_{c_i} / \sum w_i^*$$

$$\bar{\delta} = \sum w_i^* d_{c_i} / \sum w_i^* \qquad (30)$$

$$SD_{r_c}^2 = \sum w_i^* (r_{c_i} - \bar{\rho})^2 / \sum w_i^*$$

$$SD_{d_c}^2 = \sum w_i^* (d_{c_i} - \bar{\delta})^2 / \sum w_i^* \qquad (31)$$

$$\overline{SE_{r_c}^2} = \sum w_i^* SE_{r_{ci}}^2 / \sum w_i^*$$

$$\overline{SE_{d_c}^2} = \sum w_i^* SE_{d_{ci}}^2 / \sum w_i^* \qquad (32)$$

$$\tau_c^2 = SD_\rho^2 = SD_{r_c}^2 - \overline{SE_{r_c}^2}$$

$$\tau_c^2 = SD_\delta^2 = SD_{d_c}^2 - \overline{SE_{d_c}^2} \qquad (33)$$

Meta-analyses of corrected *d* values can also be computed using the r_c value for each sample. The final meta-analytic results are converted back to the δ metric using one of these formulas:

$$\bar{\delta} = \bar{\rho} / \sqrt{\overline{p^*}\,(1 - \overline{p^*})\,(1 - \overline{\rho^2})} \qquad (34)$$

$$\tau_c^2 = SD_\delta^2 = SD_\rho^2 / [\overline{p^*}(1 - \overline{p^*})(1 - \bar{\rho}^2)^3] \qquad (35)$$

Corrected effect sizes, sampling error variances, and weights are also used in other meta-analytic procedures, such as meta-regression and subgroup moderator analyses, publication-bias analyses, and sensitivity analyses. Using corrected effect sizes for these analyses is critical to prevent heterogeneity in artifacts across studies from biasing results.

It is often the case that the necessary statistics to correct for measurement error variance or selection effects (i.e., reliability coefficients, standard deviations, and reference-population standard deviations) are not reported for all the studies in a meta-analysis. Several options are available to address missing artifact data. A simple option is to impute missing artifacts by bootstrapping (i.e., sampling with replacement) artifact values from the studies for which this information is not reported. This approach works well and generally yields accurate estimates of the mean effect size, its standard error, and the random-effects variance component (Schmidt & Hunter, 2015). A more robust approach is to apply reliability or selection generalization (Vacha-Haase, 1998), wherein meta-regression is used to predict missing reliability values or selection *u* ratios on the basis of characteristics of the sample and study design, the measure used, scale properties, and other moderators. This approach can yield somewhat more accurate results than the naïve bootstrapping approach if these factors have strong impacts on observed artifact values. A third approach is to sidestep the issue of missing data by not correcting effect sizes individually, but instead to correct overall results of the meta-analysis on the basis of the artifact means and standard deviations. This approach is called the artifact-distribution method (Schmidt & Hunter, 2015) and is described next.[19]

Artifact-distribution meta-analysis

In the artifact-distribution method, studies are initially meta-analyzed using observed effect sizes, without correcting for measurement error variance or selection effects. Then, the initial meta-analysis results are corrected using the means and variances of the distributions of artifacts that are observed. The mean effect size is corrected using the means of the square-root reliabilities ($\overline{\sqrt{r_{xx'}}}$ and $\overline{\sqrt{r_{yy'}}}$), the selection *SD* ratios ($\bar{u}_x$ and $\bar{u}_y$), and the appropriate equation in Table 3. The random-effects variance component is corrected using a weighted sum of the variances of each artifact; the weights are determined by a Taylor series approximation (TSA)—essentially, a function that reflects how each artifact individually affects the value of the corrected effect size. The TSA variance estimator for each artifact model in Table 3 takes the same general form:

$$\tau_c^2 = [\tau^2 - Var_{\text{art}}] / b_{\text{metric}}^2, \qquad (36)$$

where τ_c^2 (also labeled SD_ρ^2 *or* SD_δ^2) is the random-effects variance component for corrected effect sizes, τ^2 (also labeled SD_{res}^2) is the random-effects variance component for uncorrected effect sizes, Var_{art} is the variance in effect sizes attributable to artifacts, and b_{metric} is a scaling factor that converts τ_c^2 from the metric of uncorrected effect sizes to the corrected (true-score) effect-size metric. The exact form of the TSA function varies for each artifact model. For example, the TSA estimator used when correlations are corrected for measurement error alone is

$$\tau_c^2 = \left[\tau^2 - (\bar{\rho}^2 \times \bar{q}_y^2 \times var(q_x) + \bar{\rho}^2 \times \bar{q}_x^2 \times var(q_y))\right] / (\bar{q}_x^2\ \bar{q}_y^2), \qquad (37)$$

where $q_x = \sqrt{r_{xx'}}$, $q_y = \sqrt{r_{yy'}}$, and $\bar{\rho}$ is the mean corrected correlation. TSA estimators for other artifact models have been described elsewhere (Dahlke & Wiernik, 2019a; Hunter et al., 2006; Raju, Burke, Normand, & Langlois, 1991; see also the *psychmeta* package for R (Dahlke & Wiernik, 2019b).

Taylor series artifact-distribution methods are derived from the principle of maximum likelihood (Raju & Drasgow, 2003), and the accuracy of their results is comparable to that of individual-correction approaches to meta-analysis (Hunter et al., 2006; Raju et al., 1991; Schmidt & Hunter, 2015). It should be noted that these artifact-distribution methods assume that the population artifact values (u ratios and reference-population, or unselected, square-root reliabilities) are uncorrelated; this assumption is often reasonable (Raju & Drasgow, 2003; see Yuan, Morgeson, & LeBreton, 2018, 2019, and Köhler, Cortina, Kurtessis, & Gölz, 2015, for discussions of statistical and situational factors that may create correlations among artifacts).

Artifact-distribution methods have several potential advantages over individual-correction approaches to meta-analysis. First, artifact-distribution methods are easier to apply because the meta-analysis results can be corrected after computation, so that each effect size does not need to be individually corrected beforehand. The difference in the ease of application is particularly large when some studies have missing artifact data, as no imputation is required. Second, artifact-distribution methods allow meta-analyses to be corrected using artifact distributions reported in previously published meta-analyses. For example, the reports of many studies of job performance do not include appropriate interrater-reliability estimates that can be used to correct for measurement error; the artifact distributions reported by Viswesvaran et al. (1996) have been widely used in subsequent meta-analyses to correct for unreliability in this variable. Third, in some cases, artifact-distribution methods can be more accurate than individual-correction methods (Dahlke & Wiernik, 2019a).

However, the artifact-distribution method also has several disadvantages. Unlike the individual-corrections approach, it requires that the same artifact-correction model be applied to all effect sizes (e.g., it is not possible to correct for direct range restriction in one sample and indirect range restriction in another).[20] Moreover, because the artifact-distribution method does not correct effect sizes individually, it cannot correct for the impact of differential reliability or selection across studies on results of meta-regression or publication-bias analyses. Instead, these analyses are conducted with the uncorrected effect sizes. For example, if smaller-sample studies with null results were more likely to be unpublished because of low reliability or restricted sampling, rather than publication bias, PET-PEESE analyses and cumulative meta-analyses (Leimu & Koricheva, 2004; McDaniel, 2009) would not detect this.

Conclusion

Measurement error variance and selection effects are pervasive in psychological research and research in other fields, and the detrimental impacts of these artifacts on the validity of research conclusions have been widely documented (Schmidt & Hunter, 2015). The infrequency with which measurement error variance and selection effects are considered and corrected in meta-analyses in many literatures conveys the risk of substantial bias in their results. By applying corrections as described in this article, researchers can make more accurate meta-analytic inferences, and more useful recommendations for future psychological research and practice can be realized.

Appendix A: Using *psychmeta* to Correct Artifacts and Conduct Psychometric Meta-analyses in R

The *psychmeta* package (Dahlke & Wiernik, 2019b) in R can be used to correct correlations and d values (and adjust confidence intervals and standard errors) for the artifact models described in this article. It can also be used to conduct individual-correction and artifact-distribution meta-analyses.

To correct correlations, use the `correct_r` function, as in the following example:

```
correct_r(correction = "bvirr",
          rxyi = .40,
          n = 150,
          rxx = .80,
          ryy = .80,
          ux = .90,
          uy = .80)
```

The `correction` argument is used to specify which artifact model to apply (bivariate indirect selection in this example), `rxyi` is the vector of observed correlations, `n` is the sample sizes, `rxx` and `ryy` are the observed reliability values, and `ux` and `uy` are the observed selection-effect u ratios. The function returns the correlations corrected for measurement error, selection effects, or both.

To correct d values, use the `correct_d` function, as in the following example:

```
correct_d(correction = "uvirr_y",
         d = .40,
         n1 = 75,
         n2 = 75,
         rGg = .80,
         ryy = .80,
         uy = .80)
```

The correction argument specifies which artifact model to apply (univariate indirect selection in this example), d is the vector of observed *d* values, n1 and n2 are the group sample sizes, rGg is the correlation between observed and true group membership, ryy is the observed total-sample reliability values, and uy is the observed selection-effect *u* ratios for the dependent variable. The function converts *d* values to point-biserial correlations, applies the specified correction, converts the r_c values back to d_c values, and returns the *d* values corrected for measurement error, selection effects, or both.

To conduct individual-correction meta-analyses, use the ma_r or ma_d function and specify ma_method = "ic":

```
ma_results_r <-
   ma_r(ma_method = "ic",
      rxyi = rxyi,
      n = n,
      rxx = rxxi,
      ryy = ryyi,
      ux = ux,
      uy = uy,
      correct_rxx = TRUE,
      correct_ryy = TRUE,
      correct_rr_x = TRUE,
      correct_rr_y = TRUE,
      indirect_rr_x = TRUE,
      indirect_rr_y = TRUE,
      data = data_r_bvirr)

ma_results_d <-
   ma_d(ma_method = "ic",
      d = d,
      n1 = n1,
      n2 = n2,
      ryy = ryyi,
      construct_y = construct,
      data = data_d_meas_multi)
```

The rxyi or d argument is the vectors of observed effect sizes; n, n1, and n2 are the sample sizes; rxx and ryy are the observed total-sample reliability values; and ux and uy are the observed selection-effect *u* ratios. The correct_ arguments specify whether each artifact should be corrected. The indirect_rr_ arguments specify whether selection was direct or indirect for each variable. The data argument is the data frame containing the data for the meta-analysis.

These functions will correct each effect size for measurement error variance, selection effects, or both, and then conduct Hunter-Schmidt meta-analyses for each set of construct variables. Moderators can also be specified, and a variety of follow-up analyses (sensitivity analyses, meta-regression, etc.) can be conducted. It is also possible to use the get_metafor function to extract a data frame of corrected effect sizes, corrected variances, and sample and moderator information for use with the meta-analysis models provided in the *metafor* package (Viechtbauer, 2010):

```
es_data <-
   get_metafor(ma_results_d,
      analyses = list(construct_y = "Y"),
      ma_method = "ic",
      correction_type = "ts")
```

In this code, ma_results is the results from either ma_r or ma_d, analyses specifies which analysis to extract from ma_results, ma_method specifies which type of meta-analysis to extract data for (in this example, individual-correction results), and correction_type specifies whether to extract data corrected for all artifacts ("ts") or uncorrected for measurement error in one of the variables ("vgx" or "vgy"). The function returns a data frame of effect sizes, variances, sample information, and moderator variables that can be used with *metafor*.

To conduct artifact-distribution meta-analyses, use the ma_r or ma_d function and specify ma_method = "ad":

```
ma_results_r <-
   ma_r(ma_method = "ad",
      rxyi = rxyi,
      n = n,
      rxx = rxxi,
      ryy = ryyi,
      ux = ux,
      uy = uy,
      correct_rxx = TRUE,
      correct_ryy = TRUE,
      correct_rr_x = TRUE,
      correct_rr_y = TRUE,
      indirect_rr_x = TRUE,
      indirect_rr_y = TRUE,
      data = data_r_bvirr)
```

The arguments are the same as previously specified.

Additional details about these and other functions are provided in the *psychmeta* documentation and vignettes (Dahlke & Wiernik, 2019c).

Appendix B: Accuracy of Corrections for *d* Values When Groups Have Been Misclassified

This appendix illustrates the effects of group misclassification on observed *d* values in group-comparison research. We simulated *d* values, varying the total

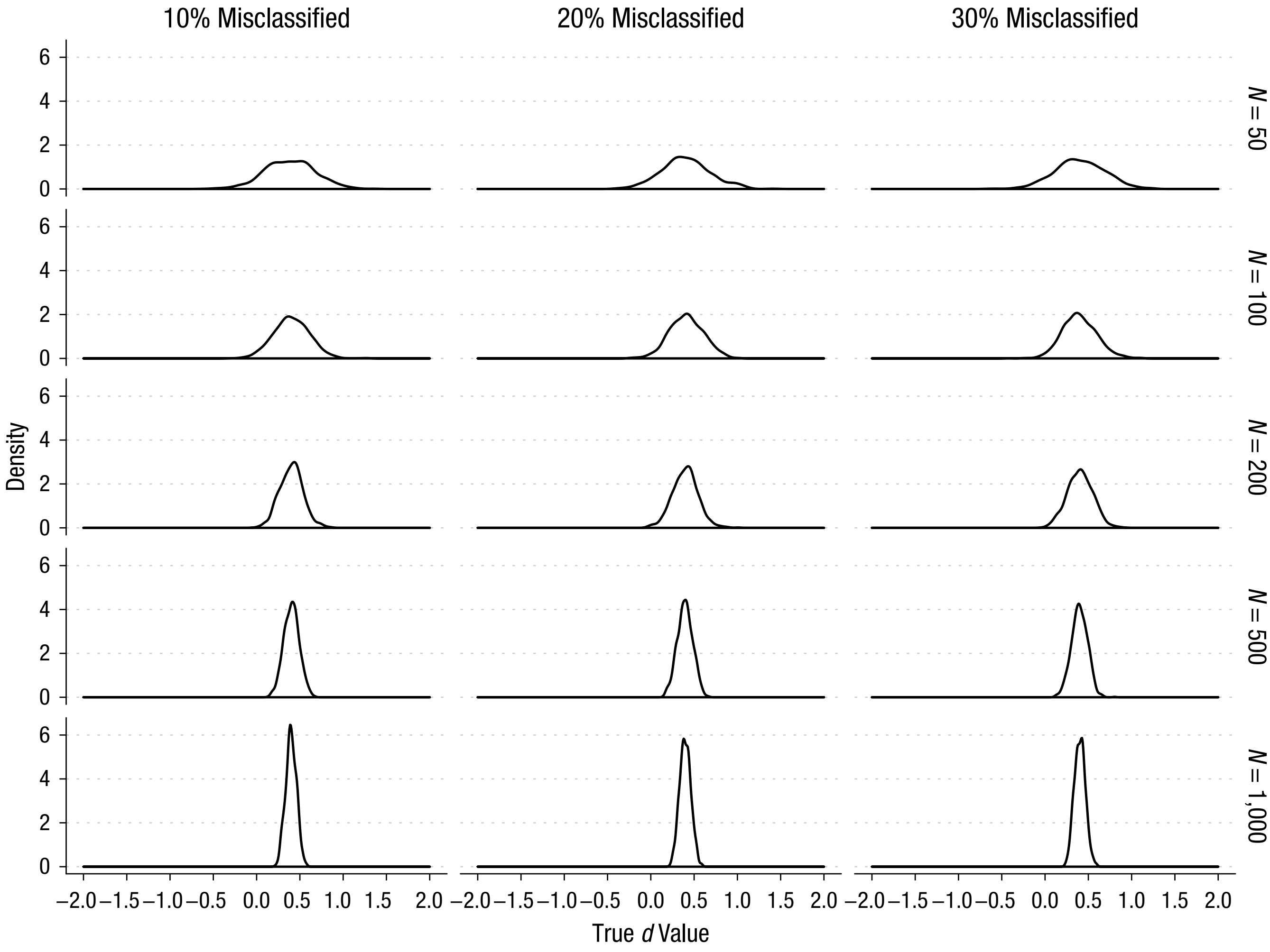

Fig. B1. Distributions of the true *d* values for the 15 conditions in the simulation.

sample size (N = 50, 100, 200, 500, or 1,000) and the misclassification rate (10%, 20%, or 30%). For each combination of sample size and misclassification rate, we simulated 1,000 samples of scores on a dependent variable y for two groups with equal sample sizes (i.e., $n_1 = n_2 = N/2$). The population distribution for Group A was specified to have a mean, $\bar{y}$, of .00 and standard deviation of 1.0. The population distribution for Group B was specified to have a mean, $\bar{y}$, of .40 and standard deviation of 1.0. Thus, the population for simulated samples was specified to be homogeneous, with δ = .40. For each simulated sample, we calculated the true d value using the standard formula: $d = (Mean_{\mathrm{A}} - Mean_{\mathrm{B}})/SD_{\mathrm{pooled}}$. The distribution of true d values for each condition is shown in Figure B1. These distributions reflect only variability introduced by sampling error variance.

To illustrate the impact of group misclassification, in each simulated sample, we created an observed-group variable by randomly reassigning a proportion of the sample equal to the misclassification rate to the opposite group, thereby ensuring equal misclassification for the two groups. We then calculated the d value for the difference between these observed groups on the dependent variable, y. The distributions of observed d values after misclassification are shown in Figure B2. These distributions reflect both variability introduced by sampling error variance and the bias and variability introduced by group misclassification.

In Figure B2, it is apparent that the mean observed d value is biased toward 0 relative to the true population value (i.e., δ = .40). Moreover, as the total sample size and misclassification rate increase, the distributions of observed d values become increasingly negatively skewed and bimodal. This is because in a minority of samples, group members with the most extreme values of y are misclassified, causing the sign of the observed d value to flip relative to the true d value.

We then corrected for group misclassification using Equations 12 and 13. We calculated r_{gG} using the correlation between the true and observed group membership for each simulated case. The distributions of the corrected d (i.e., d_c) values are shown in Figure B3.

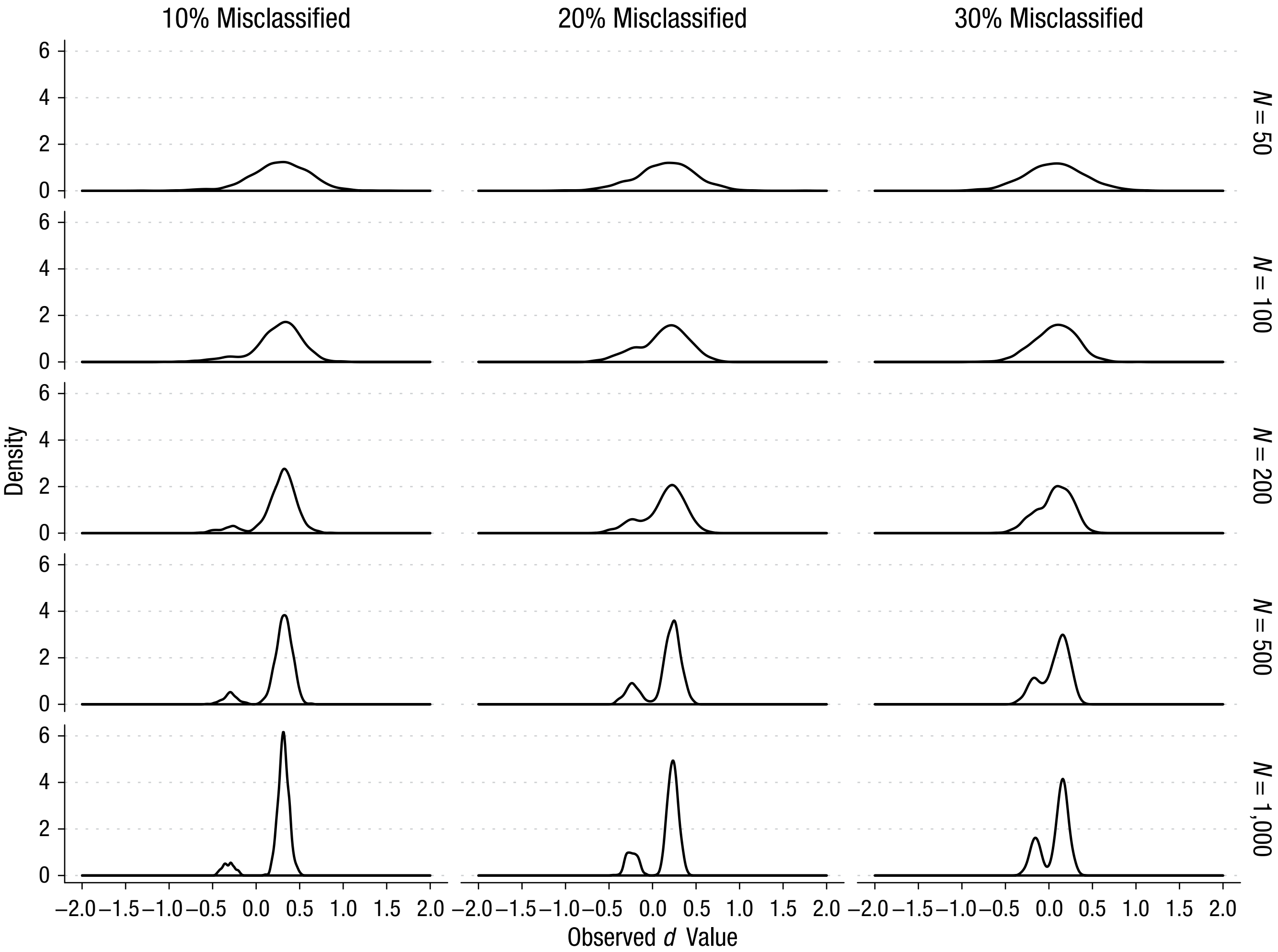

Fig. B2. Distribution of the observed *d* values for the 15 conditions in the simulation.

In Figure B3, the larger mode in the bimodal distributions is centered on the mean true *d* value, an indication that the correction model was accurate. However, the d_c values retain the negative skew and bimodal distribution of the observed *d* values because the correction cannot adjust for potential sign flipping in observed *d* values.

Summary statistics for the distributions of true *d* values, observed *d* values, and corrected *d* values (d_c values) are shown in Table B1. Because of the negative skew and bimodal distribution of the observed *d* values when groups are misclassified, the mean d_c value is negatively biased as an estimator of the mean true *d* value. Similarly, the standard deviation of the d_c values is positively biased as an estimator of the standard deviation of the true *d* values. These biases can be corrected by instead calculating the median d_c value to estimate the mean true *d* value and by calculating the median absolute deviation from the median d_c value (*MAD*) to estimate the standard deviation of the true *d* values. These estimators are much more robust to the presence of observed *d* values that are outliers because of sign flipping (cf. Lin et al., 2017). When group misclassification is a potential concern (even if corrections using r_{gG} are not made), we recommend conducting meta-analyses using the median and *MAD*, rather than mean and standard deviation, of observed or corrected *d* values.

Transparency

Action Editor: Frederick L. Oswald
Editor: Daniel J. Simons
Author Contributions
B. M. Wiernik outlined the manuscript, wrote the first draft, and prepared Figures 1, 3, and 5. J. A. Dahlke critically revised and edited the manuscript and prepared Figures 2, 4, and 6 and the Appendix B figures.
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

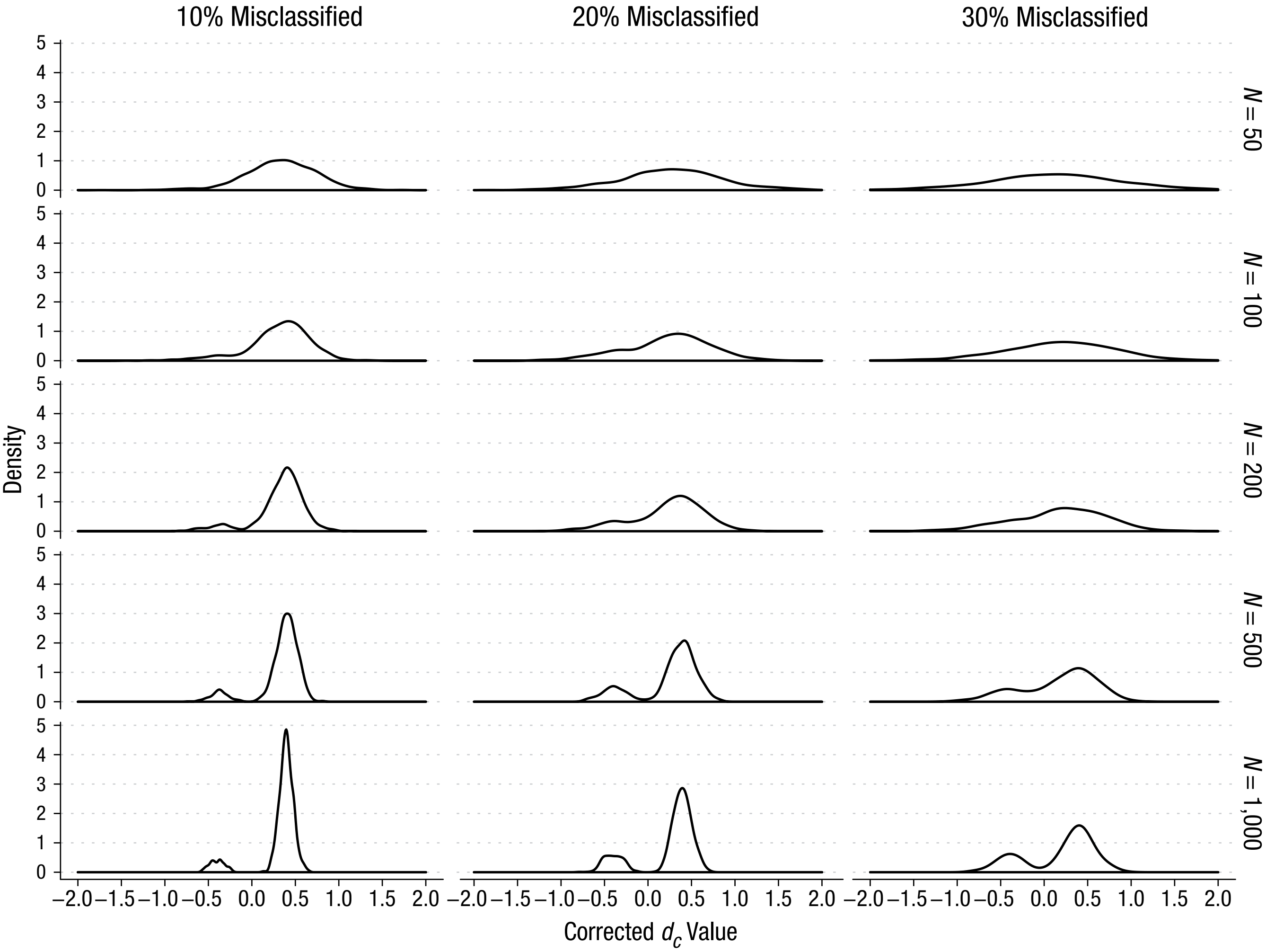

Fig. B3. Distributions of the corrected d values (i.e., d_c values) for the 15 conditions in the simulation.

Open Practices

Open Data: not applicable

Open Materials: https://osf.io/cp6rt/

Preregistration: not applicable

All materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/cp6rt/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245919885611. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Brenton M. Wiernik https://orcid.org/0000-0001-9560-6336

Jeffrey A. Dahlke https://orcid.org/0000-0003-1024-4562

Notes

1. In classical test theory, an individual's *true score* is the expected value of the individual's response to a measurement process that has been repeated an infinite number of times, such that all random errors of measurement average out to zero across observations. That is, the true score is the score on the measure without any measurement error variance. The true score does not necessarily correspond to the individual's standing on the intended latent construct—that is a question of measure validity, not reliability. If a measure has poor validity as an indicator of the intended construct, even correlations corrected for measurement error variance will poorly reflect correlations with the latent-construct variable (for discussions, see Borsboom, 2006; Schmidt, Le, & Oh, 2013).

2. Measurement error can include both systematic and random components. Systematic error (also called *bias*) affects each score in the same manner (e.g., consistent underestimates) and generally refers to the mean error across persons. Random error affects each score differently and refers to the variance of the errors across persons. Typically, only the error variance, not the mean error, affects standardized effect sizes such as correlations and Cohen's *d*. However, if a measure is differentially biased across groups or score ranges, systematic error can also affect correlations or *d* values. The methods described in this article assume that errors are uncorrelated with true scores on the included

Table B1. Summary Statistics for the Distributions of Simulated True *d* Values, Observed *d* Values, and Corrected *d* Values

	True values				Observed values				Corrected values			
N	Mean *d*	Median *d*	SD_d	MAD_d	Mean *d*	Median *d*	SD_d	MAD_d	Mean d_c	Median d_c	SD_{d_c}	MAD_{d_c}
					10% misclassification							
50	0.39	0.38	0.29	0.29	0.28	0.29	0.34	0.32	0.33	0.35	0.41	0.39
100	0.40	0.40	0.28	0.26	0.25	0.29	0.34	0.32	0.31	0.36	0.63	0.54
200	0.40	0.40	0.29	0.29	0.26	0.31	0.34	0.33	0.33	0.38	1.10	0.77
500	0.40	0.40	0.21	0.21	0.26	0.31	0.29	0.23	0.33	0.39	0.37	0.29
1,000	0.40	0.40	0.20	0.20	0.25	0.31	0.28	0.26	0.32	0.39	0.49	0.44
					20% misclassification							
50	0.40	0.39	0.20	0.20	0.15	0.16	0.25	0.25	0.27	0.27	0.74	0.63
100	0.41	0.41	0.14	0.13	0.13	0.17	0.23	0.15	0.23	0.29	0.29	0.19
200	0.40	0.40	0.15	0.14	0.14	0.19	0.24	0.20	0.24	0.31	0.41	0.35
500	0.40	0.40	0.15	0.15	0.15	0.21	0.20	0.20	0.25	0.35	0.54	0.52
1,000	0.40	0.40	0.09	0.09	0.14	0.21	0.20	0.10	0.24	0.36	0.26	0.13
					30% misclassification							
50	0.40	0.40	0.09	0.09	0.07	0.07	0.20	0.12	0.19	0.15	0.35	0.20
100	0.40	0.39	0.09	0.09	0.07	0.08	0.17	0.15	0.20	0.21	0.43	0.38
200	0.40	0.41	0.06	0.06	0.07	0.09	0.20	0.07	0.17	0.22	0.25	0.09
500	0.40	0.40	0.06	0.06	0.07	0.11	0.20	0.08	0.17	0.28	0.33	0.15
1,000	0.40	0.40	0.06	0.06	0.07	0.13	0.16	0.10	0.19	0.33	0.40	0.26

Note: For each condition, distributions are based on 1,000 simulated samples with a homogeneous true population δ value. δ = .40; group sample size (*n*) = *N*/2; *MAD* = median absolute deviation from the median.

variables, and they will not correct for such systematic errors (see the section When Should You Correct for Measurement Error Variance? for more discussion of correction assumptions).

3. The expected effect of measurement error is to bias standardized effect sizes toward the null. However, this is only the *expected* (asymptotic) effect. When sample sizes are small, measurement error can sometime have smaller- or larger-than-expected effects on observed effect sizes, or can even make observed effect sizes larger, because sampling error can produce nuisance correlations between true scores and measurement errors (Stanley & Spence, 2014; Wacholder, Hartge, Lubin, & Dosemeci, 1995). In a meta-analysis with a sufficiently large number of studies, these random deviations from the expected null-bias effect will tend to average out to zero (Stanley & Spence, 2014; Wiernik & Ones, 2017), though low reliability can exacerbate upward-biasing effects of publication bias and questionable research practices in small samples (Loken & Gelman, 2017; Simmons, Nelson, & Simonsohn, 2011).

4. R code implementing all the methods described in this article is provided in Appendix A.

5. All formulas in this article apply equally to Hedges's *g*. For Glass's Δ, $r_{yy'}$, and u_y (in corrections for selection effects, discussed later) should be computed within the control group only.

6. In a minority of cases, if group members with the most extreme values are misclassified, misclassification can also have the effect of reversing the direction of the effect (the sign of the effect size). This effect is illustrated in Appendix B. The frequency with which such sign changes occur increases as the proportion of cases misclassified grows. In individual studies, it is not possible to determine if such a sign reversal has occurred, but in meta-analyses, this possibility can be addressed by cumulating studies using the weighted median *d* value (rather than the mean) and the weighted median absolute deviation from the median *d* value (rather than the standard deviation of the *d* values; cf. Lin, Chu, & Hodges, 2017).

7. Simulation results showing the accuracy of this procedure are reported in Appendix B.

8. Schmidt and Hunter (2015, p. 265) described an alternative method for estimating treatment reliability by assuming that a treatment is equally effective for all participants (full homogeneity of experimental effects). This assumption is very strong and rarely justifiable, so we do not recommend this method.

9. Conversely, some short measures contain highly redundant items that essentially repeat the same content multiple times (e.g., the three-item job-satisfaction measure used by Judge, Boudreau, & Bretz, 1994). In these cases, internal consistency will overestimate the reliability of the scale as a measure of a broader construct. Parallel-forms reliability or test-retest reliability would be more appropriate.

10. Typically, in psychological measurement, researchers also assume that random measurement errors are normally distributed. However, the correction methods described in this article are robust to reasonable deviations from normality of the measurement-error distributions. See also Schennach (2016) for correction methods that do not rely on normality assumptions.

11. Anecdotally, in the first author's experience discussing statistical and methodological issues with other social-media researchers, range variation and selection effects are the topics that are most frequently misunderstood.

12. We focus on the Pearson family of selection-effect corrections because these require only information about variances

and covariances among variables, so they are readily usable in meta-analysis. If individual-level data are available, depending on what is known about the nature of the selection mechanism, a variety of regression- and missing-data-based selection models can also be applied to estimate relationships in the unselected population (see, e.g., Fife, Hunter, & Mendoza, 2016; Gross & McGanney, 1987; Heckman, 1976, 1979; Olson & Becker, 1983; Pfaffel, Kollmayer, Schober, & Spiel, 2016; Puhani, 2000; Yang, Sackett, & Nho, 2004). Sackett and Yang (2000) provided a detailed overview of many of these selection procedures, their data requirements and assumptions, and their relative advantages and disadvantages.

13. For observational research, p^* can instead be estimated using the group proportion in the reference population (e.g., the gender distribution in the population, base rates of disorders in epidemiological studies, the proportion of majority- and minority-group members in the job-applicant pool; Bobko, Roth, & Bobko, 2001; Li, 2015). For observational group comparisons, the difference between the observed group proportion, p, and the population proportion, p^*, reflects an adverse-impact effect: Restriction on the selection variable causes the group proportions in the selected sample to be unrepresentative of the reference population.

14. Commonly used methods for constructing confidence intervals for correlations and d values corrected for selection effects additionally assume multivariate normality. These parametric confidence intervals may be inaccurate if this assumption is seriously violated (cf. Held & Foley, 1994).

15. To correct for measurement error in one variable but not the other, set the reliability value for the uncorrected variable to 1.0. To correct for selection effects but not measurement error variance, set the reliability values for both variables to 1.0.

16. If you are correcting only for group misclassification (and potentially measurement error variance in the dependent variable), in place of p^*, use p^{true} (an estimate of the true group proportion) if that value is known or use the observed group proportion, p, if p^{true} is unknown.

17. We use the notation τ_c^2, SD_ρ^2, or SD_δ^2 to indicate the estimated random-effects variance component for a corrected effect size and the notation τ^2 or SD_{res}^2 to indicate the estimated random-effects variance component for an observed, uncorrected effect size.

18. A different set of weights is used to correct for BVIRR. See Dahlke and Wiernik (2019a) for details.

19. Another popular approach is to impute a missing artifact value using the mean artifact value from included studies. Although this approach will not bias the mean effect size, if the amount of missing artifact data is large, it can substantially reduce artifact variability and upwardly bias the random-effects variance component.

20. In artifact-distribution meta-analysis, if a specific artifact is a concern in some samples but not others, this can be accommodated by including a value of 1.0 for the artifact in the distribution for each study in which it is not a concern.

References

Aitken, A. C. (1935). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society*, *4*, 106–110. doi:10/bmqrbz

Alexander, L. K., Lopes, B., Ricchetti-Masterson, K., & Yeatts, K. B. (2014a). *Information bias and misclassification* (ERIC Notebook No. 14; 2nd ed.). Retrieved from University of North Carolina at Chapel Hill website: https://sph.unc.edu/files/2015/07/nciph_ERIC14.pdf

Alexander, L. K., Lopes, B., Ricchetti-Masterson, K., & Yeatts, K. B. (2014b). *Selection bias* (ERIC Notebook No. 13; 2nd ed.). Retrieved from University of North Carolina at Chapel Hill website: https://sph.unc.edu/files/2015/07/nciph_ERIC13.pdf

Alexander, R. A. (1990). Correction formulas for correlations restricted by selection on an unmeasured variable. *Journal of Educational Measurement*, *27*, 187–189. doi:10/cjrr9c

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–455. doi:10/gdz4f4

Beatty, A. S., Barratt, C. L., Berry, C. M., & Sackett, P. R. (2014). Testing the generalizability of indirect range restriction corrections. *Journal of Applied Psychology*, *99*, 587–598. doi:10/f6bs73

Beatty, A. S., Walmsley, P. T., Sackett, P. R., Kuncel, N. R., & Koch, A. J. (2015). The reliability of college grades. *Educational Measurement: Issues and Practice*, *34*(4), 31–40. doi:10/gckfmw

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*, 478–494. doi:10/b383m5

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of d for range restriction and unreliability. *Organizational Research Methods*, *4*, 46–61. doi:10/d9p5tg

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425–440. doi:10/bpr6h3

Carter, E., Schönbrodt, F., Gervais, W. M., & Hilgard, J. (2017). Correcting for bias in psychology: A comparison of meta-analytic methods. *PsyArXiv*. doi:10/gcmdfw

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods*, *10*, 206–226. doi:10/b8gbsf

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, *10*, Article 35. doi:10/gdb9wr

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*, 1092–1122. doi:10/dqbns8

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology*, *8*, Article 1482. doi:10/gbw956

Credé, M., Harms, P. D., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology*, *102*, 874–888. doi:10/f3wr4m

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika*, *12*, 1–16. doi:10/dcvq9g

Dahlke, J. A., & Wiernik, B. M. (2019a). Not restricted to selection research: Accounting for indirect range restriction in organizational research. *Organizational Research Methods*. Advance online publication. doi: 10.1177/1094428119859398

Dahlke, J. A., & Wiernik, B. M. (2019b). *psychmeta*: An R package for psychometric meta-analysis. *Applied Psychological Measurement*, *43*, 415–416. doi:10/gfgt9t

Dahlke, J. A., & Wiernik, B. M. (2019c). psychmeta: Psychometric meta-analysis toolkit (R package Version 2.3.3) [Computer software]. Retrieved from https://cran.r-project.org/package=psychmeta

DesJardins, S. L., McCall, B. P., Ott, M., & Kim, J. (2010). A quasi-experimental investigation of how the Gates Millennium Scholars program is related to college students' time use and activities. *Educational Evaluation and Policy Analysis*, *32*, 456–475. doi:10/cqfk98

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, *86*, 335–337. doi:10/d5cxrq

Fife, D. A., Hunter, M. D., & Mendoza, J. L. (2016). Estimating unattenuated correlations with limited information about selection variables: Alternatives to Case IV. *Organizational Research Methods*, *19*, 593–615. doi:10/f84gph

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*, 370–378. doi:10/gbf8nx

Fried, E. I., & Flake, J. K. (2018). Measurement matters. *APS Observer*, *31*(3). Retrieved from https://www.psychologicalscience.org/observer/measurement-matters

Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods*, *8*, 88–101. doi:10/bxq9r4

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, *14*, 300–306. doi:10/fdpc26

Gross, A. L., & McGanney, M. L. (1987). The restriction of range problem and nonignorable selection processes. *Journal of Applied Psychology*, *72*, 604–610. https://doi.org/10/cx6p6w

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology*, *9*, Article 998. doi:10/gfv2zs

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*, 475–492. Retrieved from https://www.nber.org/chapters/c10491

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–161. doi:10/c62z76

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504. doi:10/b29mkd

Held, J. D., & Foley, P. P. (1994). Explanations for accuracy of the general multivariate formulas in correcting for range restriction. *Applied Psychological Measurement*, *18*, 355–367. doi:10/bhkm8v

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*, 594–612. doi:10/bt4t68

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*, 309–334. doi:10/ccfxnq

Judge, T. A., Boudreau, J. W., & Bretz, R. D. (1994). Job and life attitudes of male executives. *Journal of Applied Psychology*, *79*, 767–782. doi:10/bxsqrt

Köhler, T., Cortina, J. M., Kurtessis, J. N., & Gölz, M. (2015). Are we correcting correctly?: Interdependence of reliabilities in meta-analysis. *Organizational Research Methods*, *18*, 355–428. doi:10/f7f2bq

Kulinskaya, E., & Bakbergenuly, I. (2018, July). *Meta-analysis in practice: Time for a change*. Paper presented at the Society for Research Synthesis Methodology conference, Bristol, England.

Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the Graduate Record Examination for master's and doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement*, *70*, 340–352. doi:10/dxxfg9

Lawley, D. N. (1944). A note on Karl Pearson's selection formulæ. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, *62*, 28–30. doi:10/ckc2

Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods*, *12*, 165–200. doi:10/c9qbtd

Leimu, R., & Koricheva, J. (2004). Cumulative meta-analysis: A new tool for detection of temporal trends and publication bias in ecology. *Proceedings of the Royal Society B: Biological Sciences*, *271*, 1961–1966. doi:10/bv6kfh

Li, J. C.-H. (2015). Cohen's *d* corrected for Case IV range restriction: A more accurate procedure for evaluating subgroup differences in organizational research. *Personnel Psychology*, *68*, 899–927. doi:10/gfgnqc

Lin, L., Chu, H., & Hodges, J. S. (2017). Alternative measures of between-study heterogeneity in meta-analysis: Reducing the impact of outlying studies. *Biometrics*, *73*, 156–166. doi:10/f92hb2

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*, 584–585. doi:10/gckf3c

McDaniel, M. A. (2009). *Cumulative meta-analysis as a publication bias method*. Retrieved from https://www.people.vcu.edu/~mamcdani/Publications/Cumulative%20meta-analysis%20as%20a%20publication%20bias%20method%20Final%20SIOP%202009.pdf

McNamee, R. (2009). Intention to treat, per protocol, as treated and instrumental variable estimators given non-compliance and effect heterogeneity. *Statistics in Medicine*, *28*, 2639–2652. doi:10/c4g5zm

Mealli, F., & Rubin, D. B. (2002). Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. *Health Services and Outcomes Research Methodology*, *3*, 225–232. doi:10/b5f4h5

Mendoza, J. L., Bard, D. E., Mumford, M. D., & Ang, S. C. (2004). Criterion-related validity in multiple-hurdle designs: Estimation and bias. *Organizational Research Methods*, *7*, 418–441. doi:10/d6pnwc

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, *56*, 63–75. doi:10/djg7v2

Murray, A. L., Johnson, W., McGue, M., & Iacono, W. G. (2014). How are conscientiousness and cognitive ability related to one another? A re-examination of the intelligence compensation hypothesis. *Personality and Individual Differences*, *70*, 17–22. doi:10/f6gvt6

Olson, C. A., & Becker, B. E. (1983). A proposed technique for the treatment of restriction of range of selected validation. *Psychological Bulletin*, *93*, 137–148. doi:10/ckhf5k

Ones, D. S., Wiernik, B. M., Wilmot, M. P., & Kostal, J. W. (2016). Conceptual and methodological complexity of narrow trait measures in personality-outcome research: Better knowledge by partitioning variance from multiple latent traits and measurement artifacts. *European Journal of Personality*, *30*, 319–321. doi:10/bp27

Oswald, F. L., Ercan, S., McAbee, S. T., Ock, J., & Shaw, A. (2015). Imperfect corrections or correct imperfections? Psychometric corrections in meta-analysis. *Industrial and Organizational Psychology*, *8*, e1–e4. doi:10/ggdjt7

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*. Advance online publication. doi:10/ggc7gz

Pearson, K. (1903). Mathematical contributions to the theory of evolution. —XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London, Series A*, *200*(321–330), 1–66. doi:10/dgb3gs

Pfaffel, A., Kollmayer, M., Schober, B., & Spiel, C. (2016). A missing data approach to correct for direct and indirect range restrictions with a dichotomous criterion: A simulation study. *PLOS ONE*, *11*(3), Article e0152330. doi:10/f8whvv

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, *10*, 178–192. https://doi.org/10/bp6st4

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, *14*, 53–68. doi:10/fjkhwb

Putka, D. J., Hoffman, B. J., & Carter, N. T. (2014). Correcting the correction: When individual raters offer distinct but valid perspectives. *Industrial and Organizational Psychology*, *7*, 543–548. doi:10/gckf6z

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*, 959–981. doi:10/frvzn9

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology*, *76*, 432–446. doi:10/dcrgkf

Raju, N. S., & Drasgow, F. (2003). Maximum likelihood estimation in validity generalization. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 263–285). Mahwah, NJ: Erlbaum.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*, 203–212. doi:10/djzd32

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology*, *79*, 298–301. doi:10/bc6p5h

Revelle, W. (2009). Classical test theory and the measurement of reliability. In *An introduction to psychometric theory with applications in R*. Retrieved from http://www.personality-project.org/r/book/Chapter7.pdf

Rhemtulla, M., van Bork, R., & Borsboom, D. (2019). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*. Advance online publication. doi:10/gf835w10/gfzc9k

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*, 313–345. doi:10/cq42fv

Rogers, W. M., Schmitt, N., & Mullins, M. E. (2002). Correction for unreliability of multifactor measures: Comparison of alpha and parallel forms approaches. *Organizational Research Methods*, *5*, 184–199. doi:10/df3rz4

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*, 27–42. doi:10/gcvj3r

Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Psychology*, *92*, 538–544. doi:10/ch6qk2

Sackett, P. R., & Yang, H. (2000). Corrections for range restriction: An extended typology. *Journal of Applied Psychology*, *85*, 112–118. doi:10/c6npmd

Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics*, *8*, 341–377. doi:10/gfghb8

Schmidt, F. L. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science*, *5*, 233–242. doi:10/d7dq9m

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199–223. doi:10/fww5q4

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). doi:10/b6mg

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*, *8*, 206–224. doi:10/dzmk7n

Schmidt, F. L., Le, H., & Oh, I.-S. (2009). Correcting for the distorting effects of study artifacts in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The* *handbook of research synthesis and meta-analysis* (2nd ed., pp. 317–333). New York, NY: Russell Sage Foundation.

Schmidt, F. L., Le, H., & Oh, I.-S. (2013). Are true scores and construct scores the same? A critical examination of their substitutability and the implications for research results. *International Journal of Selection and Assessment*, *21*, 339–354. doi:10/gf4867

Shuster, J. J. (2010). Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine*, *29*, 1259–1265. doi:10/bq5xzb

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10/bxbw3c

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101. doi:10/c4wwz9

Stallings, W. M., & Gillmore, G. M. (1971). A note on "accuracy" and "precision." *Journal of Educational Measurement*, *8*, 127–129. doi:10/bmd6nm

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*, 305–318. doi:10/bc3q

Ten Have, T. R., Normand, S.-L. T., Marcus, S. M., Brown, C. H., Lavori, P., & Duan, N. (2008). Intent-to-treat vs. non-intent-to-treat analyses under treatment non-adherence in mental health randomized trials. *Psychiatric Annals*, *38*, 772–783. doi:10/cfszt3

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6–20. doi:10/cx3g2m

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, Article 3. doi:10/gckfpj

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574. doi:10/c8f68f

Wacholder, S., Hartge, P., Lubin, J. H., & Dosemeci, M. (1995). Non-differential misclassification and bias towards the null: A clarification. *Occupational and Environmental Medicine*, *52*, 557–558. doi:10/bs7qdv

Waller, N. G. (2008). Commingled samples: A neglected source of bias in reliability analysis. *Applied Psychological Measurement*, *32*, 211–223. doi:10/b2kr2h

Walter, S. D. (1983). Effects of interaction, confounding and observational error on attributable risk estimation. *American Journal of Epidemiology*, *117*, 598–604. doi:10/gfgrwn

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLOS ONE*, *11*(3), Article e0152719. doi:10/f8wpvb

Wherry, R. J., Sr. (2014). *Contributions to correlational analysis*. Orlando, FL: Academic Press.

Wiernik, B. M., & Ones, D. S. (2017). Correcting measurement error to build scientific knowledge. *Science E-Letters*. Retrieved from http://science.org/content/355/6325/584/tab-e-letters

Williams, R. H., & Zimmerman, D. W. (1977). The reliability of difference scores when errors are correlated. *Educational and Psychological Measurement*, *37*, 679–689. doi:10.1177/001316447703700310

Yang, H., Sackett, P. R., & Nho, Y. (2004). Developing a procedure to correct for range restriction that involves both institutional selection and applicants' rejection of job offers. *Organizational Research Methods*, *7*, 442–455. doi:10/fjg7sr

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, *44*, 180–198. doi:10/bprznw

Yuan, Z., Morgeson, F., & LeBreton, J. M. (2018). Situational moderators and criterion reliability: Independent or interdependent? And why? *Academy of Management Proceedings*, *2018*, Article 11389. doi:10/ggdjxg

Yuan, Z., Morgeson, F. P., & LeBreton, J. M. (2019). Maybe not so independent after all: The possibility, prevalence, and consequences of violating the independence assumptions in psychometric meta-analysis. *Personnel Psychology*. Advance online publication. doi:10/ggc49d

Zimmerman, D. W. (2007). Correction for attenuation with biased reliability estimates and correlated errors in populations and samples. *Educational and Psychological Measurement*, *67*, 920–939. doi:10/fjn7rp