Article

Unreliable heterogeneity: how measurement error obscures heterogeneity in meta-analyses in psychology

Anton Olsson-Collentine[1,*], Marjan Bakker[1], and Jelte Wicherts[1]

[1]Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, the Netherlands.
[*]Correspondence: anton@olssoncollentine.com

ABSTRACT Measurement error (imperfect reliability) is present in any empirical effect size estimate and systematically attenuates observed effect sizes compared to true underlying effect sizes. Yet there exist broad concerns that proper measurement tends to be neglected in much of psychological research. We examined how measurement error in primary studies affects meta-analytic heterogeneity estimates using Monte-Carlo simulations. Our results indicate that although measurement error in primary studies can both inflate and suppress heterogeneity, under most circumstances measurement error in primary studies leads to a severe underestimate of heterogeneity in meta-analysis. Our simulations showed expected heterogeneity to be underestimated by about 15% - 60% when considering a typical effect size around r = 0.2 and true heterogeneity levels that are common in the meta-analytic literature (>0.1, in Pearson's r). The underestimate primarily depends on average reliability in primary studies (higher reliability leads to a smaller underestimate), but also worsens with smaller primary study sample sizes. We observed a positive bias in heterogeneity estimates due to measurement error only under specific and arguably uncommon circumstances of (1) actual zero heterogeneity, particularly when mean effect sizes are large, or (2) combinations of very small true heterogeneity, large variance in primary study reliabilities, large mean effect sizes, and a limited number of primary studies. Severe underestimates of heterogeneity due to measurement error may affect many meta-analyses in psychology and obscure true differences between studies that could be relevant for theory, practice, and future research efforts. Research on concrete guidance to applied meta-analysts is needed, as sophisticated methods for correcting measurement unreliability such as meta-analytic structural equation modeling (MASEM) are only applicable in exceptional cases and corrections based on classical test theory come with caveats and strong assumptions.

Estimating heterogeneity of effect sizes is usually considered the main purpose of meta-analysis, alongside estimating an average effect size. Heterogeneity of effect sizes (henceforth referred to as heterogeneity) refers to an effect size's sensitivity to variability in study design features related to the type of (1) sample, (2) setting, (3) treatment variable, and (4) measurement (e.g., Cook et al., 2002). Heterogeneity is an important outcome in meta-analysis for several reasons. First, its presence signals the existence of moderators and so can be seen as an opportunity for theoretical development (Simons et al., 2017). Second, heterogeneity affects the implementation of research by indicating that an intervention may not be equally effective under all conditions or for everyone. Third, the presence of heterogeneity changes the interpretation of the average effect size derived in a meta-analysis from the true population effect size (under homogeneity) to the average of true subpopulation effect sizes (under heterogeneity). As such, estimating heterogeneity with minimal bias is of central importance to meta-analysis (for a more extensive discussion on the importance of heterogeneity, see Olsson-Collentine et al., 2020; Simons et al., 2017).

Meta-analytic estimates depend on the quality of data in primary studies. However, there is concern that researchers in psychology tend to neglect proper measurement (Flake & Fried, 2020) to the extent that psychology can be said to be in a measurement crisis (Kane et al., 2021). For example, Flake et al. (2017) report that although latent constructs are widely examined in personality and social psychology, researchers barely report any of the validity evidence required to ascertain the extent to which implemented scales measure the constructs of interest. Such inattention to measurement can also bias heterogeneity estimates. We will focus here on the reliability of measurement [defined in classical

test theory as the ratio of true score variance to measurement error variance; e.g., Lord et al. (1968)], which is the only evidence on measurement quality that researchers report consistently (Flake et al., 2017), and which can be corrected for in meta-analysis (Hedges & Olkin, 1985; Schmidt & Hunter, 2015) .

Measurement error (imperfect reliability) is present in any empirical effect size estimate and systematically attenuates observed effect sizes compared to true underlying effect sizes (Spearman, 1904). This affects heterogeneity estimates, which from a statistical perspective, estimate the variance between true effect sizes. However, reliability is not necessarily accounted for in heterogeneity estimates, because 'true' underlying effect sizes generally refer to two different entities: Either effect sizes free from sampling error, or effect sizes free from both sampling error and measurement error. Whereas all meta-analytic models attempt to correct for sampling error in their estimates, correcting for measurement errors is explicitly modelled (alongside other measurement artifacts) only in the psychometric meta-analysis approach by Schmidt & Hunter (2015). This has consequences for heterogeneity estimates reported in many areas of psychology. Inattention to measurement reliability and how it affects heterogeneity estimates can lead meta-analysts to misinterpret their average effect size estimate, ignore the presence of theoretically or practically relevant moderators or search for (and discover) non-existent moderators, over-generalize outcomes, and implement research interventions inappropriately.

Meta-analyses in psychology rarely correct for unreliability in primary study measurements. Wiernik & Dahlke (2020) reported that among the 71 meta-analytic studies published in the journal Psychological Bulletin between 2016 and 2018, only 6/71 (8%) corrected for unreliability. Similarly, Schmidt (2010) reported that only 19/199 (10%) of the meta-analytic studies published in the same journal between 1978 - 2006 corrected for any measurement artifacts. The exception is the subfield of industrial-organizational studies, where corrections tend to be more common (e.g., Aguinis et al., 2011; Cortina, 2003).

One reason that few meta-analysts correct for unreliability could be that correcting for unreliability has long been controversial (for arguments spanning the last century, see Table 2 of LeBreton et al., 2014). On the one hand, correcting for unreliability is seen by some as conceptually problematic because it inflates effect size estimates to match a hypothetical and unachievable scenario of perfect measurement (e.g., Seymour (1988); LeBreton et al. (2014)]. Systematically correcting for unreliability may also lead to an overestimation of effect sizes, because reliability estimates are lower bound estimates of the true reliability (i.e., a reliability estimate of 0.8 implies the reliability is between 0.8 - 1). Even worse, the most popular estimate of reliability in psychology [Cronbach's Alpha; Flake et al. (2017)], as well as improvements thereof, tend to underestimate this lower bound of reliability (Sijtsma, 2008). Moreover, from a pragmatic meta-analytic perspective, effect sizes appear to often be overestimated in psychology (e.g., Kvarven et al., 2019; Schäfer & Schwarz, 2019; Scheel et al., 2021) due to selective reporting based on the significance of outcomes (Simmons et al., 2011) and publication bias (Thornton & Lee, 2000), so correcting for unreliability may inflate meta-analytic estimates even further and lead to estimates further from their true values rather than closer. On the other hand, researchers in psychology are typically interested in latent constructs rather than observed measures. As such, neglecting measurement errors means that computed estimates do not correspond to the entity of interest. From this perspective, correcting for unreliability is desirable even if doing so is challenging (e.g., Oswald et al., 2015; Schmidt & Hunter, 2015), although treating disattenuated effect sizes as directly comparable to latent scores requires strong assumptions (Borsboom & Mellenbergh, 2002). In the end, whether correcting for unreliability or not, authors and consumers of meta-analyses need to be aware of how unreliability of measurement affects heterogeneity estimates. Schmidt & Hunter (2015) assert that failure to control for variation in unreliability across studies results in a large overestimate of heterogeneity (p. 159), but little systematic research has been done to study the effects of unreliability on estimating heterogeneity.

Bias in heterogeneity estimates due to unreliability

In classical test theory (CTT), reliability of measurement (R_{xx}) is defined as the proportion true score variance (σ^2_T) to observed score variance (σ^2_X) and is presumed to be constant for all participants in a sample $R_{xx} = \sigma^2_T/\sigma^2_X$. We use unreliability of measurement to refer to $1 - R_{xx}$. Imperfect reliability in measurements leads to attenuation of observed effect sizes. For product-moment correlations the observed correlation can be computed as $r_{xy} = \rho_{xy} \times \sqrt{R_{xx'}} \times \sqrt{R_{yy'}}$ where r_{xy} is the observed correlation between the variables X and Y, ρ_{xy} is the true correlation (with or without sampling error) and $R_{xx'}$ and $R_{yy'}$ are the measurement reliabilities for X and Y. As such, correcting for unreliability is straightforward for product-moment correlations, but can also be done for other effect size types (Wiernik & Dahlke, 2020). A corrected correlation should also have its sampling variance corrected. This is inflated by a factor equal to the square of the reliabilities, such that if $a = \sqrt{R_{xx'}} \times \sqrt{R_{yy'}}$ then the corrected sampling variance $V_\rho = V_r/a^2$

(e.g., Borenstein, 2009, p. 343). Reliability is usually assumed to be known (but see Charles, 2005; Raju et al., 1991). Imperfect reliability in primary studies can both increase and suppress true heterogeneity.

Heterogeneity is inflated in heterogeneity estimates of uncorrected effect sizes due to the variability in reliability between primary studies. This is because the attenuation in effect size will then differ from study to study, which results in differences across studies in observed effect sizes beyond their true variability. This is most easily illustrated by assuming a common true effect size μ across studies (i.e., a fixed effect and no true heterogeneity) and infinite sample size within studies.

Any observed effect size $\hat{\mu}$ can then be seen as the true effect size μ multiplied by a study's reliability R_i (assuming $R_i = R_{xx'} = R_{yy'}$). Hence, the distribution of observed effect sizes is the true effect size μ multiplied by the distribution of reliabilities across studies. That is, $\hat{\mu} \sim \mu \times N(\bar{R}, \sigma_R) = N(\mu\bar{R}, \mu\sigma_R)$, where $\bar{R}$ is the average reliability across studies and σ_R its standard deviation. Because reliability lies between 0 – 1, under these conditions imperfect reliability implies that the observed average effect size will be less than the true effect size by a factor $\bar{R}$ and estimated heterogeneity will be larger than true heterogeneity (i.e., zero) by the product of $\mu \times \sigma_R$ (if heterogeneity is expressed as standard deviation rather than variance). That is, variability in reliability across studies is a force that inflates heterogeneity depending on the size of the true effect size. Table 1 illustrates this effect for three studies with differing reliability and zero true heterogeneity.

Heterogeneity is suppressed in heterogeneity estimates of uncorrected effect sizes because in absolute terms larger effect sizes are more attenuated by a given reliability. That is, although the attenuation formula $r_{xy} = \rho_{xy} \times \sqrt{R_{xx'}} \times \sqrt{R_{yy'}}$ implies that any effect size ρ_{xy} will be attenuated by the same proportion given the same reliabilities $\sqrt{R_{xx'}} \times \sqrt{R_{yy'}}$, this proportion corresponds to a larger absolute value for larger effect sizes. Consequently, in the presence of heterogeneity attenuation will move larger true effect sizes further towards zero than smaller ones, decreasing heterogeneity in observed effect sizes. This suppression can be observed directly by assuming an average effect size θ with some heterogeneity τ^2, infinite sample size within studies (i.e., no within-study sampling variance), and equal reliability across studies ($R = R_i = R_{xx'} = R_{yy'}$). Any observed effect size $\hat{\theta}_i$ can then be seen as a study's true effect size θ_i multiplied by the study reliability R. Hence, the distribution of observed effect sizes is the distribution of true effect sizes $N(\theta, \tau^2)$ multiplied by the reliability R, such that $R \times N(\theta, \tau^2) = N(R\theta, R^2\tau^2) = N(R\theta, R\tau)$. Because reliability lies between 0 – 1, under these conditions imperfect reliability implies that both the observed average effect size and the estimated heterogeneity will be less than their true value by a factor $\bar{R}$ (if heterogeneity is expressed as standard deviation rather than variance). This depiction is not entirely correct for Pearson's r as it is bounded at {-1, 1}, but holds approximately if the mean effect size θ and heterogeneity τ^2 are not too large. Hence, imperfect reliability in studies can be expected to suppress heterogeneity estimates whenever true heterogeneity is present. Consider an average reliability of 0.8 [the median reported reliability in psychology; Flake et al. (2017)]. Then, $0.8 \times N(\theta, \tau) = N(0.8\theta, 0.8\tau)$, resulting in 20% less standard deviation among observed effect sizes than in true effect sizes. Table 2 illustrates this relationship for infinite sample size (i.e., no sampling variance) and three levels of reliability across three studies.

It is difficult to predict if the total effect is a negative or positive bias in heterogeneity due to unreliability, and existing literature tends to focus on the inflation of heterogeneity (e.g., Borenstein, 2009, p. 342; Card, 2015, p. 126; Schmidt & Hunter, 2015; Wiernik & Dahlke, 2020). However, given the value of accurate heterogeneity estimates for many research outcomes in psychology and that the suppression of heterogeneity could often be 20% or more, we consider it important for meta-analysts to have insight into the bias in heterogeneity estimates that should be expected due to unreliability in a wider range of settings. The effect of unreliability depends on average effect size, true heterogeneity, mean reliability across studies, the variability in reliabilities, and sampling variance within studies. Hence, here we perform a Monte-Carlo simulation study to explore the expected bias in heterogeneity estimates due to unreliability in primary studies. In our simulation, we consider reliabilities to be known and conforming to the strict assumptions of CTT to gain a better understanding of how measurement error biases heterogeneity estimates, while acknowledging that in meta-analytic practice, heterogeneity estimation is further challenged by selective reporting, sampling error, and estimation of reliabilities, which is particularly challenging when CTT's strict assumptions are violated.

METHODS

All code and data for this project are available at (OSF LINK). For our simulations and analyses we used R 4.0.2 (R Core Team, 2020) and we used the R package 'metafor' (Viechtbauer, 2010) to perform meta-analyses and estimate heterogeneity. For parallel computing we took advantage of the R package 'parabar' (Constantin, 2023) and for data handling 'data.table' (Dowle & Srinivasan, 2021) and 'ggplot2' (Wickham, 2016).

Table 1: Variance in reliability inflates true heterogeneity

		Observed Effect Sizes			
		Study 1	Study 2	Study 3	
Meta-Analysis	ρ_{xy}	$\sqrt{R_{xx'}} \times \sqrt{R_{yy'}} = .6$	$\sqrt{R_{xx'}} \times \sqrt{R_{yy'}} = .7$	$\sqrt{R_{xx'}} \times \sqrt{R_{yy'}} = .8$	SD(ES)
I	0.0	0.00	0.00	0.00	0.00
II	0.3	0.18	0.21	0.24	0.03
III	0.5	0.30	0.35	0.40	0.05

Note: Reproduced from Olsson Collentine et al. (2020). The values under Study 1, 2 and 3 are observed effect sizes for that study given its measurement reliability $\sqrt{R_{xx'}} \times \sqrt{R_{yy'}}$ and the true effect size ρ_{xy} when within-study sample size is infinite. SD (ES) is the standard deviation of the observed effect sizes for meta-analysis I, II and III, equivalent to heterogeneity given infinite within-study sample sizes. As true heterogeneity is absent, any SD(ES) values above zero represent bias. Code to reproduce table: LINK

Table 2: Imperfect reliability suppresses true heterogeneity

	Observed Effect Sizes			
$\sqrt{R_{xx'}} \times \sqrt{R_{yy'}}$	Study 1	Study 2	Study 3	SD(ES)
1	0.1	0.15	0.2	0.05
0.8	0.08	0.12	0.16	0.04
0.6	0.06	0.09	0.12	0.03

Note: The values under Study 1, 2 and 3 are observed effect sizes for that study given the measurement reliability $\sqrt{R_{xx'}} \times \sqrt{R_{yy'}}$ and the true effect size of that study (first row) when within-study sample size is infinite. SD (ES) is the standard deviation of the observed effect sizes, equivalent to heterogeneity given infinite within-study sample sizes. Code to reproduce table: LINK

Choice of effect size type for simulations

We focus on correlations as our effect size type in these simulations because correcting them for unreliability is straightforward and because they are the effect size type most corrected in practice. However, unreliability attenuates all types of effect sizes and all can be corrected, either directly [standardized mean differences; Wiernik & Dahlke (2020)] or by first transforming them into correlations. Although effect sizes can generally be transformed into each other (e.g., Borenstein, 2009, p. 46), this should be done with care as it can sometimes affect conclusions (e.g., we found a violation of monotonicity in heterogeneity estimates when converting effect sizes into correlations in a previous project, Olsson-Collentine et al., 2020, supplement A).

Perhaps the most common way to meta-analyze correlations is to first transform them into Fisher z correlations (Borenstein, 2009, p. 41 - 43) to normalize their distribution. The Fisher z transformation is the inverse hyperbolic tangent of the product moment correlation and can be computed as $z = 0.5 \times ln\left(\frac{1+r}{1-r}\right)$ and its variance approximated as $V_z = 1/(n-3)$. There is controversy in the use of the Fisher z transformation for meta-analysis (e.g., Schmidt & Hunter, 2015, p. 220 - 223), largely related to whether Pearson's r or Fisher's z leads to less bias in average correlations (Schulze, 2007). However, estimating heterogeneity does not seem to have been the focus of most of the literature's discussion of the appropriateness of the Fisher z transformation (Field, 2005; Hafdahl & Williams, 2009; Schmidt & Hunter, 2015). Hall & Brannick (2002) do report heterogeneity estimates but focus on the coverage of prediction intervals ('credibility intervals' in Hunter & Schmidt terminology) which combine the heterogeneity estimate and point estimate, and Brannick et al. (2019) examined results only after corrections for attenuation. Given both this inconclusive evidence and the common use of the Fisher z transformation in practice, we use it as one of the effect size types in our simulations.

We also use Pearson's r. Pearson's r has the advantage that it creates heterogeneity estimates on the same scale that are more interpretable than heterogeneity estimates on the Fisher z scale. Although Fisher z correlations can be back-transformed into Pearson's r for interpretation, its heterogeneity cannot. A workaround is to compute a prediction interval and convert this interval onto the product-moment scale (Hedges & Vevea, 1998). The width of the interval can then be used as an estimate of the heterogeneity. We prefer to compute comparable heterogeneity values for Pearson's r and Fisher's z (see the methods section 'Parameter values') and report the bias in heterogeneity estimates, as we believe this is more interpretable for applied meta-analysts.

The disadvantage to using Pearson's r when estimating heterogeneity are that (1) the effect size is bounded to {-1, 1} and (2) its sampling variance $\sigma_r^2 = \frac{(1-\rho^2)^2}{n-1}$

depends on the true effect size ρ. The bounded nature of the Pearson correlation means that large levels of heterogeneity create substantial truncation of the distribution of effect sizes if the average effect size is also large. That the sampling variance covaries with effect size leads to a small positive bias in heterogeneity estimates (supplement A). This can be solved by replacing ρ with the average correlation across studies $\bar{r}$ (Schmidt & Hunter, 2015), or using the Fisher z transformation. Fisher z correlations have the additional advantage that they are not bound to {-1, 1}. A disadvantage to using Fisher z correlations, in addition to interpretability of heterogeneity estimates, is that they require back-transforming into product-moment correlations before adding measurement error, followed by another transformation to Fisher z for meta-analysis. Especially for large Fisher z values (which are common when both average effect size and heterogeneity are large) this transformation process may introduce inaccuracies in the estimates. Given the common use of both effect size types and their mix of advantages and disadvantages we report results for both Fisher's z and Pearson's r.

Meta-analytic model

We apply two meta-analytic models to ensure that our results are not model-dependent: the Hedges & Vevea (1998) random-effects model and the Schmidt & Hunter (2015) 'bare-bones' random-effects model. The Hedges and Vevea (HV) model uses inverse variance weight computed as $w_i = 1/(V_i + T^2)$ (e.g., Borenstein et al., 2010) where V_i is the sampling variance of study i and T^2 is the estimated between-studies variance. The Hunter and Schmidt (HS) model weighs studies by sample size $w_i = N_i$. The differences between these weights are relatively minor (Borenstein et al., 2010; Schmidt & Hunter, 2015, pp. 423–426). Although the HV model is typically applied to Fisher z transformed correlations (e.g., Borenstein, 2009, pp. 41–43), we apply it to both transformed and untransformed correlations to enable a more direct comparison with the HS model and facilitate interpretation. The HS model typically includes corrections for unreliability or other measurement artifacts. However, we implement it without corrections ('bare-bones,' Schmidt & Hunter, 2015) because this allows us to focus on the bias in heterogeneity estimates rather than the performance of corrections (e.g., see Zhang, 2022 for how well they perform in terms of mean estimates).

Estimating heterogeneity in both the HV and bare-bones HS model essentially consists of comparing the observed variance in effect sizes with what would be expected from sampling error alone and then ascribing any excess observed variance to heterogeneity (Borenstein et al., 2010, Box 1; Schmidt & Hunter, 2015, p. 100). However, the process for doing so differs somewhat between methods. We apply the Restricted Maximum Likelihood (REML) estimator of heterogeneity when estimating heterogeneity in the Hedges and Vevea (1998) model, because it is the generally recommended heterogeneity estimator (Veroniki et al., 2016; Viechtbauer, 2005). Veroniki et al. (2016) describe the REML estimator in detail. The Hunter and Schmidt (2015) model estimates heterogeneity as $\tau^2 = \sigma_r^2 - \sigma_e^2$ where σ_r^2 is the observed variance across correlations and σ_e^2 is the average sampling error variance across studies, computed as $\frac{\Sigma[N_i \sigma_{ei}^2]}{\Sigma N_i}$ (Schmidt & Hunter, 2015, p. 100) and σ_{ei}^2 is the usual sampling variance for the Pearson correlation but with ρ replaced with the sample size weighted average correlation $\bar{r} = \frac{\Sigma n_i r_i}{\Sigma n_i}$ instead of r_i as is common. A generalized version of the HS estimator for effect sizes other than correlations was developed by Viechtbauer and colleagues (Viechtbauer, 2005; Viechtbauer et al., 2015).

Simulation study design

Figure 1 provides an overview of the design of this simulation study. Our design was broadly the same for product-moment correlations and Fisher's z, with some minor differences we highlight in the text below detailing the procedure.

We generated data for the meta-analyses as follows. First, we sampled $i = 1, ..., k$ true study effect sizes θ_i from $Ntrunc(\theta, \tau^2)$. That is, a normal distribution with a mean of θ and a variance of τ^2 that we truncated at $\{-1, 1\}$ to stay within the bounds of the product-moment correlation. We implemented the effect size truncation in R using inverse transform sampling (link: functions.r). For Fisher's z no truncation was applied. Throughout this text we will refer to three types of truncation: (1) effect size truncation (truncation of Pearson's correlations between {-1, 1}), (2), reliability truncation (truncation of reliability between {0, 1}), and (3) heterogeneity truncation (truncation of heterogeneity estimates to be non-negative).

For each of the k sampled true study effect sizes θ_i we then sampled one observed effect size γ_i from a truncated normal distribution $Ntrunc(\theta_i, \sigma_i^2)$, using the same truncation procedure. The sampling variance σ_i^2 for each true effect size θ_i we computed using the standard formula for Pearson's r $\sigma_i^2 = \frac{(1-\theta_i^2)^2}{N_i - 1}$ where N_i is the total sample size for study $i = 1, ..., k$. For Fisher's z no truncation was applied and we approximated the sampling variance as $\sigma_i^2 = \frac{1}{N_i - 3}$. At this point we have k sampled effect sizes γ_i in Pearson's r or Fisher's z without measurement error.

To add measurement error we first sampled a reliability R_i for each study from a truncated ($\{0, 1\}$) normal

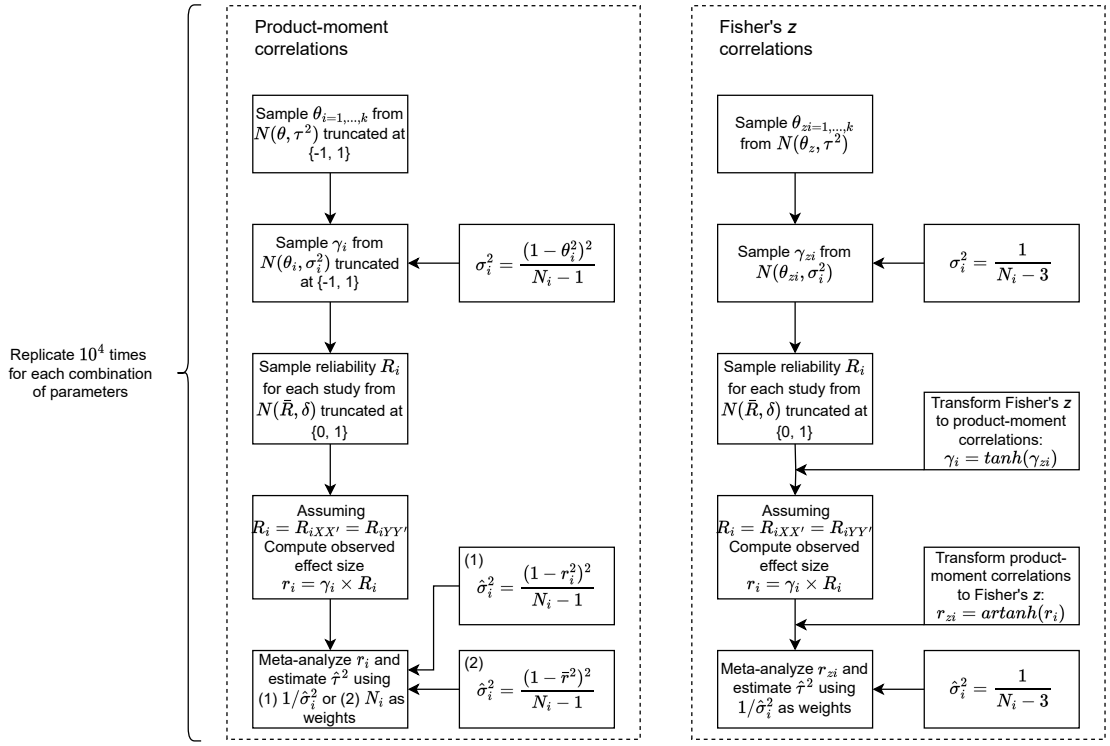

Figure 1: Figure 1. Flowchart summarizing the simulation design. θ = average true effect size (i.e., without sampling or measurement error), τ^2 = variance of true effect sizes, θ_i = study true effect size, γ_i = study effect size with sampling error σ_i^2, N_i = study sample size, R_i = study reliability, $\bar{R}$ = average reliability across studies, δ = standard deviation of reliabilities across studies, r_i = observed study effect size with sampling and measurement error, $\bar{r}$ = average observed effect size across studies. For product-moment correlations, option (1) corresponds to the Hedges and Vevea (1998) model and option (2) to the Schmidt and Hunter (2015) model (see methods section 'Meta-analytic model').

distribution $Ntrunc(\bar{R}, \delta)$, given some average reliability across studies $\bar{R}$ and standard deviation in reliability δ. We assumed that both the dependent and independent variable within a study were measured with the same reliability such that $R_i = R_{iXX'} = R_{iYY'}$. We then computed observed effect sizes r_i for each study given the attenuation formula $r_i = \gamma_i \times \sqrt{R_{XX'}} \times \sqrt{R_{YY'}}$. Because we assumed equal reliability in X and Y this simplifies to $r_i = \gamma_i \times R_i$. Fisher's z we transformed to product-moment correlations before adding measurement error and then back-transformed into Fisher's z before the next step. At this point we have k effect sizes r_i in Pearson's r or Fisher's z with sampling error and measurement error. We then applied either the HV meta-analytic model or the HS model, as described in the section 'the meta-analytic model'. The procedure described in the current section was replicated 10,000 times for each combination of parameter values.

Parameter values

We ran our analyses across a range of within-study sample sizes N and number of studies K. Within meta-analyses we used a fixed sample size across all studies such that N=N_i. We made this choice rather than treating sample size as a random variable (Brannick et al., 2019; as done by e.g., Field, 2005) to be able to observe the effect of changes in sample size on estimates, and because we are focused on the average heterogeneity estimates rather than its sampling variance. We base our range of sample sizes on empirical estimates of typical sample size in correlational research in psychology. Fraley et al. (2022) report median sample sizes for between-person studies in nine psychology journals between 2011 - 2018 which ranged between 69 - 496 depending on journal and year. Sassenberg & Ditrich (2019) find a median sample size of 110 (interquartile range, 71 - 195) across four journals and years (2009, 2011, 2016, and 2018) in social psychology, and Bakker et al. (preliminary results; https://osf.io/zsjf4) find median sample sizes between 24 - 184 for six journals in psychology for the

years 1995, 2006, 2019. Given these empirical findings, we consider the following sample size values {50, 100, 150, 200}.

When focusing on the bias of heterogeneity estimate, the number of meta-analyzed studies, K, would not matter for an unbiased estimator. However, both the heterogeneity estimator we used, restricted maximum likelihood, and most other heterogeneity estimators are truncated at zero (Viechtbauer, 2005). Lower values for K then result in a larger proportion of truncated variance estimates below zero because of increased sampling variance in τ^2. As such, values of K can affect bias, especially at lower levels of heterogeneity. We consider the following values of K: {5, 20, 40, 200}.

We vary the grand mean θ from 0 to 0.6 in steps of 0.1 to cover all realistic effect sizes and explore the boundaries of the interaction between effect size and reliability variance. For Fisher's z, we transformed these values to the z scale. For context to these values, we considered the empirical estimates of typical correlational effect sizes provided by Nuijten et al. (2020) and Schäfer & Schwarz (2019). Nuijten et al. (2020) performed a meta-meta-analysis on the fields of intelligence research and reported a median meta-analytic correlation of 0.24 across 779 predictive validity and correlational studies. This estimate is likely positively biased due to publication bias and selective reporting. Schäfer & Schwarz (2019) reported a median r of 0.16 amongst preregistered research in psychology in general, with an 'upper median' (i.e., the 83.35% quantile) of 0.41. This estimate is likely less affected by positive bias but should not be expected to be unbiased as there appears to be some positive bias also in preregistered research on average Scheel et al. (2021).

We defined the between-studies standard deviation (τ) to cover a wide range of variability in true effect sizes and empirical estimates reported in the psychological literature. Van Erp et al. (2017) provided empirical heterogeneity estimates from 747 meta-analyses in 61 articles published in the journal Psychological Bulletin between 1990 - 2013. The median (uncorrected) Pearson's r τ value in these data was 0.17 (interquartile range 0.1 - 0.24). The point of our study is that empirical heterogeneity estimates may not be accurately estimated. Nonetheless, we selected approximately the interquartile range of τ-values reported in Van Erp et al. (2017) for the product-moment correlation, both to make sure we covered the empirically reported range of values and because we considered them to represent reasonable variability on the product-moment correlation scale. The values we selected were {0, 0.1, 0.15, 0.2}. We selected the largest heterogeneity level to avoid excessive effect size truncation for our maximum average effect size of 0.6: $\theta_i \sim N(\theta = 0.6, \tau = 0.2)$ implies approximately 95% of effect sizes will be below 1. Our values are similar to those used by Brannick et al. (2019) in their simulation study on the performance of different heterogeneity estimators after corrections for unreliability (they used $\tau = 0$, 0.08, 0.13, 0.2).

Fisher's z is measured at a different scale and it is not possible to directly convert τ-values expressed in Pearson's r into corresponding τ-values on the Fisher's z scale. To compare results between Fisher's z and the product-moment correlation we defined heterogeneity on the I^2 scale. The I^2 index is a relative measure of heterogeneity describing the percentage of total variance that is due to heterogeneity (Higgins, 2003; Higgins & Thompson, 2002), and as such is measured from 0 - 100%. The I^2 index can be defined as $I^2 = \frac{\tau^2}{s^2+\tau^2}$ where $s^2 = \frac{\Sigma w_i(k-1)}{(\Sigma w_i)^2 - \Sigma w_i^2}$ and w_i is the precision $w_i = 1/\sigma^2$. This means that the I^2 value generally depends on both the number of studies K and the sample size N_i in these studies (Borenstein et al., 2017). However, because we keep N_i fixed across studies, the value of I^2 depends only on $N = N_i$ and the heterogeneity.

We thus computed the I^2 index for each sample size condition and τ-values defined in Pearson's r, and then computed corresponding τ-values on the Fisher's z scale given these I^2-values and sample sizes. Two complications were that the value for σ^2 varies with effect size for Pearson's r and that effect size varies across studies because of τ. We resolved this by setting $\theta = 0$ and computing the expected σ^2 given τ and sample size and then using this expected value to compute I^2. We used the 'law of the unconscious statistician' to compute the expected value of σ^2, which says that the expected value of a function $g(X)$ of a random variable can be expressed in terms of the probability distribution of X: $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$. In our case, $g(x)$ corresponds to the sampling variance of the product-moment correlation, $f(x) \sim N(\theta, \tau^2)$ and because the product-moment correlation is bounded at {-1, 1} we set ∞ instead to 0.999. The resulting τ values for Fisher z were only minimally different from {0, 0.1, 0.15, 0.2}. For example, for $N = 50$ the corresponding Fisher z τ-values were {0, 0.1031263, 0.1566007, 0.2123514} and as N increased heterogeneity values were more similar (see supplement B).

Finally, we defined average reliability levels and their standard deviation based on empirical estimates from the literature. Flake et al. (2017) documented 245 estimates of Cronbach's Alpha in psychology and found an average of 0.79 and a standard deviation of 0.13. The interquartile range was approximately 0.68 – 0.87 for studies using (*ad hoc*) scales that lacked a reference to an earlier source and 0.79 – 0.88 for scales that did have a reference. Sánchez-Meca et al. (2012) reported the reliability estimates based on five so-called reliability generalization studies. These five studies summarized reli-

ability in 25 – 51 primary studies (184 total), and the mean reliability ranged from 0.767 to 0.891 with standard deviations ranging between 0.034 – 0.133. Given some likely positive bias in such empirical values caused by selective reporting (Hussey et al., 2023), we examined the following mean reliabilities {0.6, 0.7, 0.8, 0.9}, and standard deviations {0, 0.05, 0.1, 0.15}.

RESULTS

Overall, the net bias in heterogeneity estimates due to imperfect reliability in primary studies was negative. We found that the inflation of heterogeneity due to variance in reliability across studies was less than the suppression due to average unreliability across all levels of heterogeneity in our primary analysis. For clarity of communication, we report here only the condition with maximum variance in reliability (SD = 0.15), as lower levels of variance will generate inflation between that observed for zero average effect size (i.e., zero inflation) and that generated under the maximum reliability variance. Consequently, analyses with smaller variances in reliabilities across primary studies entail more severe underestimates of heterogeneity for larger effect sizes than those presented here (supplement C). In addition, although a smaller number of studies in a meta-analysis (k) leads to a larger positive bias in the absence of heterogeneity, as would be expected, there was no practical difference in the bias once true heterogeneity exceeded $\tau = 0.1$ (Supplement C). Hence, we primarily report results for k = 20 in the main text, and only report results using other values for k (5, 40, 200) when considering heterogeneity below $\tau = 0.1$. Finally, for the sake of succinctness we similarly report only results for N = 150 in the main text, except when we discuss the effects of sample size on heterogeneity estimates (Figure 4).

Figure 2 shows the net bias in heterogeneity estimates for different levels of heterogeneity (measured in τ) and meta-analytic models. Each column of Figure 2 corresponds to one nominal level of actual heterogeneity ($\tau \in \{0, 0.1, 0.15, 0.2\}$) and each row to a different analysis (HV with Fisher z, HV with Pearson's r, and HS). The black solid lines indicate actual heterogeneity, which may differ from nominal heterogeneity because of effect size truncation or, in the case of Fisher z, because the comparable heterogeneity levels were slightly different (as discussed in the method section). The different dashed lines in Figure 2 show the $\hat{\tau}$ estimates for four different levels of average reliability across primary studies, and the x-axis shows the average (superpopulation) effect size. As can be seen, the bias is marginally worse for Fisher z than the other two methods (although Fisher z is on a different scale), but the results are approximately the same for all three methods. As such, in the remainder of the results, we will focus on Pearson's r (row two).

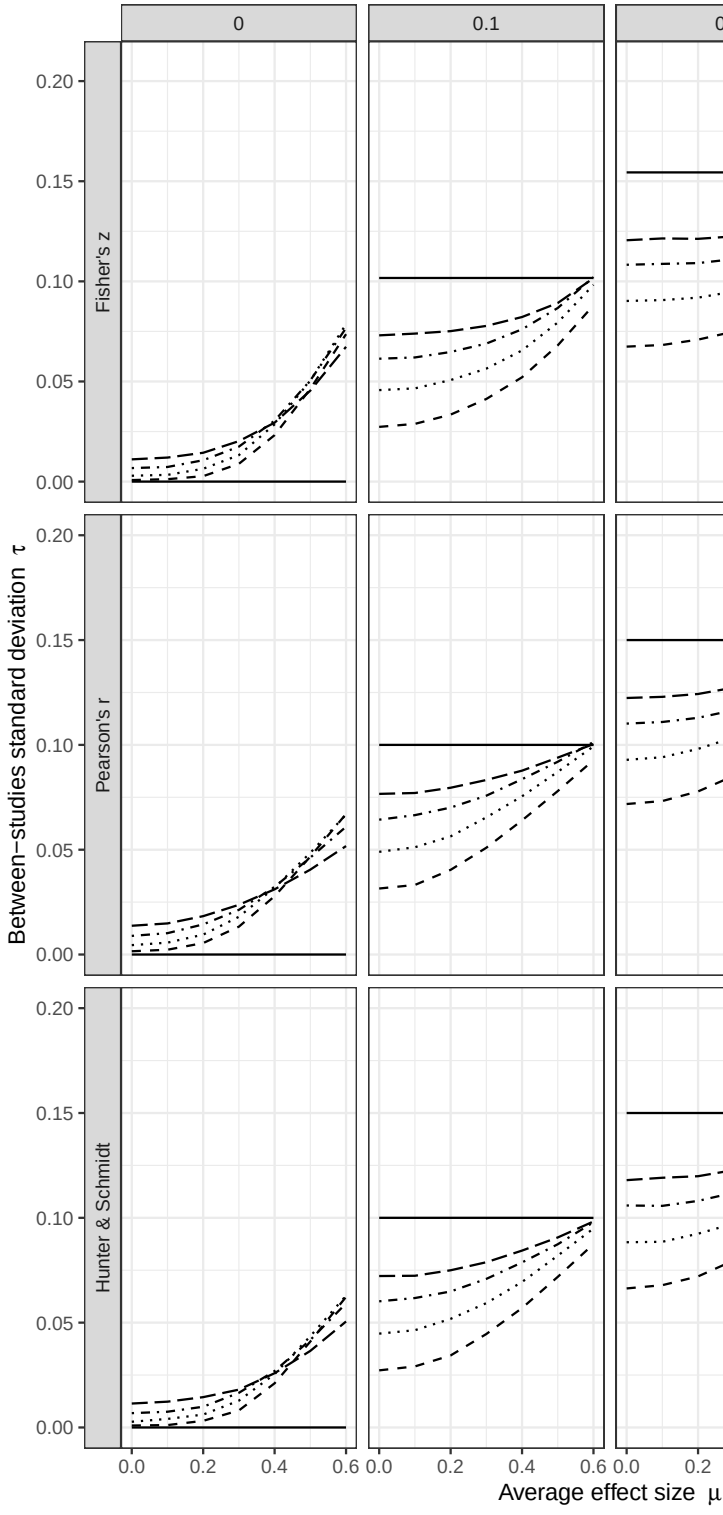

Figure 2. Unreliability in primary studies leads to a net negative bias when heterogeneity is positive. The x-axis indicates average effect size and the y-axis estimated heterogeneity in standard deviations averaged across 10,000 replications. Columns indicate the nom-

inal true heterogeneity standard deviation τ. Due to truncation in Pearson's r or translation to Fisher's z these values may differ from actual true heterogeneity standard deviation (black solid lines). Each dashed line corresponds to an average reliability in primary studies. Results are for standard deviation in reliability = 0.15, within-study sample size = 150, and 20 studies within each meta-analysis. Code to reproduce figure: LINK.

As expected, for zero heterogeneity (leftmost column Figure 2) we see an overestimate of heterogeneity that increases with average effect size. Generally, there are two sources that explain the positive bias in the absence of true heterogeneity, although the use of Pearson's r as the type of effect size additionally inflates heterogeneity estimates to a small degree for the HV model (row 2, Figure 2) due to the dependence between effect size and sampling variance (supplement A). First, and the only of these two upwardly biasing sources that applies when $\mu = 0$, the heterogeneity estimate is inflated due to the truncation of negative heterogeneity estimates (Viechtbauer, 2005). Second, there is inflation due to the variance in reliability across studies in the fixed effect scenario. For example, for $\mu = 0.2$ (the closest value to the median correlation reported by Schäfer & Schwarz, 2019) and $\bar{R} = 0.8$ the bias for Pearson's r compared to true zero heterogeneity is 0.014. For $\mu = 0.4$ (the upper median reported by Schäfer & Schwarz) and $\bar{R} = 0.8$ the bias is 0.032. For the highest average reliability (0.9) the slope is less inclined. This is because the standard deviation for this reliability level is severely truncated due to the reliability being bounded at 1. The upward force on estimations of heterogeneity with larger effect size can also be seen in the presence of heterogeneity (all other columns). However, this upward force is superseded by the suppression of heterogeneity that unreliability creates.

For all true heterogeneity levels in Figure 2 above zero, we see an overall underestimate of heterogeneity due to unreliability in the primary studies. This negative bias can be relatively severe, is worse with smaller effect sizes and with lower average reliability, and worsens in an absolute sense as true heterogeneity increases. For example, for $\bar{R} = 0.8$, $\mu = 0.2$, $\tau = 0.1$, the bias is -0.03 (30%) for Pearson's r, for $\bar{R} = 0.8$, $\mu = 0.2$, $\tau = 0.15$ the bias is -0.037 (25%), and for $\bar{R} = 0.8$, $\mu = 0.2$, $\tau = 0.2$ the bias is -0.046 (23%). For an effect size of $\mu = 0.2$, true heterogeneity between 0.1 - 0.2 is underestimated by between 16% ($\tau = 0.2$, $\bar{R} = 0.9$) and 60% ($\tau = 0.1$, $\bar{R} = 0.6$). In other words, the underestimation of true heterogeneity due to unreliability can be quite severe.

To explore further under what conditions the forces of inflation and suppression of heterogeneity reach an equilibrium, we added an analysis for several conditions with $\tau < 0.1$. We used the following values {0.02, 0.04, 0.06, 0.08}. Figure 3 presents the results of this analysis for Pearson's r. Because these lower heterogeneity levels can be expected to be affected by heterogeneity truncation to a larger extent (which results in inflation), we present results for four different numbers of studies in the meta-analysis (rows Figure 3, lower k is associated with larger inflation of estimates). The standard deviation in reliabilities is again 0.15. Figure 3 shows that even when only little heterogeneity is present, it will generally be underestimated. If < 0.3 then imperfect reliability leads to a negative bias in heterogeneity estimates for all $\tau < 0.06$ in Pearson's r. Even for the lower levels of true heterogeneity, estimates of heterogeneity will generally be positively biased only when average effect size is large, or the number of meta-analyzed studies is small.

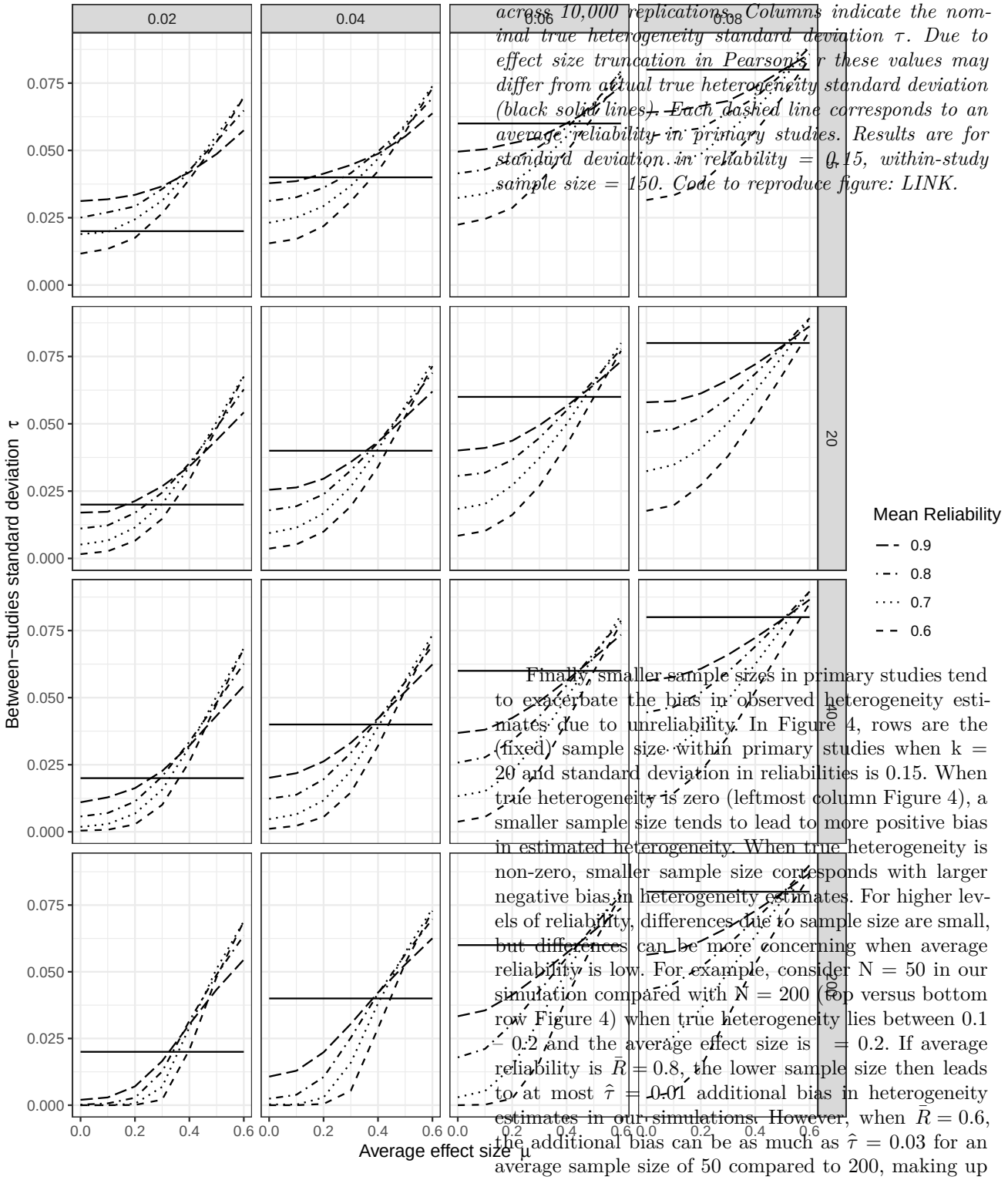

Figure 3. Unreliability in primary studies typically leads to a net negative bias even for small degrees of heterogeneity. Effect size is Pearson's r and rows indicate the number of studies in the meta-analysis. The x-axis indicates average effect size and the y-axis estimated heterogeneity in standard deviations averaged across 10,000 replications. Columns indicate the nominal true heterogeneity standard deviation τ. Due to effect size truncation in Pearson's r these values may differ from actual true heterogeneity standard deviation (black solid lines). Each dashed line corresponds to an average reliability in primary studies. Results are for standard deviation in reliability = 0.15, within-study sample size = 150. Code to reproduce figure: LINK.

Finally, smaller sample sizes in primary studies tend to exacerbate the bias in observed heterogeneity estimates due to unreliability. In Figure 4, rows are the (fixed) sample size within primary studies when k = 20 and standard deviation in reliabilities is 0.15. When true heterogeneity is zero (leftmost column Figure 4), a smaller sample size tends to lead to more positive bias in estimated heterogeneity. When true heterogeneity is non-zero, smaller sample size corresponds with larger negative bias in heterogeneity estimates. For higher levels of reliability, differences due to sample size are small, but differences can be more concerning when average reliability is low. For example, consider N = 50 in our simulation compared with N = 200 (top versus bottom row Figure 4) when true heterogeneity lies between 0.1 – 0.2 and the average effect size is = 0.2. If average reliability is $\bar{R} = 0.8$, the lower sample size then leads to at most $\hat{\tau} = 0.01$ additional bias in heterogeneity estimates in our simulations. However, when $\bar{R} = 0.6$, the additional bias can be as much as $\hat{\tau} = 0.03$ for an average sample size of 50 compared to 200, making up approximately one third of the total bias (-0.097) in this condition (N = 50, $\tau = 0.15$). Generally, for an effect size of = 0.2, true heterogeneity between 0.1 – 0.2 is underestimated by between 16% ($\tau = 0.2$, $\bar{R} = 0.9$) and 73% ($\tau = 0.1$, $\bar{R} = 0.6$) in our simulations when sample size is 50, as compared to the 16-59% we observed when sample size was 150.

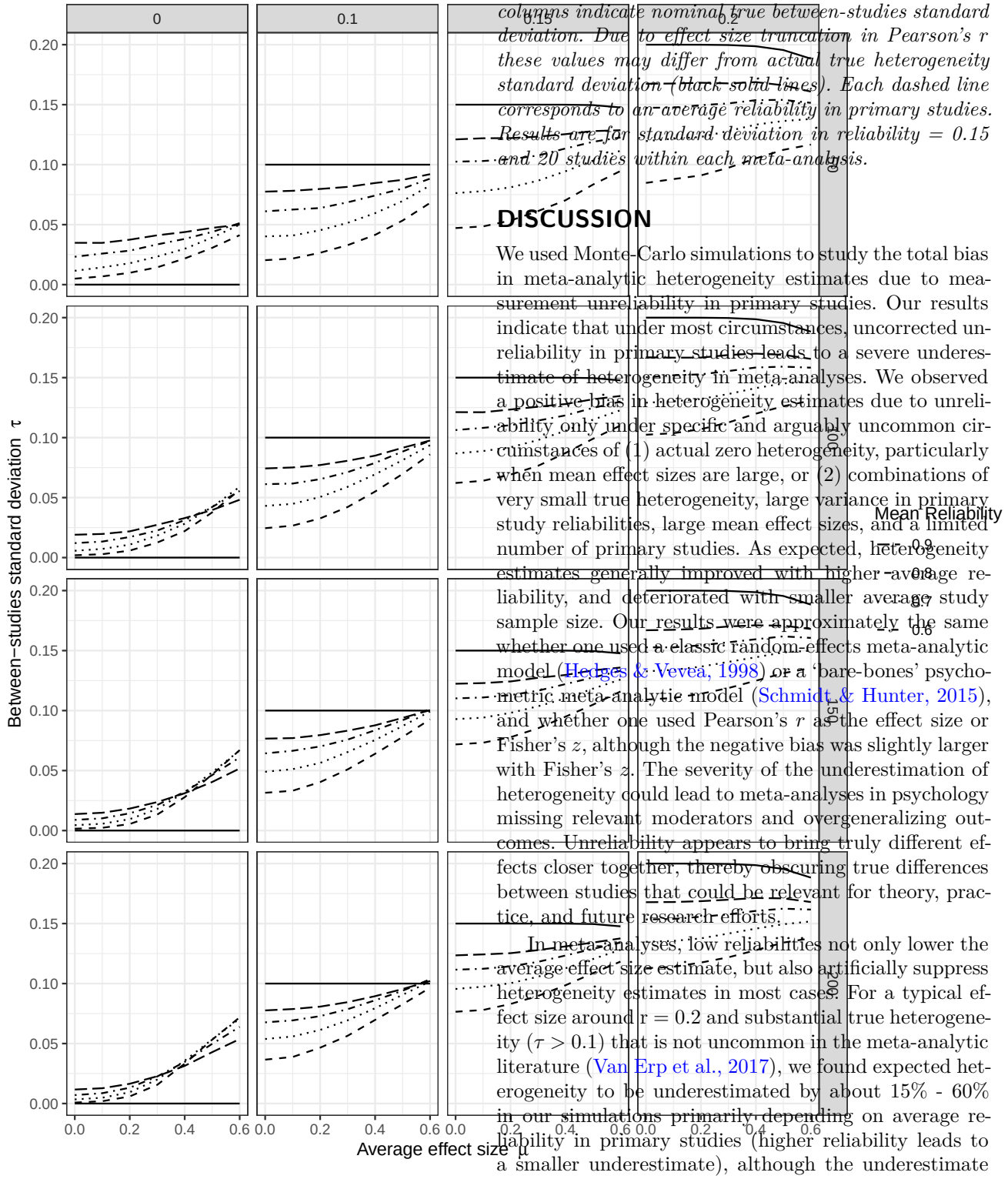

Figure 4. A smaller sample size in primary studies tends to exacerbate the bias in observed heterogeneity estimates due to unreliability. The x-axis indicates average effect size and the y-axis estimated heterogeneity in standard deviations averaged across 10,000 replications. Rows indicate the (fixed) sample size within studies and columns indicate nominal true between-studies standard deviation. Due to effect size truncation in Pearson's r these values may differ from actual true heterogeneity standard deviation (black solid lines). Each dashed line corresponds to an average reliability in primary studies. Results are for standard deviation in reliability = 0.15 and 20 studies within each meta-analysis.

DISCUSSION

We used Monte-Carlo simulations to study the total bias in meta-analytic heterogeneity estimates due to measurement unreliability in primary studies. Our results indicate that under most circumstances, uncorrected unreliability in primary studies leads to a severe underestimate of heterogeneity in meta-analyses. We observed a positive bias in heterogeneity estimates due to unreliability only under specific and arguably uncommon circumstances of (1) actual zero heterogeneity, particularly when mean effect sizes are large, or (2) combinations of very small true heterogeneity, large variance in primary study reliabilities, large mean effect sizes, and a limited number of primary studies. As expected, heterogeneity estimates generally improved with higher average reliability, and deteriorated with smaller average study sample size. Our results were approximately the same whether one used a classic random-effects meta-analytic model (Hedges & Vevea, 1998) or a 'bare-bones' psychometric meta-analytic model (Schmidt & Hunter, 2015), and whether one used Pearson's r as the effect size or Fisher's z, although the negative bias was slightly larger with Fisher's z. The severity of the underestimation of heterogeneity could lead to meta-analyses in psychology missing relevant moderators and overgeneralizing outcomes. Unreliability appears to bring truly different effects closer together, thereby obscuring true differences between studies that could be relevant for theory, practice, and future research efforts.

In meta-analyses, low reliabilities not only lower the average effect size estimate, but also artificially suppress heterogeneity estimates in most cases. For a typical effect size around r = 0.2 and substantial true heterogeneity ($\tau > 0.1$) that is not uncommon in the meta-analytic literature (Van Erp et al., 2017), we found expected heterogeneity to be underestimated by about 15% - 60% in our simulations primarily depending on average reliability in primary studies (higher reliability leads to a smaller underestimate), although the underestimate could be as high as 73% with low sample sizes. As true heterogeneity increases, the proportional underestimate decreases to some degree, but the underestimate in absolute terms will worsen. Variability in reliabilities across primary studies is a counter-acting force that inflates heterogeneity estimates as average effect size increases. Consequently, the underestimate of heterogeneity can

become less severe as effect size increases, although the overall bias is likely to remain negative. Small degrees of true heterogeneity ($\tau < .08$) can be overestimated under certain conditions, but this would typically require a combination of a small number of studies, large variance in study reliabilities, and a large average effect size. The severity and generality of the underestimate in heterogeneity due to unreliability means that meta-analysts should not ignore the effect of reliability on their heterogeneity estimates and that research is needed to provide practical guidance to meta-analysts.

Without any information on primary studies' reliabilities, it would often be impossible to know whether a small observed heterogeneity is due to overestimated true zero heterogeneity or underestimated small heterogeneity. Moreover, it is certainly possible to observe a zero heterogeneity estimate under larger true heterogeneity levels, because the sampling variability in heterogeneity estimates is often substantial in practice (Ioannidis et al., 2007; Olsson-Collentine et al., 2020). Generally, distinguishing between zero and small heterogeneity in empirical meta-analyses is statistically challenging considering the typical sample sizes and numbers of studies in meta-analyses in psychology (Olsson-Collentine et al., 2020). The power to detect small true heterogeneity in most psychological meta-analyses is low. In our view, the distinction between zero and small heterogeneity in meta-analysis is primarily a hypothetical one as true zero heterogeneity is not expected in meta-analytic practice to begin with (see also Hedges, 1987), except perhaps in specific sets of very similar direct replications (Olsson-Collentine et al., 2020).

We found that using Pearson's r or Fisher's z as our effect size resulted in very similar degrees of bias in heterogeneity estimates due to uncorrected unreliability in primary studies. The underestimate is slightly worse for Fisher's z than for Pearson's r. The usage of Pearson's r inflates heterogeneity estimates slightly (due to its bounded nature and effect size dependent sampling variance), which is apparently beneficial for heterogeneity estimates when not correcting for unreliability. However, the additional negative bias when using Fisher's z is small and we consider the difference negligible in the light of the severe underestimates generated by unreliability.

Should meta-analysts correct for unreliability in primary studies?

Our simulations show that heterogeneity estimates are biased when unreliability is ignored, so in principle one should correct for unreliability. However, even done perfectly this would not mean that heterogeneity estimates end up to be accurate, because few meta-analyses are able to include sufficient studies (Davey et al., 2011; Van Erp et al., 2017) to measure heterogeneity without large uncertainty and wide confidence intervals (Ioannidis et al., 2007; Olsson-Collentine et al., 2020). In practice, attempting to correct for unreliability comes with many caveats and assumptions that must be considered. We first consider challenges with correcting for unreliability in the CTT framework presuming that CTT assumptions hold. Next, we consider potential concerns with the CTT assumptions and whether structural equation modeling (SEM) is a viable alternative in the meta-analytic context.

There are several concerns with correcting for unreliability in the CTT framework even when its assumptions hold. First, reliability estimates are lower bound estimates, meaning that systematically correcting for unreliability can be expected to lead to overestimated mean effect sizes and subsequent overestimated heterogeneity. Unfortunately, the most popular reliability estimate (Cronbach's Alpha, Flake et al., 2017; Hogan et al., 2000) is known to underestimate the lower bound even further than several alternatives, particularly when the strong assumptions underlying CTT are not met (Sijtsma, 2008). Second, CTT corrections of correlations as commonly applied (Schmidt & Hunter, 2015; Charles, 2005; but see Raju et al., 1991) do not consider that reliability estimates themselves are subject to substantial sampling error (Fan & Thompson, 2001) that add additional uncertainty to corrected effect sizes and subsequent estimates of heterogeneity. Third, there are several types of reliability estimates (e.g., "internal consistency", test-retest, coefficient of equivalence) that do not necessarily derive from the same type of measurement error and may result in different reliability estimates (Dimitrov, 2002; Revelle & Condon, 2019). If the reliability estimate applied does not account for all relevant sources of measurement error, it will overestimate reliability and heterogeneity will remain underestimated despite unreliability corrections. Fourth, reliability estimates are sometimes not reported in primary studies (Flake et al., 2017). For meta-analysis, missing estimates are then often imputed based on average observed reliability or based on some 'typical' reliability (Gnambs & Sengewald, 2023; Schmidt & Hunter, 2015). However, what a 'typical' reliability should be is open to debate (LeBreton et al., 2014), and mean imputation leads to a biased pooled estimate, especially so if missing statistics are related to the value of the statistic (and it seems likely that non-reporting is more common when reliability is low). Mean imputation is seen by experts in missing data as "possibly the worst missing data handling method available" (Enders, 2010, p. 43; Schafer & Graham, 2002, p. 159; see also Van Buuren, 2018, p. 12).

Correcting for unreliability in the CTT framework

comes with several assumptions that may be hard to motivate in practice. First, most estimates of reliability in CTT assume unidimensionality (e.g., Borsboom & Mellenbergh, 2002; Savalei & Reise, 2019), and violations of this assumption can severely bias reliability estimates (Savalei & Reise, 2019). Second, correcting for unreliability in the CTT framework assumes linear associations between scale and latent construct, which implies that corrections may result in biased estimates when ordinal scales (such as Likert-scales) are short (Zhang, 2022). Based on their simulations, Zhang (2022) recommends that ordinal scales should have at least 12 items (with ≥ 5 answer categories per item) to use Cronbach's Alpha to correct for unreliability in meta-analysis. Problematically, Flake et al. (2017) report that the average scale length in psychology (excluding 1-item scales) is 4.69 items (SD = 6.35). Third, true scores X and Y are assumed to be normally distributed. Zhang (2022) demonstrated in simulation studies that when this assumption does not hold average effect size estimates will be inaccurate, especially when the average effect size is large (although they did not examine heterogeneity). Fourth, estimators such as Cronbach's Alpha assume uncorrelated errors (e.g., Dimitrov, 2002; Teo & Fan, 2013) and essential tau-equivalence (i.e., items have equal factor loadings on a latent construct, e.g., Revelle & Condon, 2019). Correlation of errors can lead to a substantial overestimate of reliability (e.g., Dimitrov, 2002; Teo & Fan, 2013), and may occur in many measurement situations, such as when items on a scale refer to the same stimuli (e.g., multiple reading comprehension items referring to the same paragraph, Green & Yang, 2009). Overestimation of reliability would leave heterogeneity estimates with a negative bias remaining even after CTT correction. The assumption of essential tau-equivalence is empirically unrealistic, and violations lead to an underestimate of the reliability statistic (Sijtsma, 2008, p. 111). This bias may often be small enough to be neglected (Green & Yang, 2009; Savalei & Reise, 2019), but can interact with the assumption of normal distributed true scores and result in inaccurate estimates of average effect sizes even for ordinal scales with more than 12 items (Zhang, 2022). Extending the study of Zhang (2022) also to examine the effects on heterogeneity estimates would be a useful avenue of future research.

When all of the CTT assumptions hold, and if practical concerns can be overcome, CTT corrections of unreliability appear to function well in meta-analysis of correlations (Brannick et al., 2019; e.g., Hall & Brannick, 2002; Schmidt & Hunter, 2015). However, research appears to be lacking on the consequences of CTT corrections for meta-analysis when assumptions do not hold (but see Zhang, 2022). Moreover, because latent constructs are defined by latent variable models (Borsboom & Mellenbergh, 2002), the strong assumptions of CTT regarding latent associations, factor loadings, uncorrelated errors, and the distribution of true scores may be unpalatable to many researchers. Modern latent models based on Structural Equation Modelling (SEM, e.g., Lei & Wu, 2007) may then be an option, as they permit delving into measurements and latent construct structures rather than relying on assumptions.

Meta-analysis can be conceptualized as a SEM model (e.g., Cheung, 2015), which holds promise that the benefits of modern psychometric theory might be applied also to latent constructs summarized across multiple studies. In Meta-Analytic Structural Equation Modeling (MASEM, e.g., Cheung & Chan, 2005) unreliability in primary studies can be corrected in three ways (Gnambs & Sengewald, 2023): (1) by estimating a SEM model in each primary study and then pooling estimates (parameter-based MASEM, Cheung, 2015, p. 241), (2) by directly modeling latent variables in the MASEM, or (3) by first correcting correlations before introducing them into the MASEM framework (i.e., fundamentally identical to the CTT approach). The first option is limited because it is rarely the case that the same SEM model can be fit across multiple independent datasets, except in highly structured cases (e.g., Brunner et al., 2022). The second option is the preferable option, but requires item-level statistics rather than summary values, which are rarely available to meta-analysts. The third option, naturally, suffers the same challenges and assumptions as corrections under classical test theory. As such, the potential to use MASEM to take advantage of modern psychometric theory when correcting for unreliability is challenging in practice. That said, multi-lab replication projects are increasingly common in psychology (e.g., Klein et al., 2014; McCarthy et al., 2021) and consist of a set of direct replication studies where analysts have access to all data. As such, they are well-placed to take full advantage of MASEM for their analyses.

In the end, meta-analysis depends on the quality of the primary studies that make up its data. As long as researchers in psychology underappreciate the importance of measurement, meta-analytic estimates will be poor. Meta-analysts that consider the challenges and assumptions of CTT corrections of unreliability acceptable should acknowledge the many caveats inherent in doing so and report average effect size and heterogeneity estimates based on both corrected and uncorrected estimates. Meta-analysts who consider CTT corrections untenable need to consider that heterogeneity estimates may be severely underestimated. In either case, meta-analysts should get used to extracting reliability information whenever available from primary studies.

Limitations

Arguably the biggest challenge in estimating heterogeneity in meta- practice is the large degree of sampling variance in heterogeneity estimates that can be expected in almost any meta-analysis (Ioannidis et al., 2007; Olsson-Collentine et al., 2020). Consider a meta-analysis without measurement error of k =12 studies (the median number in psychology, Van Erp et al., 2017) with N = 150 in all studies, $\tau = 0.17$ (the median observed τ for correlations in Van Erp et al., 2017), and $\mu = 0$. Across 10,000 replications the median width of the 95% confidence interval around the heterogeneity estimate (using default settings in metafor) is then 0.195 in Pearson's r units. In our study, we did not examine sampling variance in heterogeneity estimates or their root mean squared error. This is because although we apply different meta-analytic models and effect sizes, the purpose was not to compare the efficiency of these different modes of estimation. As the effect of unreliability in primary studies is simply to change the heterogeneity of effect sizes, the results of previous studies on the comparable performance of different heterogeneity estimators applies also here (e.g., Langan et al., 2017; Veroniki et al., 2016; Viechtbauer, 2005). That said, it would be useful to examine the variance of heterogeneity estimates in future research when the focus is on developing practical guidance for meta-analysts. Doing so could be helpful to gain a better idea of how likely a particular observed heterogeneity estimate is to belong to different distributions of true heterogeneity in the presence of measurement unreliability. As this would depend on N, K, mean reliability, variability in reliability, average effect size, observed heterogeneity, and the true heterogeneity levels being compared, implementing such an analysis would probably be best done on a case-by-case basis (e.g., through a webapp such as developed for sensitivity analyses of the effect of publication bias on heterogeneity by Augusteijn et al., 2019) . Given the challenges in correcting for unreliability, such a sensitivity analysis could be a useful tool to reason around the impact of unreliability for applied meta-analysts. An additional challenge when providing practical advice to meta-analysts is how to handle selective reporting of study reliabilities (Hussey et al., 2023)) in primary studies, and the problems that p-hacking and publication bias create in estimating heterogeneity. We know from Augusteijn et al. (2019) that publication bias alone creates complex outcomes, but when meta-analysis is affected by unreliability and p-hacking true heterogeneity is additionally obscured.

Our simulations come with some caveats. First, we assume reliabilities to be known. This is never the case, and any estimate of reliability is at best a lower bound estimate subject to sizeable sampling errors. By assuming reliabilities to be known, we were able to show the bias of observed heterogeneity estimates compared to the ideal case of no measurement error. However, this may mislead readers to believe that a simple correction for unreliability will leave them with an unbiased estimate. As discussed in the section 'should meta-analysts correct for unreliability?' correcting for reliability under realistic circumstances and using estimated reliability is not straightforward and whether the resulting estimate will be unbiased is contingent on several factors that remain to be studied. Examining the effect of corrections based on estimated reliabilities under realistic conditions in future studies could offer practical guidance to meta-analysts.

Second, our simulations assumed equal reliability for the measures X and Y (i.e., $R_i = R_{xx'} = R_{yy'}$) within a study, which is not realistic. However, as we never use the reliabilities individually (i.e., only expressed as $\sqrt{R_{xx'}} \times \sqrt{R_{yy'}}$) we do not expect major impact on our results. Consider that if $R_i = 0.7$ then $\sqrt{R_{xx'}} \times \sqrt{R_{yy'}} = 0.7$ whereas if if $R_{xx'}$ and $R_{yy'}$ took other values from a distribution centered on 0.7 this value would be lower (e.g., if $R_{xx'} = 0.5$ and $R_{yy'} = 0.9$ then $\sqrt{R_{xx'}} \times \sqrt{R_{yy'}} \approx 0.67$). As we use the results of $\sqrt{R_{xx'}} \times \sqrt{R_{yy'}}$ in our simulations to generating observed effect sizes from true effect sizes (i.e., $r_{xy} = \rho_{xy} \times \sqrt{R_{xx'}} \times \sqrt{R_{yy'}}$, where ρ_{xy} is the true effect size after sampling error), letting the reliability for X and Y vary would lead to slightly larger attenuation of observed effect sizes on average. That is, letting the reliability of X and Y vary within studies would lead to larger negative bias in observed heterogeneity estimates than in our simulations, indicating that the effect might even be somewhat more severe than in our results.

Third, we used a fixed sample size (N) across studies within a meta-analysis in our simulations, which is unrealistic. We showed that increasing the primary study sample size led to better heterogeneity estimates, especially for low reliability. There is large variability in how simulation studies focused on heterogeneity implement sample sizes (Langan et al., 2017). Most studies sample from a uniform distribution of sample sizes, but there are also studies that have used conditions with fixed sample sizes like us (Langan et al., 2019). There appears to be no systematic exploration of the effect of different distributions of sample sizes on heterogeneity estimation, and the effect of sample size is generally considered in view of its average value (e.g., Viechtbauer, 2005). The range of sample sizes in a meta-analysis appears to affect the performance of heterogeneity estimation (Langan et al., 2017), but we believe this relates only to the efficiency of heterogeneity estimates rather than bias, given a sufficiently large number of repetitions in the simulations. That said, implementing a re-

alistic distribution of sample sizes would be important for research attempting to provide concrete guidance to meta-researchers on how to interpret observed heterogeneity in the presence of measurement error as variability in estimates then is important. Implementing a distribution of sample sizes based on relevant empirical literature (e.g., as done by Brannick et al., 2019) could be one way to increase realism in simulated sample sizes.

Finally, we only considered bias in heterogeneity due to unreliability in primary studies, but there exist other measurement artifacts such as restriction of range, dichotomization, and more (Schmidt & Hunter, 2015). Although measurement error is the only measurement artifact that is always present, many of these other measurement artifacts may impact heterogeneity estimates when they are present. Based on Schmidt & Hunter (2015) 's arguments that measurement artifacts all tend to attenuate effect sizes in a similar way to unreliability (p. 78-79), we expect that their net effect on heterogeneity estimates may similarly be a negative bias. The extent to which this prediction holds true (and whether these measurement artifacts can be corrected for in MASEM, Gnambs & Sengewald, 2023) awaits further study.

Conclusion

Imperfect measurement reliability in primary studies generally leads to a severe underestimate of observed meta-analytic heterogeneity. Unreliability may thereby obscure true differences between studies that could be relevant for theory, practice, and future research efforts. As few meta-analyses in psychology correct for unreliability in primary studies, heterogeneity is likely underestimated in a large proportion of meta-analyses in psychology. Yet, sophisticated methods for correcting measurement unreliability such as meta-analytic structural equation modeling (MASEM) are only applicable in exceptional cases and corrections based on classical test theory come with caveats and strong assumptions that are often unrealistic and currently do not fully consider insights from modern test theory. Accurate estimation of meta-analytic heterogeneity is difficult and will remain so unless measurement concerns (Flake et al., 2017; Flake & Fried, 2020; Kane et al., 2021) are taken seriously in primary research. The good news is that study designs (multi-lab replication studies) which can apply more sophisticated versions of MASEM are becoming increasingly common in psychology, allowing us to study how to best deal with measurement errors when estimating variation of true effect sizes in meta-analyses

REFERENCES

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2011). Meta-Analytic Choices and Judgment Calls: Implications for Theory Building and Testing, Obtained Effect Sizes, and Scholarly Impact. *Journal of Management*, *37*(1), 5–38. https://doi.org/10.1177/0149206310377113

Augusteijn, H. E. M., van Aert, R. C. M., & van Assen, M. A. L. M. (2019). The effect of publication bias on the Q test and assessment of heterogeneity. *Psychological Methods*, *24*(1), 116–134. https://doi.org/10.1037/met0000197

Borenstein, M. (2009). *Introduction to meta-analysis*. Wiley-Blackwell.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*(6), 505–514. https://doi.org/10.1016/S0160-2896(02)00082-X

Brannick, M. T., Potter, S. M., Benitez, B., & Morris, S. B. (2019). Bias and Precision of Alternate Estimators in Meta-Analysis: Benefits of Blending Schmidt-Hunter and Hedges Approaches. *Organizational Research Methods*, *22*(2), 490–514. https://doi.org/10.1177/1094428117741966

Card, N. A. (2015). *Applied meta-analysis for social science research*. Guilford Publications.

Charles, E. P. (2005). The Correction for Attenuation Due to Measurement Error: Clarifying Concepts and Creating Confidence Sets. *Psychological Methods*, *10*, 206–226. https://doi.org/10.1037/1082-989X.10.2.206

Cheung, M. W. L. (2015). *Meta-analysis : A structural equation modeling approach*. John Wiley & Sons, Inc.

Cheung, M. W. L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods*, *10*(1), 40–64. https://doi.org/10.1037/1082-989X.10.1.40

Constantin, M. (2023). *Parabar: Progress bar for parallel tasks* [Manual].

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA.

Cortina, J. M. (2003). Apples and Oranges (and

Pears, Oh My!): The Search for Moderators in Meta-Analysis. *Organizational Research Methods*, *6*(4), 415–439. https://doi.org/10.1177/1094428103257358

Davey, J., Turner, R. M., Clarke, M. J., & Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: A cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, *11*(1), 160. https://doi.org/10.1186/1471-2288-11-160

Dimitrov, D. M. (2002). Reliability: Arguments for Multiple Perspectives and Potential Problems with Generalization across Studies. *Educational and Psychological Measurement*, *62*(5), 783–801. https://doi.org/10.1177/001316402236878

Dowle, M., & Srinivasan, A. (2021). *Data.table: Extension of 'data.frame'* [Manual].

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Fan, X., & Thompson, B. (2001). Confidence Intervals About Score Reliability Coefficients, Please: An EPM Guidelines Editorial. *Educational and Psychological Measurement*, *61*(4), 517–531. https://doi.org/10.1177/00131640121971365

Field, A. P. (2005). Is the Meta-Analysis of Correlation Coefficients Accurate When Population Correlations Vary? *Psychological Methods*, *10*(4), 444–467. https://doi.org/10.1037/1082-989X.10.4.444

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

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*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fraley, R. C., Chong, J. Y., Baacke, K. A., Greco, A. J., Guan, H., & Vazire, S. (2022). Journal N-Pact Factors From 2011 to 2019: Evaluating the Quality of Social/Personality Journals With Respect to Sample Size and Statistical Power. *Advances in Methods and Practices in Psychological Science*, *5*(4), 25152459221120217. https://doi.org/10.1177/25152459221120217

Gnambs, T., & Sengewald, M.-A. (2023). Meta-Analytic Structural Equation Modeling With Fallible Measurements. *Zeitschrift Für Psychologie*, *231*(1), 39–52. https://doi.org/10.1027/2151-2604/a000511

Green, S. B., & Yang, Y. (2009). Commentary on Coefficient Alpha: A Cautionary Tale. *Psychometrika*, *74*(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

Hafdahl, A. R., & Williams, M. A. (2009). Meta-analysis of correlations revisited: Attempted replication and extension of Field's (2001) simulation studies. *Psychological Methods*, *14*, 24–42. https://doi.org/10.1037/a0014697

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology*, *87*(2), 377–389. https://doi.org/10.1037/0021-9010.87.2.377

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist*, *42*(5), 443–455. https://doi.org/10.1037/0003-066X.42.5.443

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Higgins, J. P. T. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability Methods: A Note on the Frequency of Use of Various Types. *Educational and Psychological Measurement*, *60*(4), 523–531. https://doi.org/10.1177/00131640021970691

Hussey, I., Alsalti, T., Bosco, F., Elson, M., & Arslan, R. C. (2023). *An aberrant abundance of Cronbach's alpha values at .70*. PsyArXiv. https://doi.org/10.31234/osf.io/dm8xn

Ioannidis, J. P. A., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*, *335*(7626), 914–916. https://doi.org/10.1136/bmj.39343.408449.80

Kane, M. J., Smeekens, B. A., Meier, M. E., Welhaf, M. S., & Phillips, N. E. (2021). Testing the construct validity of competing measurement approaches to probed mind-wandering reports. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01557-x

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Kvarven, A., Strømland, E., & Johannesson, M. (2019). Comparing meta-analyses and

preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*, 423–434. https://doi.org/10.1038/s41562-019-0787-z

Langan, D., Higgins, J. P. T., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W., & Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*, *10*(1), 83–98. https://doi.org/10.1002/jrsm.1316

Langan, D., Higgins, J. P. T., & Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. *Research Synthesis Methods*, *8*(2), 181–198. https://doi.org/10.1002/jrsm.1198

LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for Criterion Reliability in Validity Generalization: A False Prophet in a Land of Suspended Judgment. *Industrial and Organizational Psychology*, *7*(4), 478–500. https://doi.org/10.1111/iops.12184

Lei, P.-W., & Wu, Q. (2007). Introduction to Structural Equation Modeling: Issues and Practical Considerations. *Educational Measurement: Issues and Practice*, *26*(3), 33–43. https://doi.org/10.1111/j.1745-3992.2007.00099.x

Lord, M., F., & Novick, M. R. (1968). *Statistical theories of mental test scores / frederic M. Lord and melvin R. Novick ; with contributions by allan birnbaum.* Addison-Wesley Pub. Co Reading, Mass.

McCarthy, R., Gervais, W., Aczel, B., Al-Kire, R. L., Aveyard, M., Marcella Baraldo, S., Baruh, L., Basch, C., Baumert, A., Behler, A., Bettencourt, A., Bitar, A., Bouxom, H., Buck, A., Cemalcilar, Z., Chekroun, P., Chen, J. M., del Fresno- Díaz, Á., Ducham, A., … Zogmaister, C. (2021). A Multi-Site Collaborative Study of the Hostile Priming Effect. *Collabra: Psychology*, *7*(1), 18738. https://doi.org/10.1525/collabra.18738

Nuijten, M. B., van Assen, M. A. L. M., Augusteijn, H. E. M., Crompvoets, E. A. V., & Wicherts, J. M. (2020). Effect Sizes, Power, and Biases in Intelligence Research: A Meta-Meta-Analysis. *Journal of Intelligence*, *8*(4), 36. https://doi.org/10.3390/jintelligence8040036

Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*(10), 922–940. https://doi.org/10.1037/bul0000294

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*(2), e1–e4. https://doi.org/10.1017/iop.2015.17

R Core Team. (2020). *R: A language and environment for statistical computing* [Manual]. R Foundation for Statistical Computing.

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology*, *76*(3), 432–446. https://doi.org/10.1037/0021-9010.76.3.432

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2012). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology*, n/a–n/a. https://doi.org/10.1111/j.2044-8317.2012.02057.x

Sassenberg, K., & Ditrich, L. (2019). Research in Social Psychology Changed Between 2011 and 2016: Larger Sample Sizes, More Self-Report Measures, and More Online Studies. *Advances in Methods and Practices in Psychological Science*, *2*(2), 107–114. https://doi.org/10.1177/2515245919838781

Savalei, V., & Reise, S. P. (2019). Don't Forget the Model in Your Model-based Reliability Coefficients: A Reply to McNeish (2018). *Collabra: Psychology*, *5*(1), 36. https://doi.org/10.1525/collabra.247

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00813

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 251524592110074. https://doi.org/10.1177/25152459211007467

Schmidt, F. (2010). Detecting and Correcting the Lies That Data Tell. *Perspectives on Psychological Science*, *5*(3), 233–242. https://doi.org/10.1177/1745691610369339

Schmidt, F., & Hunter, J. E. (2015). *Methods of Meta-Analysis : Correcting Error and Bias in Research Findings* (Third edition). SAGE.

Seymour, R. T. (1988). Why plaintiffs' counsel challenge tests, and how they can successfully challenge the theory of "validity generalization." *Journal of Vocational Behavior*, *33*(3), 331–364. https://doi.org/10.1016/0001-8791(88)90043-7

Sijtsma, K. (2008). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107. https:

//doi.org/10.1007/s11336-008-9101-0

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*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72. https://doi.org/10.2307/1412159

Teo, T., & Fan, X. (2013). Coefficient Alpha and Beyond: Issues and Alternatives for Educational Research. *The Asia-Pacific Education Researcher*, *22*(2), 209–213. https://doi.org/10.1007/s40299-013-0075-z

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology*, *53*(2), 207–216. https://doi.org/10.1016/S0895-4356(99)00161-4

Van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition* (Second). Chapman and Hall/CRC. https://doi.org/10.1201/9780429492259

Van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of Between-Study Heterogeneity for 705 Meta-Analyses Reported in *Psychological Bulletin* From 1990–2013. *Journal of Open Psychology Data*, *5*(1), 4. https://doi.org/10.5334/jopd.33

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*(1), 55–79. https://doi.org/10.1002/jrsm.1164

Viechtbauer, W. (2005). Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293. https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting Meta-Analyses in *R* with the **Metafor** Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods*, *20*(3), 360–374. https://doi.org/10.1037/met0000023

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining Unbiased Results in Meta-Analysis: The Importance of Correcting for Statistical Artifacts. *Advances in Methods and Practices in Psychological Science*, *3*(1), 94–123. https://doi.org/10.1177/2515245919885611

Zhang, Q. (2022). Meta-analysis of correlation coefficients: A cautionary tale on treating measurement error. *Psychological Methods*. https://doi.org/10.1037/met0000498