Title:

“Unreliable heterogeneity: why most meta-analyses in psychology are likely to underestimate heterogeneity”

Abstract:

One of the main purposes of meta-analysis is to estimate heterogeneity of effect sizes. The quality of meta-analytic estimates depends in large part on the quality of evidence in primary studies, yet there exist broad concerns that proper measurement tends to be neglected in much of psychological primary research. Such inattention to measurement can also bias heterogeneity estimates. We performed a Monte-Carlo simulation study to examine the expected bias in heterogeneity estimates due to unreliability in primary studies. Although unreliability in primary studies can both inflate and suppress heterogeneity, our results indicate that under most circumstances unreliability in primary studies leads to a severe underestimate of heterogeneity in meta-analyses. For a typical effect size heterogeneity can be expected to be underestimated by 15-60%, or more if average reliability in primary studies is below 0.6. 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. 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. Research on concrete guidance to applied meta-analysts is urgently 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) measurements (e.g., Shadish, Cook, & Campbell, 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, 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 & Novick, 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; Hunter & Schmidt; 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 Hunter & Schmidt (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. Wiernick and 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., Cortina, 2003; Anguinis, 2011).

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., 2017). 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., 2017). 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 the 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 & Scharz, 2019; Scheel et al., 2020) 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; Hunter & Schmidt, 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.

## Bias in heterogeneity estimates due to unreliability

In classical test theory (CTT), reliability of measurement () is defined as the proportion of true score variance () to observed score variance () and is presumed to be constant for all participants in a sample . We use unreliability of measurement to refer to . Imperfect reliability in measurements leads to attenuation of observed effect sizes. For product-moment correlations the observed correlation can be computed as where is the observed correlation between the variables and , is the true correlation (with or without sampling error) and and are the measurement reliabilities for and . As such, correcting for unreliability is straightforward for product-moment correlations, but can also be done for other effect size types (Wiernick and 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 then the corrected sampling variance (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 across studies (i.e., a fixed effect and no true heterogeneity) and infinite sample size within studies. Any observed effect size can then be seen as the true effect size multiplied by a study’s reliability (assuming ). Hence, the distribution of observed effect sizes is the true effect size multiplied by the distribution of reliabilities across studies. That is, , where is the average reliability across studies and 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 and estimated heterogeneity will be larger than true heterogeneity (i.e., zero) by the product of (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.

Table 1.

*Variance in reliability inflates heterogeneity in combination with increased average effect size and leads to overestimated heterogeneity in a fixed effects model*

[1] "Table 1 approximately here. Cannot be printed nicely to Word, see tables.pdf"

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 implies that any effect size will be attenuated by the same proportion given the same reliabilities , this proportion corresponds to a larger absolute value for larger true 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 supression can be observed directly by assuming an average effect size with some heterogeneity , infinite sample size within studies (i.e., no within-study sampling variance), and equal reliability across studies (). Any observed effect size can then be seen as a study’s true effect size multiplied by the study reliability . Hence, the distribution of observed effect sizes is the distribution of true effect sizes multiplied by the reliability , such that . 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 (if heterogeneity is expressed as standard deviation rather than variance). This depiction is not entirely correct for Pearson’s as it is bounded at {-1, 1}, but holds approximately if the mean effect size and heterogeneity 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, , 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.

Table 2.

*Imperfect reliability suppresses true heterogeneity*

[1] "Table 2 approximately here. Cannot be printed nicely to Word, see tables.pdf"

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; Hunter & Schmidt, 2015; Wiernick & 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 sampling error and estimation of reliabilities, particularly 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 (REF) and we used the R package ‘metafor’ (REF) to perform meta-analyses and estimate heterogeneity. For parallel computing we took advantage of the R packages ‘parallel’ (REF) and ‘parabar’ (REF).

## 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; Wiernick & 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, 2020, supplement A).

Perhaps the most common way to meta-analyze correlations is to first transform them into Fisher correlations (Borenstein et al., 2009, p. 41 - 43) to normalize their distribution. The Fisher transformation is the inverse hyperbolic tangent of the product-moment correlation and can be computed as and its variance approximated as . There is controversy in the use of the Fisher transformation for meta-analysis (e.g., Hunter & Schmidt, 2015, p. XX, ch. 5 “Use of Fisher’s z in Meta-Analysis of Correlations”), largely related to whether Pearson’s or Fisher’s 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 transformation (Field, 2005; Hafdahl & Williams, 2009, Hunter & Schmidt, 2015). Hall and 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 andthe common use of the Fisher transformation in practice, we use it as one of the effect size types in our simulations.

We also use Pearson’s . Pearson’s has the advantage that it creates heterogeneity estimates on the same scale that ate more interpretable hetereogeneity estimates on the Fisher scale. Although Fisher correlations can be back-transformed into Pearson’s 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 and Fisher’s (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 when estimating heterogeneity are that (1) the effect size is bounded to {-1, 1} and (2) its sampling variance depends on the true effect size . The bounded nature of the Pearson correlation means that large levels of heterogeneity create substantial truncation 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 XX). This can be solved by replacing with the average correlation across studies (Hunter & Schmidt, 2015), or using the Fisher transformation. Fisher correlations have the additional advantage that they are not bounded to {-1, 1}. A disadvantage to using Fisher 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 for meta-analysis. Especially for large Fisher values (which are common when the 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*.

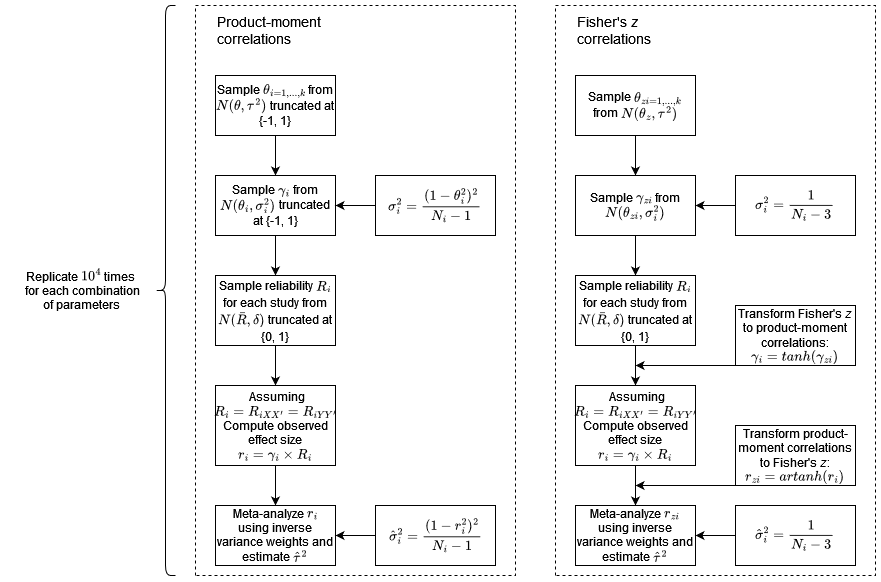
## Meta-analytic model

We apply two meta-analytic models to ensure that our results are not model-dependent: the Hedges and Vevea (1998) random-effects model and the Hunter and Schmidt (2015) ‘bare-bones’ random-effects model. The Hedges and Vevea (HV) model uses inverse variance weight computed as (e.g., Borenstein et al., 2010) where is the sampling variance of study and is the estimated between-studies variance. The Hunter and Schmidt (HS) model weighs studies by sample size . The differences between these weights are relatively minor (Borenstein et al., 2010; Hunter and Schmidt, 2015, p. XX). Although the HV model is typically applied to Fisher transformed correlations (e.g., Borenstein et al., p. 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’; Hunter & Schmidt, 2015) because this allows us to focus on the bias in heterogeneity estimates rather than the performance of corrections.

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; Hunter & Schmidt, 2015, p. XX). 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 $\tau^2 = \sigma^2\_r - \sigma^2\_e$ where $\sigma^2\_r$ is the observed variance across correlations and $\sigma^2\_e$ is the average sampling error variance across studies, computed as $ \frac{\Sigma[N\_i \sigma^2\_{ei}]}{\Sigma N\_i}$ (p. 87, 2nd edition) and $\sigma^2\_{ei}$ is the usual sampling variance for the Pearson correlation but with $\rho$ 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 can be found in Viechtbauer (2005; 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 , with some minor differences we highlight in the text below detailing the procedure.



(FIGURE 1: a flowchart summarizing the simulation design)

We generated data for the meta-analyses as follows. First, we sampled true study effect sizes from . That is, a normal distribution with a mean of and a variance of that we truncated at to stay within the bounds of the product-moment correlation. We implemented the truncation in R using inverse transform sampling (link: functions.r). For Fisher’ no truncation was applied.

For each of the sampled true study effect sizes we then sampled one observed effect size from a truncated normal distribution , using the same truncation procedure. The sampling variance for each true effect size we computed using the standard formula for Pearson’s where is the total sample size for study . For Fisher’s no truncation was applied and we approximated the sampling variance as . At this point we have sampled effect sizes in Pearson’s or Fisher’s without measurement error.

To add measurement error, we first sampled a reliability for each study from a truncated () normal distribution , given some average reliability across studies and standard deviation in reliability . We assumed that both the dependent and independent variable within a study were measured with the same reliability such that . We then computed observed effect sizes for each study given the attenuation formula . Because we assumed equal reliability in and this simplifies to . Fisher’s we transformed to product-moment correlations before adding measurement error and then back-transformed into Fisher’s before the next step. At this point we have effect sizes in Pearson’s or Fisher’s 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 and number of studies . Within meta-analyses we used a fixed sample size across all studies such that . We made this choice rather than treating sample size as a random variable (as done by e.g., Field, 2005; Brannick et al., 2019) 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 9 psychology journals between 2011 - 2018 which ranged between 69 - 496 depending on journal and year. Sassenber and Ditrich (2019) find a median sample size of 110 (interquartile range, 71 - 195) across 4 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 6 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, , 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 then result in a larger proportion of truncated estimates below zero because of increased sampling variance in . As such, values of can affect bias, especially at lower levels of heterogeneity. We consider the following values of : {5, 20, 40, 100}.

We vary the grand mean 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 , we transformed these values￼ scale. For context to these values, we considered the empirical estimates of typical correlational effect sizes provided by Nuijten et al. (2020) and Schäfer & Scharz (2019). Nuijten et al. performed a meta-meta-analysis on the fields of intelligence research and reported a median meta-analytic correlation of 0.24 for predictive validity and correlational studies. This estimate is likely positively biased due to publication bias and selective reporting. Schäfer & Scharz reported￼ 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, 2021).

We defined the between-studies standard deviation () 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 value in these data was 0.17 (interquartile range 0.1 - 0.24). The point of this study is that empirical heterogeneity estimates may not be accurately estimated. Nonetheless, we selected approximately the interquartile range of -values reported in van Erp et al. 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 truncation for our maximum average effect size of 0.6: 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 $= $ 0, 0.08, 0.13, 0.2).

Fisher’s is measured at a different scale and it is not possible to directly convert -values expressed in Pearson’s into corresponding -values on the Fisher’s scale. To compare results between Fisher’s and the product-moment correlation we defined heterogeneity on the scale. The index is a relative measure of heterogeneity describing the percentage of total variance that is due to heterogeneity (Higgins, 2002; 2003), and as such is measured from 0 - 100%. The index can be defined as where and is the precision . This means that the value generally depends on both the number of studies and the sample size in these studies (Borenstein, 2017). However, because we keep fixed across studies, the value of depends only on and the heterogeneity.

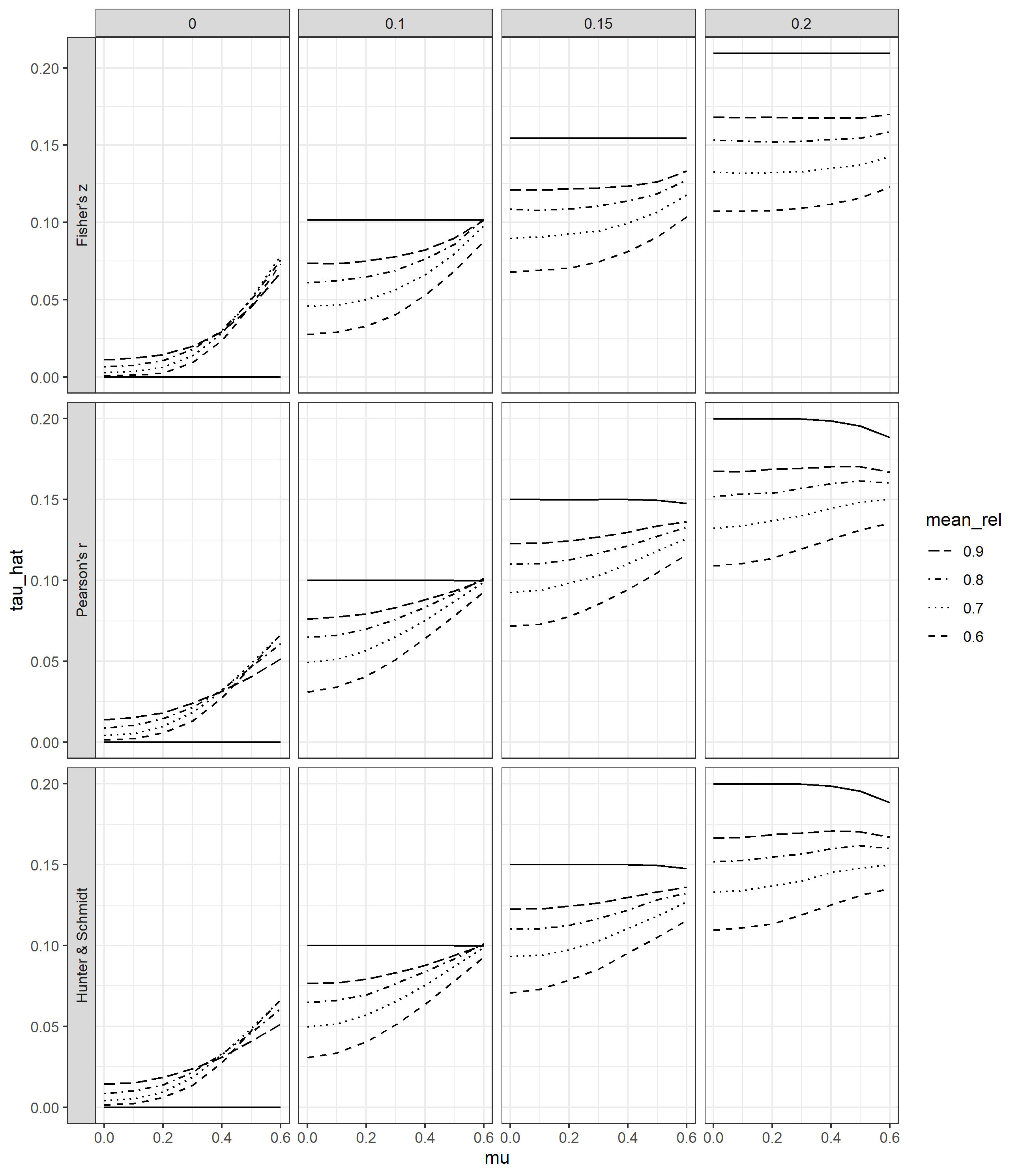
We thus computed the index for each sample size condition and -values defined in Pearson’s , and then computed corresponding -values on the Fisher’s scale given these -values and sample sizes. Two complications were that the value for varies with effect size for Pearson’s and that effect size varies across studies because of >0￼￼￼￼￼￼￼￼￼￼￼￼￼￼￼￼ increased heterogeneity values were more similar (see supplement XX).

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. (2003) reported the reliability estimates based on five so-called reliability generalization studies. These five studies summarized reliability 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 , and standard deviations .

# Results

Overall, the net bias in heterogeneity estimates due to imperfect reliability in primary studies tended to be 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 XX). 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 $\tau = 0.1$. Hence, we primarily report results for k = 20 in the main manuscript, and only report results using other values for k (5, 40, 200) when considering heterogeneity below $\tau = 0.1$ and in supplement XX. Finally, the effect on bias in heterogeneity estimates of varying within-study sample size was relatively small, and we similarly report only results for N = 150 in the main manuscript and present results for other sample sizes in supplement XX.

Figure 2 shows the net bias in heterogeneity estimates for different levels of heterogeneity (measured in ) and meta-analytic models. Each column of Figure 2 corresponds to one nominal level of actual heterogeneity ( {0, 0.1, 0.15, 0.2}) and each row to a different analysis (HV with Fisher , HV with Pearson’s , and HS). The black solid lines indicate actual heterogeneity, which may differ from nominal heterogeneity because of truncation or, in the case of Fisher , because the comparable heterogeneity levels were slightly different (as discussed in the method section). The different dashed lines in Figure 2 show the 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 than the other two methods (although Fisher is on a different scale) but the results are approximately the same for all three methods. As such, we will focus in the remainder of the results on Pearson’s (row two).

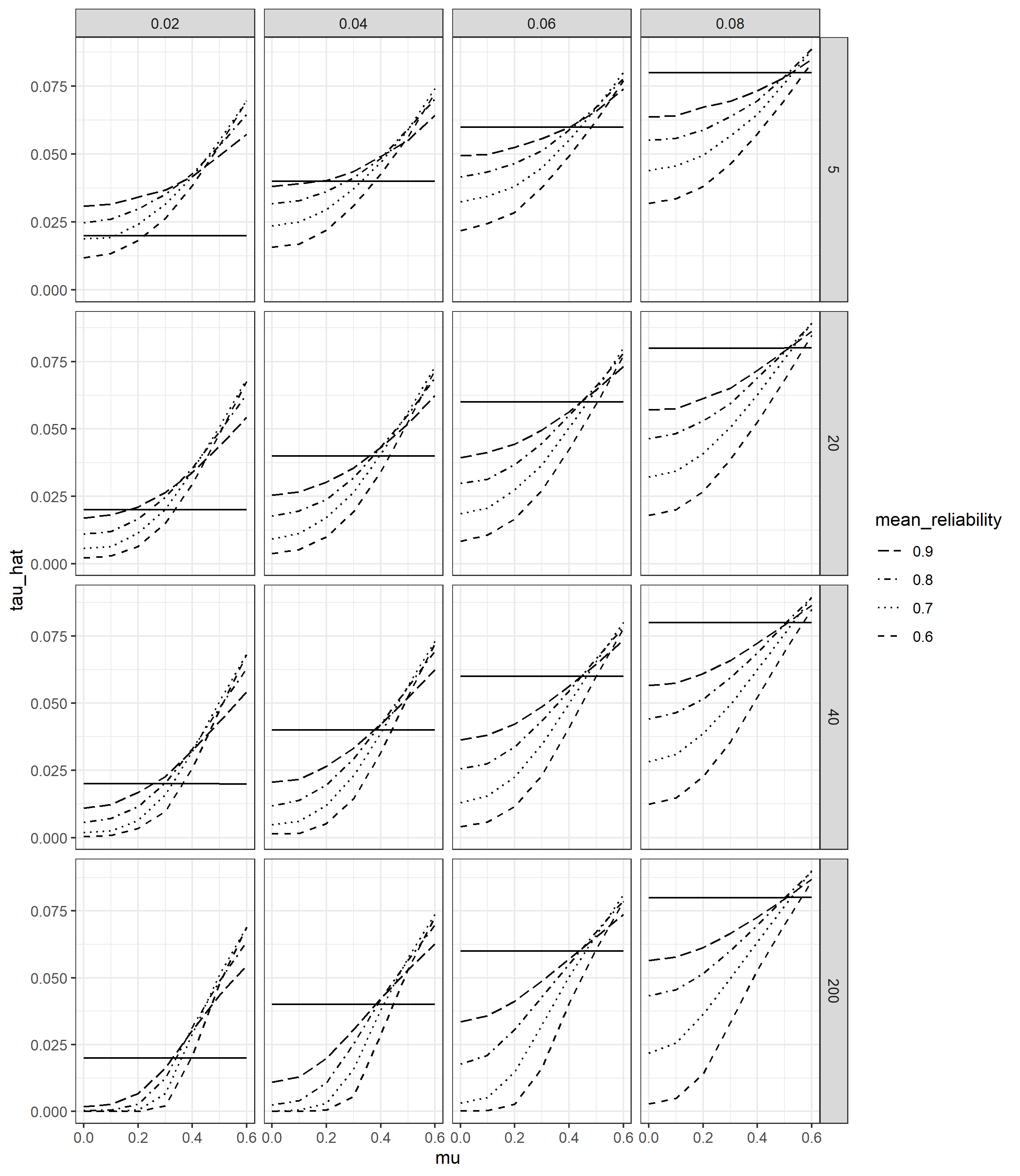


*Figure 2.* Unreliability in primary studies leads to a net negative bias when heterogeneity is positive. The x-axis indicates average effect size and the y-axis estimated heterogeneity in standard deviations. Columns indicate the nominal true heterogeneity standard deviation . Due to truncation in Pearson’s or translation to Fisher’s these values may differ from actual true heterogeneity standard deviation (black solid lines). Each dashed line corresponds to an average reliability in primary studies. Standard deviation in reliability was 0.15. 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 as the type of effect size additionally inflates heterogeneity estimates to a small degree due to the dependence between effect size and sampling variance (supplement XX), . First, and the only of these two that applies when , 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. For example, for (the closest value to the median correlation reported by Schäfer & Schwarz) and the bias for Pearson’s compared to true zero heterogeneity is 0.006. For (the upper median reported by Schäfer & Schwarz) and the bias is 0.027. 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 when the reliability is bounded at 1. The inflation of heterogeneity with effect size can also be seen in the presence of heterogeneity (all other columns). However, it is superseded by the suppression of heterogeneity due to unreliability.

For all true heterogeneity levels in Figure 2 above zero, we see an overall underestimate of heterogeneity due to unreliability in 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 , , , the bias is -0.03 (30%) for Pearson’s , for , , the bias is -0.037 (25%), and for , , the bias is -0.046 (23%). For an effect size of true heterogeneity between 0.1 - 0.2 is underestimated by between 16% (, ) and 59% (, ). 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 . We used the following values {0.02, 0.04, 0.06, 0.08}. Figure 3 presents the results of this analysis for Pearson’s . Because these lower heterogeneity levels can be expected to be affected by 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 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 then imperfect reliability leads to a negative bias in heterogeneity estimates for all in Pearson’s . Even for lower levels of heterogeneity, 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 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. Columns indicate the nominal true heterogeneity standard deviation . Due to truncation in Pearson’s or translation to Fisher’s these values may differ from actual true heterogeneity standard deviation (black solid lines). Each dashed line corresponds to an average reliability in primary studies. Standard deviation in reliability was 0.15. Code to reproduce figure: LINK.

# 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 expeted, estimates generally improved with higher average reliability. 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 (Hunter & Schmidt, 2015), and whether one used Pearson’s as the effect size or Fisher’s , although the negative bias was slightly larger with Fisher’s . 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 estimate, but also artifically suppress heterogeneity estimates. For a typical effect size around = 0.2 and substantial true heterogeneity () that is not uncommon in the meta-analytic literature (van Erp et al., 2017), heterogeneity can be expected to be underestimated by about 15% - 60% primarily depending on average reliability in primary studies (higher reliability leads to a smaller underestimate). 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 bias is likely to remain negative. Small degrees of true heterogeneity () 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.

It is often impossible to know whether a small degree of observed heterogeneity is due to overestimated true zero heterogeneity or underestimated small heterogeneity. Moreover, it is possible to observe a zero heterogeneity estimate under larger true heterogeneity levels as sampling variability in heterogeneity estimates is often substantial in practice (Ioannidis et al., 2007; Olsson-Collentine et al., 2020). Generally, separating between zero and small heterogeneity is infeasible in empirical research considering realistic sample sizes and typical numbers of studies in meta-analyses in psychology (Olsson-Collentine et al., 2020), and 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 (see also Hedges, 1987).

We found that using Pearson’s or Fisher’s 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 than for Pearson’s . The usage of Pearson’s 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 is small and we consider the difference negligible in the context 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 were accurate, because few meta-analyses are able to include sufficient studies (van Erp et al., 2017; Davey et al., 2011) to measure heterogeneity without large uncertainty and wide confidence intervals (Ioannidis et al., 2007, Olsson-Collentine, 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 assumptions hold. First, reliability estimates are lower bound estimates, meaning systematically correcting for unreliability can be expected to lead to overestimated mean effect sizes and overestimated heterogeneity. Unfortunately, the most popular reliability estimate (Cronbach’s Alpha; Flake et al., 2017; Hogan et al., 2002) 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 (Hunter & Schmidt, 2015; but see Raju et al., 1991; Charles, 2005) 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 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 (Hunter & Schmidt, 2015; Gnambs & Sengewald, 2023). 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; see also van Buuren, 2018, p. 12; Schafer, 2002, p. 159).

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 & Mellenberh, 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 (Zheng, 2022). Based on their simulations, Zheng (2022) recommends that ordinal scales should have at least 12 items (with $\geq$ 5 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. Zheng (2022) demonstrated in simulation studies that when this assumption does not hold average 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, 2009, p. 111). This bias may often be small enough to be neglected (Green & Yang, 2009; Savale & Reise, 2019), but can interact with the assumption of normal true scores and result in inaccurate estimates even for ordinal scales with more than 12 items (Zheng, 2022).

When all assumptions hold, and if practical concerns can be overcome, CTT corrections of unreliability appear to function well in meta-analysis of correlations (e.g., Hall & Brannick, 2002; Brannick et al., 2019; Hunter & Schmidt, 2015). Research appears to be lacking on the consequences of CTT corrections for meta-analysis when assumptions do not hold (although 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 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 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 and future research

Perhaps the major challenge in estimating heterogeneity is the large degree 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 $\tau$ 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., 2017). 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.

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 so simple 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 would be useful to do in future studies to provide practical guidance to meta-analysts.

Second, our simulations assumed equal reliability for the measures and (i.e., ) within a study, which is not realistic. However, as we never use the reliabilities individually (i.e., only expressed as ) we do not expect major impact on our results. Consider that if then whereas if if and took other values from a distribution centered on 0.7 this value would always be lower (e.g., if and then ). As we use the results of in our simulations to generate observed effect sizes from true effect sizes (i.e., , where is the true effect size after sampling error), letting the reliability for and vary would lead to slightly larger attenuation of observed effect sizes on average. That is, letting the reliability of and vary within studies would lead to larger negative bias in observed heterogeneity estimates than in our simulations, indicating that the effect might be somewhat more severe than in our estimates.

Third, we used a fixed sample size (N) across studies within a meta-analysis in our simulations, which is unrealistic. We showed in Supplement XX that increasing the fixed sample size led to slightly better observed heterogeneity estimates 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., 2018). 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 realistic 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 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 (Hunter & Schmidt, 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 Hunter & Schmidt’s arguments that measurement artifacts all tend to attenuate effect sizes in a similar way to unreliability (p. XX), we expect that their net effect on heterogeneity estimates may similarly be a negative bias. The extent to which this prediction is true (and whether these measurement artifacts can be corrected for in MASEM; Gnambs & Sengewalds, 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 be obscuring 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. Accurate estimation of meta-analytic heterogeneity is difficult and will remain so unless measurement concerns (Flake & Fried, 2020; Flake et al., 2017; 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.