Estimating heterogeneity of effect sizes is usually considered the main purpose of meta-analysis in addition to 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, or more specifically to sensitivity to differences in four study factors: 1) sample population, 2) settings, 3) treatment variables and 4) measurement variables (e.g., Campbell & Stanley, 2015). Heterogeneity is a primary outcome in meta-analysis because its presence signals the existence of moderators and so can be seen as an opportunity for theoretical development (Simons et al., 2017), because it affects the implementation of research by indicating that an intervention may not be equally effective under all conditions or for all people, and because its presence changes the interpretation of the average effect sized derived in meta-analysis from the true population effect size (under homogeneity) to the average of true subpopulation effect sizes (under heterogeneity). As such, estimating heterogeneity without 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 are highly dependent 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 reliability of measurement (defined in classical test theory as the ratio of true score variance to measurement error variance; e.g., REF), which is the only measurement validity evidence researchers report consistently (Flake et al., 2017), and which can be corrected for in meta-analysis (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. 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 can refer to two different entities. Either effect sizes free from sampling error, or effect sizes free from both sampling error and measurement artifacts (such as imperfect reliability, restriction of range, and dichotomization; Hunter & Schmidt, 2015). Whereas all meta-analytic models attempt to correct for sampling error in their estimates, correcting for measurement artifacts is implicit only in the models of 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 researchers to misinterpret their meta-analytic average effect size estimate, ignore the presence of moderators or search for (and discover) non-existent moderators, and to implement research interventions inappropriately.

Meta-analyses in psychology rarely correct for unreliability in primary study measurements. Wiernick and Dahlke (2020) report 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) report 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 is generally controversial and has been so since it was proposed (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 be expected to lead to overestimation of effect sizes, because reliability estimates are lower bound estimates (i.e., a reliability estimate of 0.8 implies the reliability is between 0.8 - 1). Even worse, the most commonly reported estimate of reliability in psychology (Chronbach’s alpha; Flake et al., 2017) as well as improvements thereof tend to underestimate the lower bound of reliability (REF Klaas). 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 (REF?), 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 artifacts 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). In the end, whether correcting for unreliability or not, authors and consumers of meta-analyses need to be aware of how this decision affects heterogeneity estimates.

## Bias in heterogeneity estimates due to unreliability

In classical test theory (CTT), reliability of measurement () is defined as the proportion 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). Imperfect reliability in primary studies imbues heterogeneity estimates with both a positive and a negative bias.

A positive bias can be expected 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 seen 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 leads to a positive bias in heterogeneity that depends 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 leads to positive bias in heterogeneity estimates which increases with average effect size*

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

A negative bias can be expected 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 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 negative bias 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 create a negative bias in 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 leads to negative bias in heterogeneity estimates*

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

It is difficult to predict the total effect of the negative and positive bias in heterogeneity due to unreliability, and existing literature tends to focus on the positive bias (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 negative bias could often be 20% or more, we consider it important for meta-analysts to have insight into the total bias in heterogeneity estimates that should be expected due to unreliability. This depends on average effect size, true heterogeneity, mean reliability across studies, the variability in reliabilities, and sampling variance within studies. Hence, we perform a Monte-Carlo simulation study to explore the expected bias in heterogeneity estimates due to unreliability in primary studies.

# 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 commonly 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). 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 in Brannick et al., (2019) they examine results only after corrections for attenuation. Given the inconclusive evidence and that the Fisher transformation is often applied in practice, we use it as one of the effect size types in our simulations.

The other effect size type we us is Pearson’s . Pearson’s has the advantage that resulting heterogeneity estimates on this scale are more interpretable that 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 is that 1) the effect size is bounded to {-1, 1} and 2) the sampling variance depends on the true effect size . The bounded nature of the Pearson correlation means that large levels of heterogeneity will lead to 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 backtransforming into product-moment correlations before adding measurement error, and then transforming to Fisher again for meta-analysis. Especially for large Fisher values (which can happen if 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.

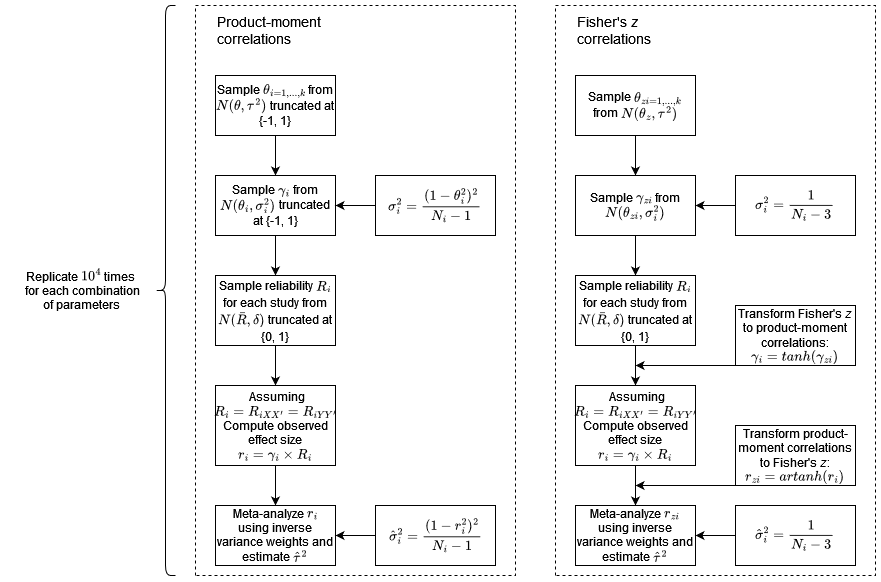
## Meta-analytic model

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 correcting 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). Describing the REML estimator in detail is beyond the scope of this paper and we refer interested readers to the description in Veroniki et al., (2016). The Hunter and Schmidt (2015) model estimates heterogeneity as where is the observed variance across correlations and is the average sampling error variance across studies, computed as $ $ (p. 87, 2nd edition) and is the usual sampling variance for the Pearson correlation but with replaced with the sample size weighted average correlation instead of 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)

When generating data for meta-analysis we proceeded 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., 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. (personal communication, 2023) 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 use, restricted maximum likelihood, and most other heterogeneity estimators are truncated at zero (Viechtbauer, 2005). Lower values for then results 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 value into equivalent effect sizes at the Fisher’s scale. For context to these values consider 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 reports a median meta-analytic effect size of 0.24 for predictive validity and correlational studies. This estimate is likely positively biased due to publication bias and selective reporting. Schäfer & Scharz report a median 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) provide empirical heterogeneity estimates from 747 meta-analyses in 61 article 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 excissive 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.

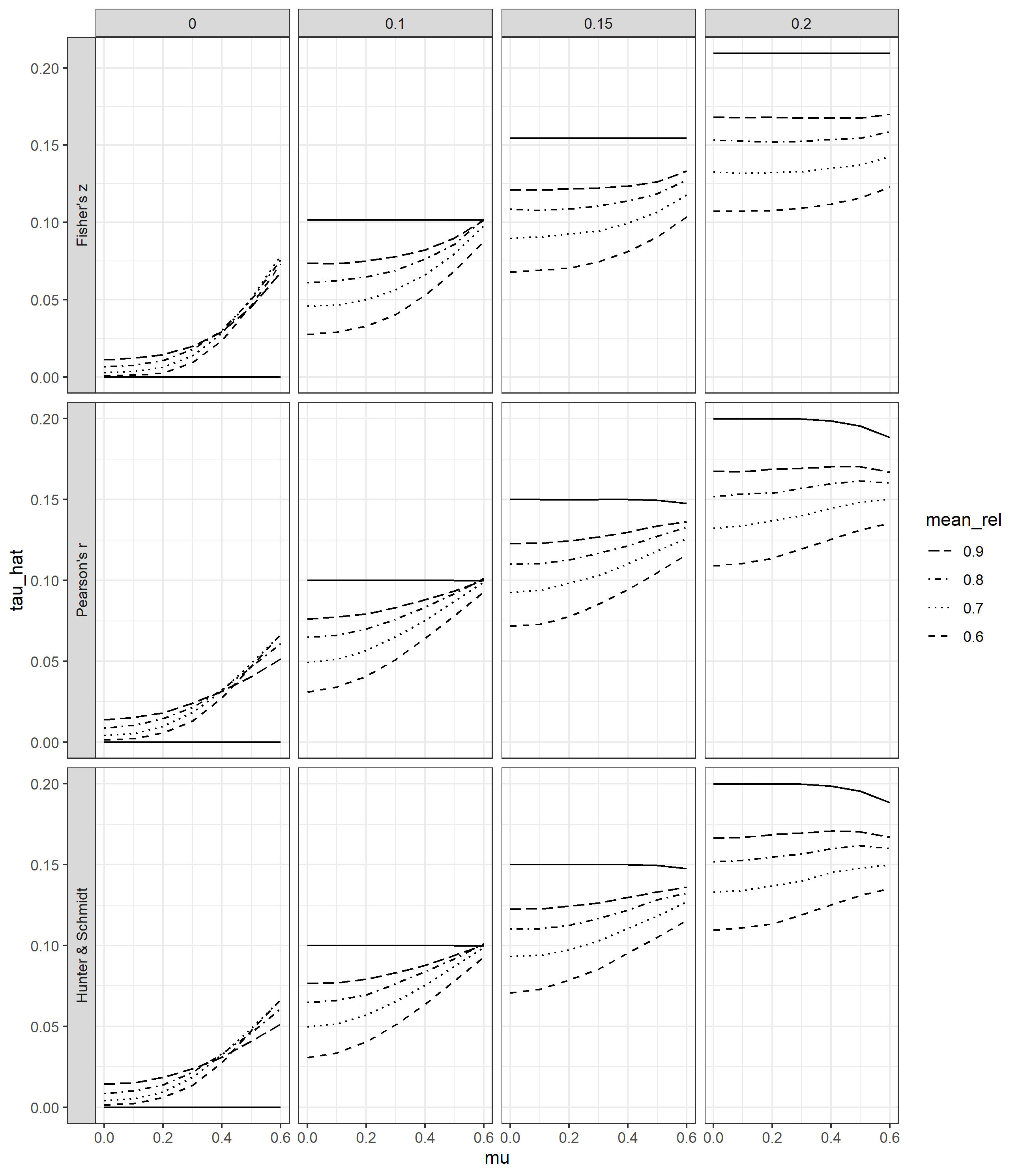
We thus computed the index for each sample size condition and -values defined in Pearson’s , and then computed corresponding -values on the Fisher’s scale given these -values and sample sizes. Two complications were that the value for varies with effect size for Pearson’s and that effect size varies across studies because of . We resolved this by setting and computing the expected given and sample size and then using this expected value to compute . We used the ‘law of the unconscious statistician’ to compute the expected value of , which says that the expected value of a function of a random variable can be expressed in terms of the probability distribution of X: . In our case, corresponds to the sampling variance of the product-moment correlation, and because the product-moment correlation is bounded at {-1, 1} we set instead to 0.999. The resulting values for Fisher were only minimally different from {0, 0.1, 0.15, 0.2}. For example, for the corresponding Fisher -values were {0, 0.1031263, 0.1566007, 0.2123514} and as 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) report that based on 245 estimates of Cronbach’s alpha in psychology the average estimate was 0.79 with a standard deviation of 0.13. The interquartile range was approximately 0.68 - 0.87 for studies wherein the associated scale lacked a reference and 0.79 - 0.88 for scales that did have a reference. Sánchez-Meca et al. (2003) report the reliability estimates based on five ‘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 (Hussey et al., 2023), we examined the following mean reliabilities , and standard deviations .

# Results

The overall bias in heterogeneity estimates due to imperfect reliability in primary studies tended to be negative. We found that the positive bias due to variance in reliability across studies was less than the negative bias due to average unreliability across all levels of heterogeneity in our primary analysis. As such, we report here only the condition with maximum variance in reliability (SD = 0.15). Analyses with lower levels of reliability variance can be found in 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 heterogeneity was above 0.1. As such, we report only results for k = 20 in the main manuscript. Analyses with other values for k (5, 40, 200) can be found 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 1 show the absolute and relative bias in heterogeneity estimates for different levels of heterogeneity (measured in ) and meta-analytic models. Each column of Figure 1 corresponds to one nominal level of 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 unbroken 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 (see methods). The different dashed lines in Figure 1 shows the estimates for four different levels of average reliability across 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, we will focus in the remainder of the results on Pearson’s r (row two).

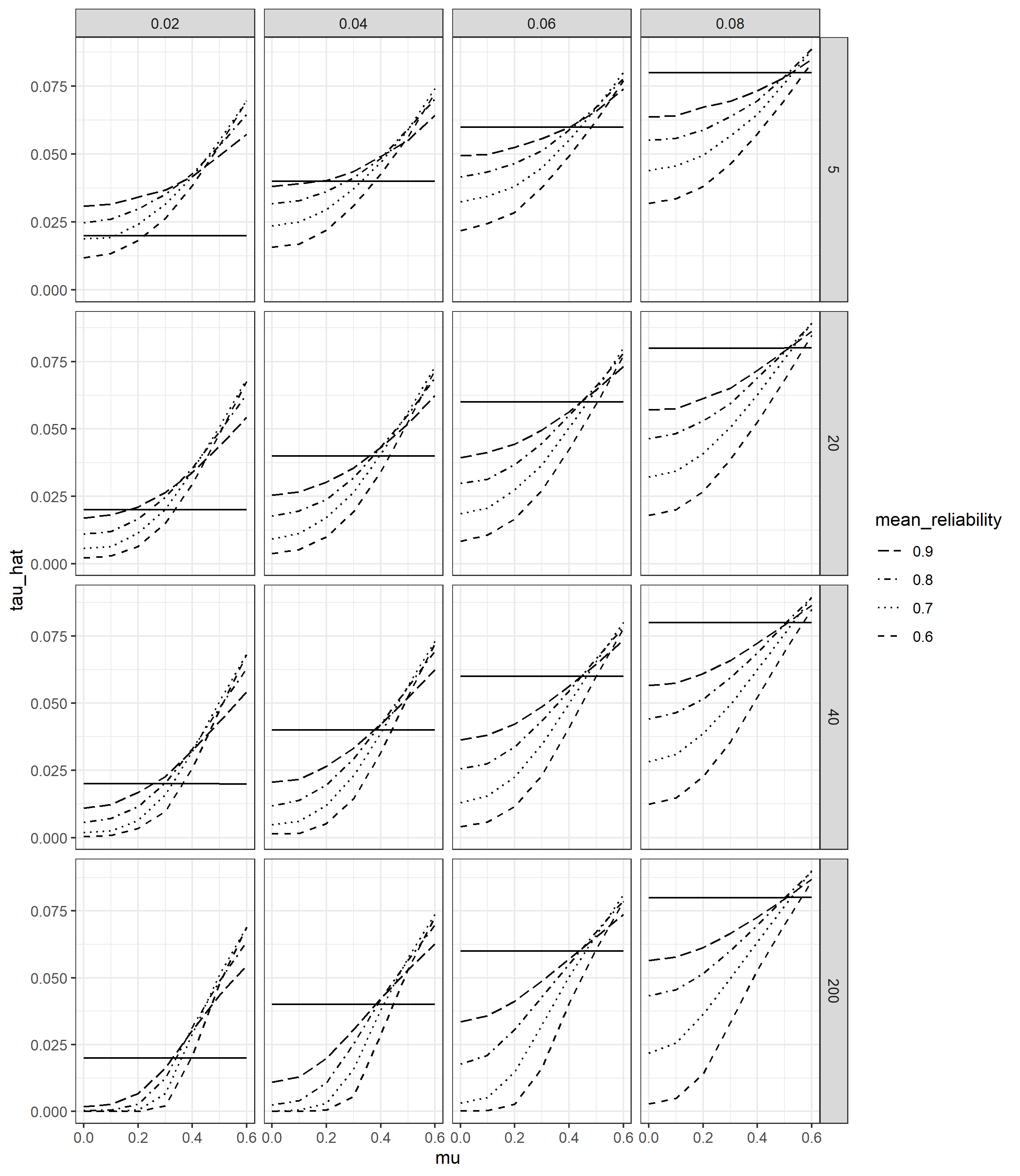


*Figure 1.* EXPLANATION. Code to reproduce figure: LINK.

As expected, for zero heterogeneity (leftmost column Figure 1) we see an overestimate of heterogeneity that increases with effect size. There are two sources of positive bias in the absence of heterogeneity. First, and the only one that applies when , there is positive bias due to the truncation of negative heterogeneity estimates (Viectbauer, 2005). Second, there is bias due to the variance in reliability. For example, for (the closest value to the median correlation reported by Schäfer & Schwarz) and the bias for Pearson’s is XX. For (the upper median reported by Schäfer & Schwarz) and the bias is XX. For the highest average reliability (0.9) a somewhat less inclined slope can be seen. This is because the standard deviation for this reliability level is severely truncated when the reliability is bounded at 1. The increase in positive bias with effect size can also be seen in the presence of heterogeneity (all other columns). However, it is superseded by the negative bias.

For all heterogeneity levels in Figure 1 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 lower average reliability, and is worse in an absolute sense but better in a relative sense as heterogeneity increases. For example, for , , , the bias is XX (XX%) for Pearson’s r, for , , the bias is XX (XX%), and for , , the bias is XX (XX%).

To explore further under what conditions the positive and negative bias in heterogeneity estimates would 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 2 presents the results of this analysis for Pearson’s r. Because these lower heterogeneity levels can be expected to be affected by truncation to a larger extent (which results in a positive bias), we present results for four different numbers of studies in the meta-analysis (rows Figure 2, lower is associated with larger positive bias). The standard deviation in reliabilities is again 0.15. Figure 2 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 2.* EXPLANATION. Code to reproduce figure: LINK.