Estimating heterogeneity of effect sizes is usually considered the second main purpose of meta-analysis, after 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). From a more statistical perspective, heterogeneity is defined as the variance between ‘true’ subpopulation effect sizes. However, true subpopulation 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 measurement error, restriction of range, and dichotomization; Hunter & Schmidt, 2015). Of these measurement artifacts, measurement error is present in any empirical effect size estimate and systematically attenuates effect sizes. It is usually measured as reliability (defined in classical test theory as the ratio of true score variance to measurement error variance; e.g., REF) and is relatively common to report in primary studies in psychology (Flake et al., 2017). However, whereas all meta-analytic models attempt to correct for sampling error in their estimates, correcting for imperfect reliability is implicit only in the models of Hunter & Schmidt (2015). This has consequences for heterogeneity estimates reported in many areas of psychology and in this article we use Monte-Carlo simulations to explore the consequences for heterogeneity estimates in meta-analysis from not correcting for unreliability.

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. Although at the meta-analysis level rather than study level, Cortina (2003) showed that about 80% of the 1,294 meta-analyses published across 59 meta-analytic studies in the Journal of Applied Psychology between 1978 - 1997 corrected for measurement artifacts, and Anguinis (2011) that around 50% of meta-analyses identified in studies published across several management journal between 1989 - 2009 corrected for unreliability of measurement.

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 (e.g., REF1, REF2, REF3), 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 should be aware of how this decision affects heterogeneity estimates.

Imperfect reliability in primary studies imbues heterogeneity estimates with both a positive and a negative bias. The positive bias is discussed by several highly cited meta-analytic handbooks, most prominently by Hunter and Schmidt (2015; 12,235 citations on google scholar as of 2023-05-11) who introduced the concept of psychometric meta-analysis (i.e., meta-analysis corrected for measurement artifacts such as unreliability), but also by Borenstein et al. (2009, p. 342; 19,020 citations), Card (2015, p.126; 1,716 citations) as well as in recent literature (Wiernick and Dahlke, 2020). However, there appears to be a lack of acknowledgement that imperfect reliability can also lead to a negative bias in heterogeneity estimates. The only one of the above sources to refer to a negative bias is Hunter and Schmidt who mention it shortly in their 2nd edition of *Methods of Meta-analysis* (2004, p. 142) in a paragraph later absent from the 3rd edition (Hunter & Schmidt, 2015). Below we detail the argument for both a positive and negative bias in 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).

A positive bias can be expected in heterogeneity estimates of uncorrected effect sizes due to the variance 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 infinite sample size (i.e., no within-study variance) and in the context of a fixed effect model (i.e., with a common underlying mean across studies and no true heterogeneity). Table 1 illustrates how under these conditions three studies with differing reliability will observe variability in effect sizes beyond their true variability (i.e., zero). Table 1 also shows how this effect increases as the true effect size increases.

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 given the same unreliability, the attenuation in absolute value increases with effect size. 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. As a consequence, in the presence of heterogeneity larger true effect sizes will move further towards zero than smaller ones, decreasing heterogeneity in observed 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"

Attenuation of effect size can also be understood as a downwards shift of the distribution of true effect sizes. To illustrate, assume an infinite sample size (i.e., no within-study sampling variance), some heterogeneity , equal reliability in outcome and measurement such that , and that all studies within a population of studies have the same reliability. Observed effect sizes given measurement error then corresponds to multiplying all true effect sizes by a factor equal to the reliability (i.e., multiplying the distribution of effect sizes by a constant ). If the distribution of effect sizes is normally distributed with a mean (i.e., the assumption of the random effects model), then . Because reliability is always a value < 1, it can be seen that the square root of the observed heterogeneity for uncorrected effect sizes is less than its true value by a factor . This depiction is not entirely correct for Pearson’s as it is bounded at {-1, 1}, but holds approximately as long as the mean effect size and heterogeneity are not too large.

To make this concrete, 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. We can consider the positive bias in the same vein. Assuming a common true effect size across studies and infinite sample size within studies we can view as a constant. In this case, attenuation consists of a distribution of reliabilities with some mean and standard deviation across studies, that is multiplied by the constant to yield a distribution of observed effect sizes . To make this concrete, consider some values , , which then yield . Again, this depiction is not entirely correct as reliability is bounded at {0, 1}, but holds approximately as long as the mean reliability and its standard deviation are not too large.

It is difficult to predict the total effect of the negative and positive bias in heterogeneity due to unreliability. This depends on average effect size, the amount of 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 when meta-analyzing primary studies with imperfect measurement reliabilities.