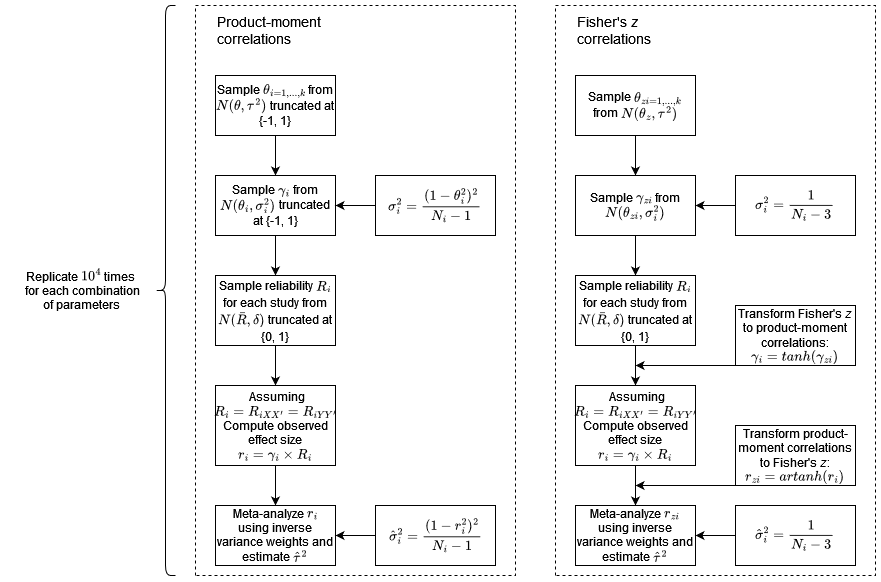
# Methods

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).

## 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. For Fisher’ no truncation was applied.

Practically, we implemented the truncation as follows. First, we sampled probabilities from an uniform distribution where and are the values given by the cumulative distribution function (CDF) of the normal distribution at . Second, we then back-transformed the sampled values onto the distribution using the inverse of the CDF.

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, ready for meta-analysis.

We meta-analyzed the effect sizes using a random effects model with inverse variance weights and the Restricted Maximum Likelihood (REML) estimator of heterogeneity. We choose the REML estimator because it is a heterogeneity estimator that performs well and is widely implemented (e.g., REF, REF2), but also show our results using the classic DerSimonian-Laird (DL) estimator in supplement X. Similarly, inverse variance weights are widely used, perform similarly to other common weights (e.g., REF1 REF2) and we see no reason other implementations would change our results substantially. The procedure described in this section was replicated 10,000 times for each combination of parameter values.

## Parameter values

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 the interaction between effect size and reliability variance. For Fisher’s we transformed these value into equivalent effect sizes at the Fisher’s scale. To put these values into context we turn to the empirical estimates of typical correlational effect sizes provided by Nuijten et al., (REF) and Schäfer & Scharz (REF). Nuijten et al. performed a meta-meta-analysis on the fields of 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 FIELD, 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 (REF Scheel).

We defined the between-studies variance () to cover a wide range of variance in true effect sizes and empirical estimates reported in the psychological literature. Van Erp et al., (REF) provide empirical heterogeneity estimates from a large number of meta-analyses. The median Pearson’s value in these data was 0.03 (interquartile range 0.01 - 0.0576). As the point of this study is that empirical heterogeneity estimates may not be accurately estimated we selected a range of estimates for the product-moment that we considered reasonable and that also covered most of the empirical estimates from van Erp et al. These were {0, …, VALUES}.

Fisher’s z 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 be able 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 (REF Higgins), an as such is measured from 0 - 100%. The index can be defined as where and is the precision . This means that its value generally depends on both the number of studies and the sample size in these studies (REF: Higgins, I2 is not an absolute measure of heterogeneity). However, because we keep fixed across studies (see below), the value of depends only on and the heterogeneity. We thus compute the index for each sample size condition and -values defined in Pearson’s , and then compute corresponding -values on the Fisher’s scale given these -values and sample sizes.

We ran our analyses across a range of 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 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 by XXX (still needs lookup) and consider the following values {VALUES}.

When focusing on the bias of heterogeneity estimate, the value of 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 (REF Viechtbauer). Lower values for then results in a larger proportion of truncated estimates below zero because of increased sampling variance. As such, the values of can affect bias, especially at lower levels of heterogeneity. In their review of meta-analyses in psychology, van Erp et al. (YEAR) find values ranging from XX to XX, and in our simulations we consider the following values {VALUES}.

Finally, we defined average reliability levels and their standard deviation based on empirical estimates from the literature. Flake et al. (YEAR) 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. Sanchez-Meca et al. (REF) 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, we examined the following mean reliabilities , and standard deviations .