Supplement C: varying reliability SD and the effect of K

Lower standard deviation in reliability leads to a more severe underestimate for large effect sizes. Figure C1 demonstrates this effect for four levels of between-studies standard deviation (columns Figure C1) and four levels of reliability standard deviation (rows Figure C1) when sample size (N) is set to 150 within studies and the number of meta-analyzed studies (K) is 20. All results in this manuscript are based on the outcome variable Pearsons's r estimated with Hedges & Vevea (1998) approach and Restricted Maximum Likelihood (REML) and averaged across 10,000 replications. As can be seen in Figure C1, once reliability standard deviation reaches 0.05 (third row) there is practically no upward inflation of heterogeneity remaining, only just sufficient to compensate for the suppression due to the truncation of Pearson's r at {-1, 1}. A such, especially in the case of large effect sizes heterogeneity can be expected to be somewhat more negatively biased as reliability standard deviation decreases.

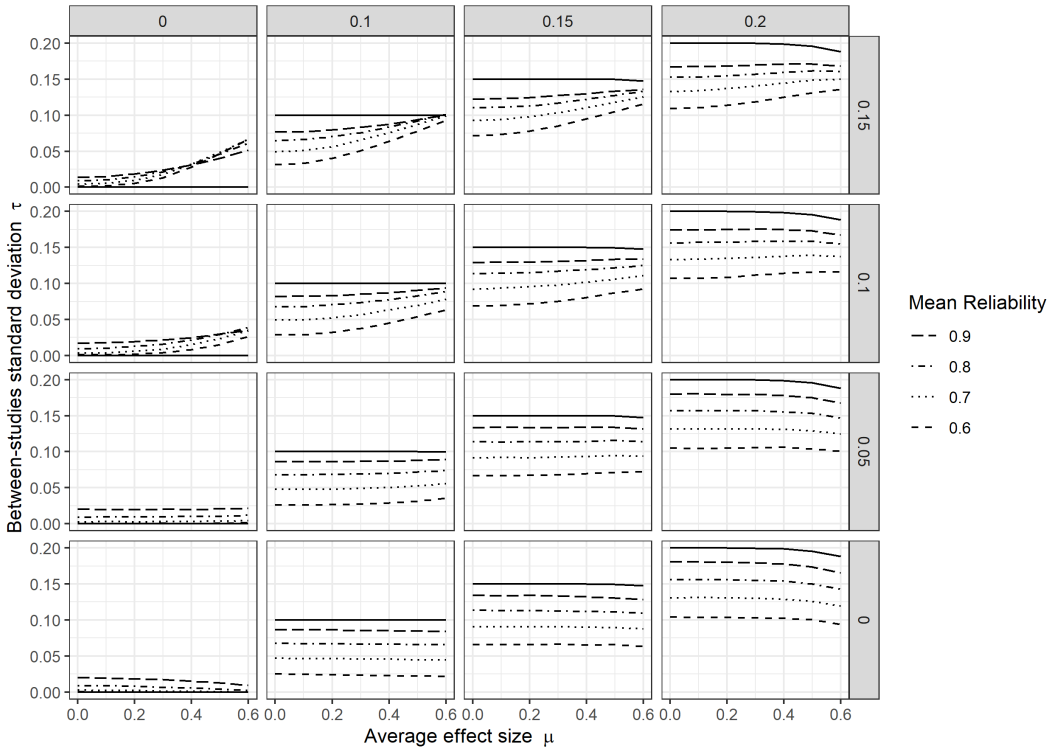

Figure C1. Inflation of effect sizes decreases with lower reliability SD leading to larger net negative bias in heterogeneity, especially for large effect sizes and moderate heterogeneity. The x-axis indicates average effect size and the y-axis estimated heterogeneity in standard deviations. Columns indicate nominal true between-studies standard deviation and rows indicate reliability standard deviation. Due to truncation in Pearson's r these values may differ from actual true heterogeneity standard deviation (black solid lines). Each dashed line correspond to an average reliability in primary studies. Code to reproduce figure: osf.io/8ygwj.

Lower number of meta-analyzed studies leads to a larger inflation of heterogeneity at low levels of true heterogeneity (Figure C2). In Figure C2, columns are again nominal true heterogeneity, whereas rows in Figure C2 are the number of studies in each meta-analysis. Generally, we would expect a lower number of studies to lead to larger sampling variance of heterogeneity estimates, hence a larger proportion of estimates truncated at zero, and consequently more inflated average heterogeneity estimates when number of studies is low (e.g., first row Figure C2) and heterogeneity is small or absent (first column Figure C2). This is also what we see in Figure C2. In addition, the number of studies mostly affects heterogeneity estimates when average effect size is small.

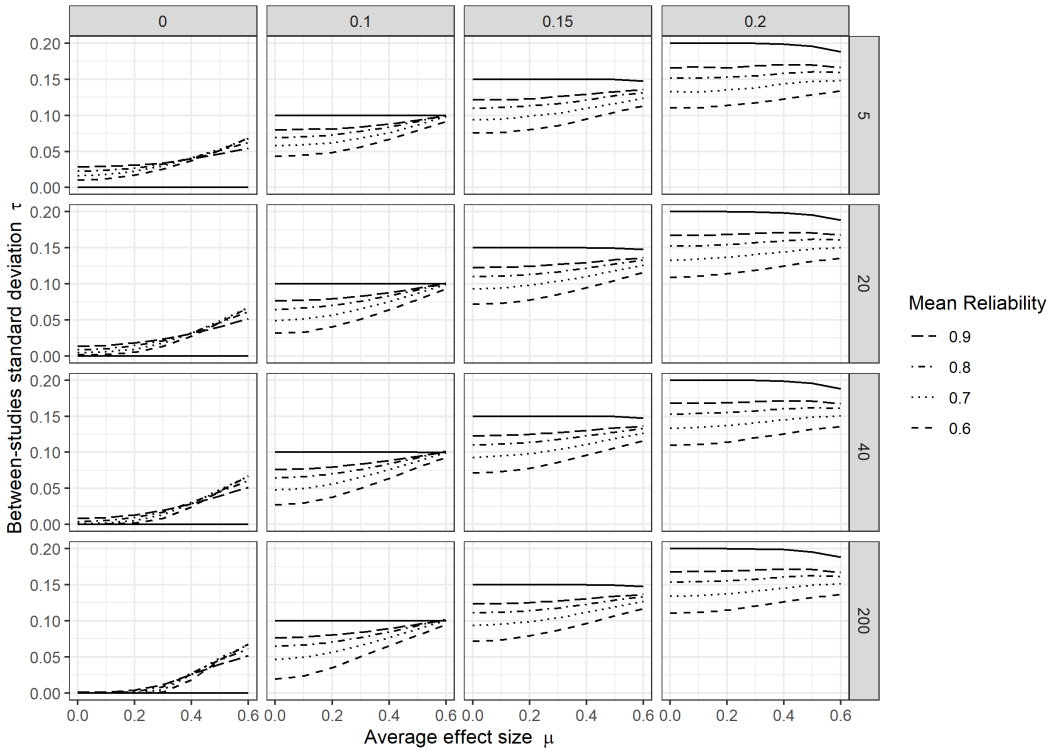

Figure C2. Expected inflation of effect sizes decreases with larger number of meta-analyzed when heterogeneity is absent or small and average effect size is small. The x-axis indicates average effect size and the y-axis estimated heterogeneity in standard deviations. Columns indicate nominal true between-studies standard deviation and rows indicate number of meta-analyzed studies. Due to truncation in Pearson's r these values may differ from actual true heterogeneity standard deviation (black solid lines). Each dashed line correspond to an average reliability in primary studies. Code to reproduce figure: osf.io/8ygwj.