Supplement: New Statistical Metrics for Multisite Replication Projects

Contents

1. Theory and statistical details

1.1. Agreement in “statistical significance”

Suppose the original study tested the null hypothesis $H_0 : \mu = \mu_N$, where μ is an unknown population parameter. Consider for now a single replication study, and let $\widehat{\theta}_{\text{orig}}$ and $\widehat{\theta}_{\text{rep}}$ be estimates of μ from the original and replication study, respectively. Assume that under both the null and alternative hypotheses, $\widehat{\theta}_{\text{orig}}$ and $\widehat{\theta}_{\text{rep}}$ are approximately and independently normal with a common mean but potentially different standard errors:

$$\begin{aligned}
\widehat{\theta}_{\text{orig}} \sim N\left(\mu, SE^2_{\text{orig}}\right) \;\amalg\; \widehat{\theta}_{\text{rep}} \sim N\left(\mu, SE^2_{\text{rep}}\right) \\
\Rightarrow \widehat{\theta}_{\text{rep}} - \widehat{\theta}_{\text{orig}} \sim N\left(0, SE^2_{\text{rep}} + SE^2_{\text{orig}}\right)
\end{aligned} \tag{1.1}$$

where $\amalg$ denotes statistical independence. (Critically, this setup does not allow for heterogeneity in that it assumes that the replication and original studies measure the same population effect, μ. Section 1.4 demonstrates the impact of this stringent assumption.)

Considering first the case in which the original estimate is above the null (i.e., $\widehat{\theta}_{\text{orig}} - \mu_N > 0$), we can derive the probability of a “significant” replication estimate that is also above the null ($\widehat{\theta}_{\text{rep}} - \mu_N > 0$) given the original estimate and its true standard error ($\widehat{\theta}_{\text{orig}}$ and SE_{orig}). Let $c_\alpha = \Phi^{-1}(1 - \alpha/2)$ be the critical value of the normalized test statistic (e.g., 1.96 for $\alpha = 0.05$). Standardize $\widehat{\theta}_{\text{rep}}$ to construct the usual standard-normal test statistic and express the desired probability as:

$$\begin{aligned}
P\left(\frac{\widehat{\theta}_{\text{rep}} - \mu_N}{SE_{\text{rep}}} > c_\alpha \;\middle|\; \widehat{\theta}_{\text{orig}}, SE_{\text{orig}}\right) &= P\left(\widehat{\theta}_{\text{rep}} > c_\alpha SE_{\text{rep}} + \mu_N \;\middle|\; \widehat{\theta}_{\text{orig}}, SE_{\text{orig}}\right) \\
&= P\left(\underbrace{\frac{\widehat{\theta}_{\text{rep}} - \widehat{\theta}_{\text{orig}}}{\sqrt{SE^2_{\text{rep}} + SE^2_{\text{orig}}}}}_{N(0,1)} > \frac{c_\alpha SE_{\text{rep}} + \mu_N - \widehat{\theta}_{\text{orig}}}{\sqrt{SE^2_{\text{rep}} + SE^2_{\text{orig}}}}\right) \\
&= 1 - \Phi\left(\frac{c_\alpha SE_{\text{rep}} + \mu_N - \widehat{\theta}_{\text{orig}}}{\sqrt{SE^2_{\text{rep}} + SE^2_{\text{orig}}}}\right) \qquad (1.2) \\
&\approx 1 - \Phi\left(\frac{c_\alpha \widehat{SE}_{\text{rep}} + \mu_N - \widehat{\theta}_{\text{orig}}}{\sqrt{\widehat{SE}^2_{\text{rep}} + \widehat{SE}^2_{\text{orig}}}}\right) \qquad (1.3)
\end{aligned}$$

where the second equality follows from Eq. 1.1 and the final approximation follows by substituting estimated standard errors for their true counterparts. Similarly, considering the case in which the original estimate is below the null ($\widehat{\theta}_{\text{orig}} - \mu_N < 0$), the probability of a "significant" replication estimate that is also below the null is:

$$\begin{aligned}
P\left(\frac{\widehat{\theta}_{\text{rep}} - \mu_N}{SE_{\text{rep}}} < -c_\alpha \,\middle|\, \widehat{\theta}_{\text{orig}}, SE_{\text{orig}}\right) &= P\left(\widehat{\theta}_{\text{rep}} < -c_\alpha SE_{\text{rep}} + \mu_N \,\middle|\, \widehat{\theta}_{\text{orig}}, SE_{\text{orig}}\right) \\
&= P\left(\underbrace{\frac{\widehat{\theta}_{\text{rep}} - \widehat{\theta}_{\text{orig}}}{\sqrt{SE^2_{\text{rep}} + SE^2_{\text{orig}}}}}_{N(0,1)} < \frac{-c_\alpha SE_{\text{rep}} + \mu_N - \widehat{\theta}_{\text{orig}}}{\sqrt{SE^2_{\text{rep}} + SE^2_{\text{orig}}}}\right) \\
&\approx \Phi\left(\frac{-c_\alpha \widehat{SE}_{\text{rep}} + \mu_N - \widehat{\theta}_{\text{orig}}}{\sqrt{\widehat{SE}^2_{\text{rep}} + \widehat{SE}^2_{\text{orig}}}}\right) \qquad (1.4)
\end{aligned}$$

When there are multiple replications (in either a many-to-one or one-to-one design), one can simply apply either Equation 1.3 or 1.4 to each replication study depending on the sign of the relevant original estimate.

1.2. Estimating the population effect distribution for P_{orig}

We assume that the replication studies estimate (with statistical error) potentially different population effect sizes that follow a normal distribution. The distribution of population effects is distinct from the *observed* distribution of replication estimates; the latter is more variable due to statistical error in the replication studies. The proposed analyses therefore begin by using the replication studies to estimate the mean and variance of the distribution of population effects using, for example, one of two straightforward modeling approaches. Both approaches begin with shared assumptions. Let $\widehat{\theta}_{\text{rep,i}}$ denote the point estimate in the i^{th} replication such that $\widehat{\theta}_{\text{rep,i}} = \mu + \gamma_i + \epsilon_i$, where $\gamma_i \sim N(0, \tau^2)$ denotes deviations of site-specific population effects from the grand mean (μ) and $\epsilon_i \sim N\left(0, SE^2_{\text{rep,i}}\right)$ denotes statistical error. Assume that γ_i and ϵ_i are independent. That is, the population effect in replication site i is $\mu + \gamma_i$, which is normal with mean μ and variance τ^2. Its estimate, incorporating additional error due to ϵ_i, is $\widehat{\theta}_{\text{rep,i}}$ and is marginally normal with mean μ and variance $\tau^2 + SE^2_{\text{rep,i}}$.

To estimate μ and τ^2, one option is compute an point estimate within each site (for example, using the same model as in the original study) and then to conduct a random-effects meta-analysis on these site-level summary measures. Such analyses are already commonplace in many-to-one designs. One can then use the meta-analytic pooled estimate as $\widehat{\mu}$ and the heterogeneity estimate of the variance of the population effects as $\widehat{\tau}^2$. A second option, which avoids aggregating data by site prior to analysis, is to fit a mixed model to the observation-level data with independent, identically normal random intercepts and slopes by site; this is a form of individual participant data meta-analysis (Stewart et al., 2012). For example, suppose the original study used ordinary least squares regression to estimate the effect (β_1) of a binary experimental manipulation X on a continuous dependent variable Y with the usual specification $Y_j = \beta_0 + \beta_1 X_j + \epsilon_j$ for subjects $j = 1, \cdots, N_{\text{orig}}$ and with the error terms ϵ_j assumed independent and identically ("iid") normal. Then, for the replications, one possible mixed model specification is:

$$\begin{aligned}
&Y_{ij} = \alpha_0 + \zeta_{0i} + \alpha_1 X_{ij} + \zeta_{1i} X_{ij} + \epsilon^*_{ij} \\
&\zeta_{0i} \sim_{iid} N\left(0, \sigma^2_{\zeta_0}\right) \amalg \zeta_{1i} \sim_{iid} N\left(0, \sigma^2_{\zeta_1}\right) \amalg \epsilon^*_{ij} \sim_{iid} N\left(0, \sigma^2_{\epsilon^*}\right)
\end{aligned}$$

where i indexes sites. Then, we can estimate μ (the average population effect of X across all sites) using the usual maximum likelihood or restricted maximum likelihood estimate, $\widehat{\alpha}_1$. We can estimate τ^2 (the variance of the population effects of X across all sites) with $\widehat{\sigma}^2_{\zeta_1}$. Depending on the experimental design, of course, a different mixed model specification may be warranted (for example, with additional random terms by subject) as long as it retains the normal assumption on the effect sizes across sites and yields unbiased, approximately normally distributed, and approximately independent estimates of μ and τ^2. Specifications that do not pre-aggregate data within sites may often be more statistically efficient than the meta-analytic approach, but the meta-analysis method may sometimes provide more flexibility because it models the effect sizes rather than the dependent variable itself. Lastly, a third possible modeling approach could simply ignore site and fit the same analysis model as was used in the original study, but we do not recommend this approach because clustering within sites will likely violate statistical assumptions regarding conditionally independent residuals, such specifications preclude estimation of τ^2, and they can lead to bias due to Simpson's Paradox (Rücker & Schumacher, 2008).

1.3. Derivation of P_{orig}

Given the estimates $\widehat{\mu}$ and $\widehat{\tau}^2$ from the above development, we can derive the probability that, if the original study and replications come from the same, potentially heterogeneous distribution of population effects, the original study's estimate would be as extreme or more extreme than its actual estimate. As above, let $\widehat{\theta}_{\text{orig}}$ be the point estimate in the original study and SE_{orig} its standard error. Letting $\widehat{\theta}^*$ be a random variable denoting the effect estimate in an arbitrary study with the same standard error as the original, we first consider the distribution of $\widehat{\theta}^* - \widehat{\mu}$. Assume that $\widehat{\mu} \sim N\left(\mu, SE^2_{\widehat{\mu}}\right)$; that is, the estimate is approximately unbiased and normal. (This holds for both the meta-analysis and the mixed model approaches above under standard assumptions.) Since $\widehat{\theta}^* \sim N\left(\mu, \tau^2 + SE^2_{\text{orig}}\right)$ independently of $\widehat{\mu}$, we can derive the first proposed metric as follows:

$$
\begin{aligned}
\widehat{\theta}^* - \widehat{\mu} &\sim N\left(0, \tau^2 + SE^2_{\text{orig}} + SE^2_{\widehat{\mu}}\right) \\
P\left(|\widehat{\theta}^* - \widehat{\mu}| \geq |\widehat{\theta}_{\text{orig}} - \widehat{\mu}|\right) &= P\left(\widehat{\theta}^* - \widehat{\mu} \geq |\widehat{\theta}_{\text{orig}} - \widehat{\mu}|\right) + P\left(\widehat{\theta}^* - \widehat{\mu} \leq -|\widehat{\theta}_{\text{orig}} - \widehat{\mu}|\right) \\
&= P\left(\underbrace{\frac{\widehat{\theta}^* - \widehat{\mu}}{\sqrt{\tau^2 + SE^2_{\text{orig}} + SE^2_{\widehat{\mu}}}}}_{\sim N(0,1)} \geq \frac{|\widehat{\theta}_{\text{orig}} - \widehat{\mu}|}{\sqrt{\tau^2 + SE^2_{\text{orig}} + SE^2_{\widehat{\mu}}}}\right) + \\
&\quad P\left(\underbrace{\frac{\widehat{\theta}^* - \widehat{\mu}}{\sqrt{\tau^2 + SE^2_{\text{orig}} + SE^2_{\widehat{\mu}}}}}_{\sim N(0,1)} \leq \frac{-|\widehat{\theta}_{\text{orig}} - \widehat{\mu}|}{\sqrt{\tau^2 + SE^2_{\text{orig}} + SE^2_{\widehat{\mu}}}}\right) \\
&= 1 - \Phi\left(\frac{|\widehat{\theta}_{\text{orig}} - \widehat{\mu}|}{\sqrt{\tau^2 + SE^2_{\text{orig}} + SE^2_{\widehat{\mu}}}}\right) + \Phi\left(\frac{-|\widehat{\theta}_{\text{orig}} - \widehat{\mu}|}{\sqrt{\tau^2 + SE^2_{\text{orig}} + SE^2_{\widehat{\mu}}}}\right) \\
&= 2 \times \left(1 - \Phi\left(\frac{|\widehat{\theta}_{\text{orig}} - \widehat{\mu}|}{\sqrt{\tau^2 + SE^2_{\text{orig}} + SE^2_{\widehat{\mu}}}}\right)\right) \qquad (1.5)
\end{aligned}
$$

We arrive at the approximation in the main text (i.e., P_{orig}) by substituting the estimates $\widehat{SE}_{\text{orig}}$ and $\widehat{SE}_{\widehat{\mu}}$ for the true parameters.

We now show that P_{orig} subsumes Patil et al. (2016)'s prediction interval in the sense that if there is a single replication study, if we assume there is no heterogeneity, and if we dichotomize P_{orig} at $\alpha = 0.05$, we mathematically recover the prediction interval. In Equation 1.5, set the left-hand side equal to 0.05 (for a 95% prediction interval) and $\tau^2 = 0$ (for no heterogeneity). Let $\theta^*_{0.05}$ be the hypothetical value for the replication point estimate that marks the lower or upper boundary of the 95% prediction interval. Since the prediction interval concerns a single replication study, set $\widehat{\mu} = \theta^*_{0.05}$ and $\widehat{SE}^2_{\widehat{\mu}} = \widehat{SE}^2_{\text{rep}}$. Thus, we can solve for the boundary values of the prediction interval, i.e., the pair of hypothetical replication estimates that would be sufficiently extreme to make the probability on the left-hand side equal to 0.05:

$$\begin{aligned}
0.05 &= 2 \times \left(1 - \Phi\left(\frac{|\widehat{\theta}_{\text{orig}} - \theta^*_{0.05}|}{\sqrt{\widehat{SE}^2_{\text{orig}} + \widehat{SE}^2_{\text{rep}}}}\right)\right) \\
|\widehat{\theta}_{\text{orig}} - \theta^*_{0.05}| &= \Phi^{-1}(0.975)\sqrt{\widehat{SE}^2_{\text{orig}} + \widehat{SE}^2_{\text{rep}}} \\
\widehat{\theta}_{\text{orig}} - \theta^*_{0.05} &= \pm\Phi^{-1}(0.975)\sqrt{\widehat{SE}^2_{\text{orig}} + \widehat{SE}^2_{\text{rep}}} \\
\theta^*_{0.05} &= \widehat{\theta}_{\text{orig}} \pm \Phi^{-1}(0.975)\sqrt{\widehat{SE}^2_{\text{orig}} + \widehat{SE}^2_{\text{rep}}}
\end{aligned}$$

which is exactly Patil et al. (2016)'s prediction interval.

1.4. Impact of ignoring heterogeneity in existing metrics

We now show that ignoring heterogeneity when estimating the expected "significance agreement" can underestimate or overestimate consistency when there truly is heterogeneity. We begin by generalizing Equation 1.1 (which ignores heterogeneity) to accommodate heterogeneity via the same framework developed in Section 1.2, assuming normally distributed population effects:

$$\begin{aligned}
\widehat{\theta}_{\text{orig}} \sim N\left(\mu, \tau^2 + SE^2_{\text{orig}}\right) \ \amalg \ \widehat{\theta}_{\text{rep}} \sim N\left(\mu, \tau^2 + SE^2_{\text{rep}}\right) \\
\Rightarrow \widehat{\theta}_{\text{rep}} - \widehat{\theta}_{\text{orig}} \sim N\left(0, 2\tau^2 + SE^2_{\text{rep}} + SE^2_{\text{orig}}\right)
\end{aligned}$$

For an original estimate above the null, we can compute the probability of "significance agreement" allowing for heterogeneity as:

$$\begin{aligned}
P\left(\frac{\widehat{\theta}_{\text{rep}} - \mu_N}{SE_{\text{rep}}} > c_\alpha \,\middle|\, \widehat{\theta}_{\text{orig}}, SE_{\text{orig}}\right) &= P\left(\widehat{\theta}_{\text{rep}} > c_\alpha SE_{\text{rep}} + \mu_N \,\middle|\, \widehat{\theta}_{\text{orig}}, SE_{\text{orig}}\right) \\
&= P\left(\underbrace{\frac{\widehat{\theta}_{\text{rep}} - \widehat{\theta}_{\text{orig}}}{\sqrt{2\tau^2 + SE^2_{\text{rep}} + SE^2_{\text{orig}}}}}_{N(0,1)} > \frac{c_\alpha \widehat{SE}_{\text{rep}} + \mu_N - \widehat{\theta}_{\text{orig}}}{\sqrt{2\tau^2 + SE^2_{\text{rep}} + SE^2_{\text{orig}}}}\right) \\
&\approx 1 - \Phi\left(\frac{c_\alpha \widehat{SE}_{\text{rep}} + \mu_N - \widehat{\theta}_{\text{orig}}}{\sqrt{2\tau^2 + \widehat{SE}^2_{\text{rep}} + \widehat{SE}^2_{\text{orig}}}}\right)
\end{aligned}$$

The only difference between this expression and Equation 1.2 (which had assumed no heterogeneity) is the $2\tau^2$ term in the denominator. Since the presence of heterogeneity implies that $2\tau^2 > 0$, this probability exceeds that in Equation 1.2 when $c_\alpha \widehat{SE}_{\text{rep}} + \mu_N - \widehat{\theta}_{\text{orig}} > 0$ and otherwise is less than that in Equation 1.2. Thus, when $c_\alpha \widehat{SE}_{\text{rep}} + \mu_N - \widehat{\theta}_{\text{orig}} > 0$, analyses that ignore heterogeneity will underestimate consistency between the replication and the original study, and otherwise, they will overestimate consistency. The case in which the original estimate is below the null is symmetrical, so is omitted.

When there is heterogeneity, the prediction interval will be too narrow. Using the previous result showing equivalence of P_{orig} with the prediction interval when there is no heterogeneity, we can set $\tau^2 = 0$ in Equation 1.5 to yield the p-value counterpart to the prediction interval. Since Equation 1.5 is strictly increasing in τ^2, constraining $\tau^2 = 0$ in this expression yields a lower p-value than allowing $\tau^2 > 0$. Thus, if there is heterogeneity, the p-value counterpart to the prediction interval is an underestimate. By the duality of p-values and intervals, the prediction interval is therefore too narrow when $\tau^2 > 0$.

2. Methods for choosing an effect size threshold

Much existing work, spanning a variety of disciplinary perspectives, has discussed how to choose thresholds for meaningfully strong effect sizes. Crosby et al. (2003) provide an excellent review and examples of numerous methods in the context of health outcomes. In particular, they discuss a variety of "anchoring-based" methods in which an effect size threshold is chosen by relating the outcome measure to external benchmarks bearing immediate scientific or policy relevance. Within psychology, this approach may be particularly relevant for applied or intervention studies; for example, when investigating effects of educational interventions, a minimum effect size threshold could be determined in relation to differences in the outcome (academic achievement) between naturally-occurring subject groups (such as children attending low- versus high-performing schools or children of different ages; Hill et al. (2008)). Numerous other types of external "anchoring" criteria have also been used in the health outcomes literature (Crosby et al., 2003).

When the aggregate public impact of an outcome (such as juvenile delinquency) is the primary concern, investigators could draw upon the extensive literature on cost-effectiveness decision rules in selecting an effect size threshold. For example, much existing work has discussed or empirically quantified the cost threshold at which societies or individuals are willing to pay for a specific improvement in physical or mental health, such as an addition of one quality-adjusted life-year (e.g., Braithwaite et al. (2008); Eichler et al. (2004)). Such findings could be used to "convert" hypothetical statistical effect sizes for a given outcome to a concrete financial scale, such as dollars. A minimum effect size threshold could then be defined in relation to the utility, expressed in dollars, of the intervention or exposure of interest.

In contrast, in disciplines such as clinical psychology, the original study may investigate an effect in which individuals' subjective experience of distress or pain is the primary concern (instead of, or in addition to, aggregate public impact). In this case, it may be useful to set the threshold as the minimum effect size that is subjectively perceptible (Jaeschke et al. (1989); Lakens et al. (2018); Norman et al. (2003); Redelmeier et al. (1996)). A systematic review considered 62 studies that attempted to estimate such thresholds for a wide variety of health outcomes, for example by relating patients' subjective self-assessments to objective measurements of health condition severity (Norman et al., 2003). This review found that $d = 0.50$ was a surprisingly consistent minimally detectable effect size for health outcomes, perhaps reflecting fundamental mechanisms of human sensory discrimination or constraints on categorical discrimination due to working memory capacity. For ease of comparison to other statistical measures of effect size, the threshold $d = 0.50$ is approximately equivalent (under some distributional assumptions) to an odds ratio of 2.5 or to a risk ratio of 1.6 (Chinn (2000); VanderWeele (2017)). However, it is important to note that an intervention that has only small effects on the individual level, even ones that are not subjectively perceptible, may still have very substantial impacts on a population level; thus, as described above, much lower thresholds might often be considered.

While the above considerations may work well for applied or interventional psychology, many replication efforts to date have focused on classic experimental psychology, conducted using stylized tasks (such as a Stroop task or Monin & Miller (2001)'s hypothetical hiring task) in order to examine basic mechanisms of, for example, cognition or perception. Although some of the above considerations are harder to apply in these classic experimental contexts, external benchmarks could still be determined using effect sizes on similar experimental tasks, preferably those estimated by meta-analyses of existing literature. For example, a meta-analysis of the enormous literature on intergroup contact and prejudice had a pooled estimate of $r = -0.21$ among all study designs and $r = -0.33$ among experimental studies (Pettigrew & Tropp, 2008). We might treat experimental intergroup-contact interventions as a "gold standard" representing the effect sizes on prejudice that are achievable through purposefully designed interventions. In contrast, the proposed moral credentialing effect is not a designed intervention on prejudice but rather a specific, potentially more subtle, cognitive mechanism of prejudice. Thus, to select an effect size threshold for moral credentialing, we might somewhat reduce the magnitude of the gold-standard interventions to, for example, $|r| = 0.20$ or $|r| = 0.10$. (Additionally, the latter threshold is often considered a standard benchmark for a "small" effect size (Cohen, 1977).)

3. Extended simulation methods and results

3.1. Methods

Figure S1 shows population effects simulated from each of the four distributions for each value of τ^2.

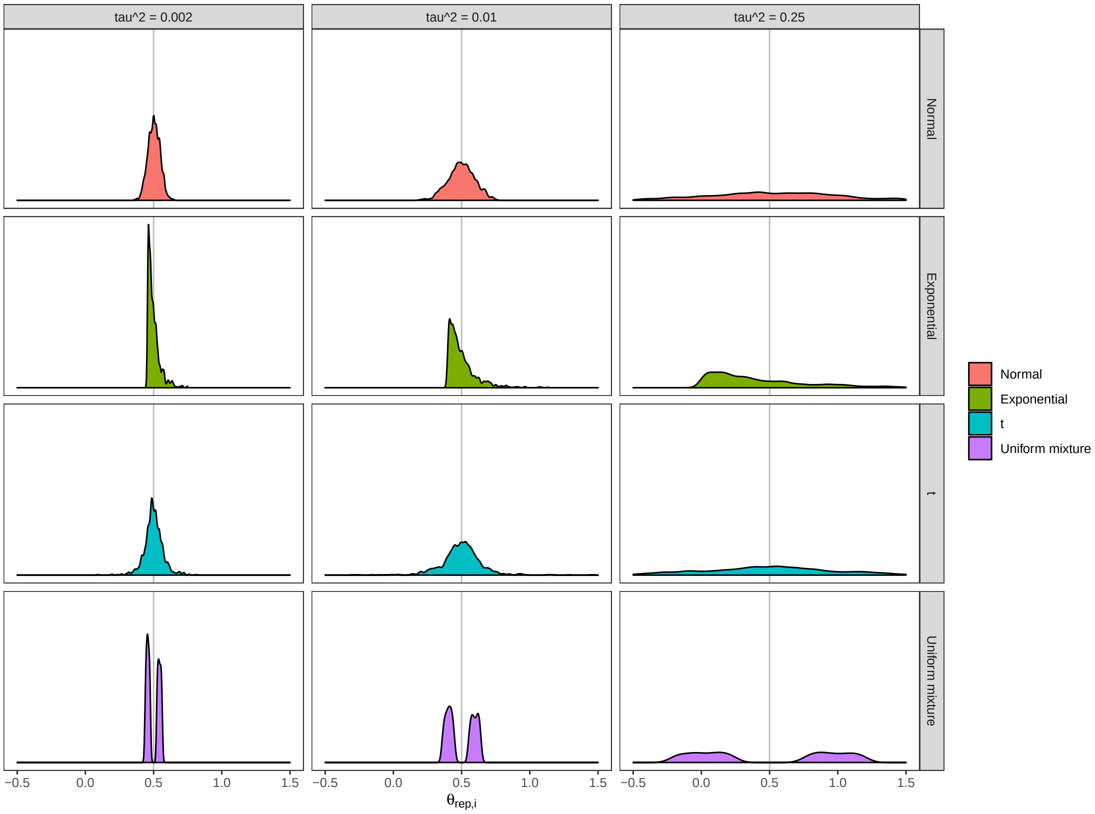

Figure S1: *Simulated population effect sizes from the four population effect distributions with varying heterogeneity*

3.2. Type I error and power of P_{orig}

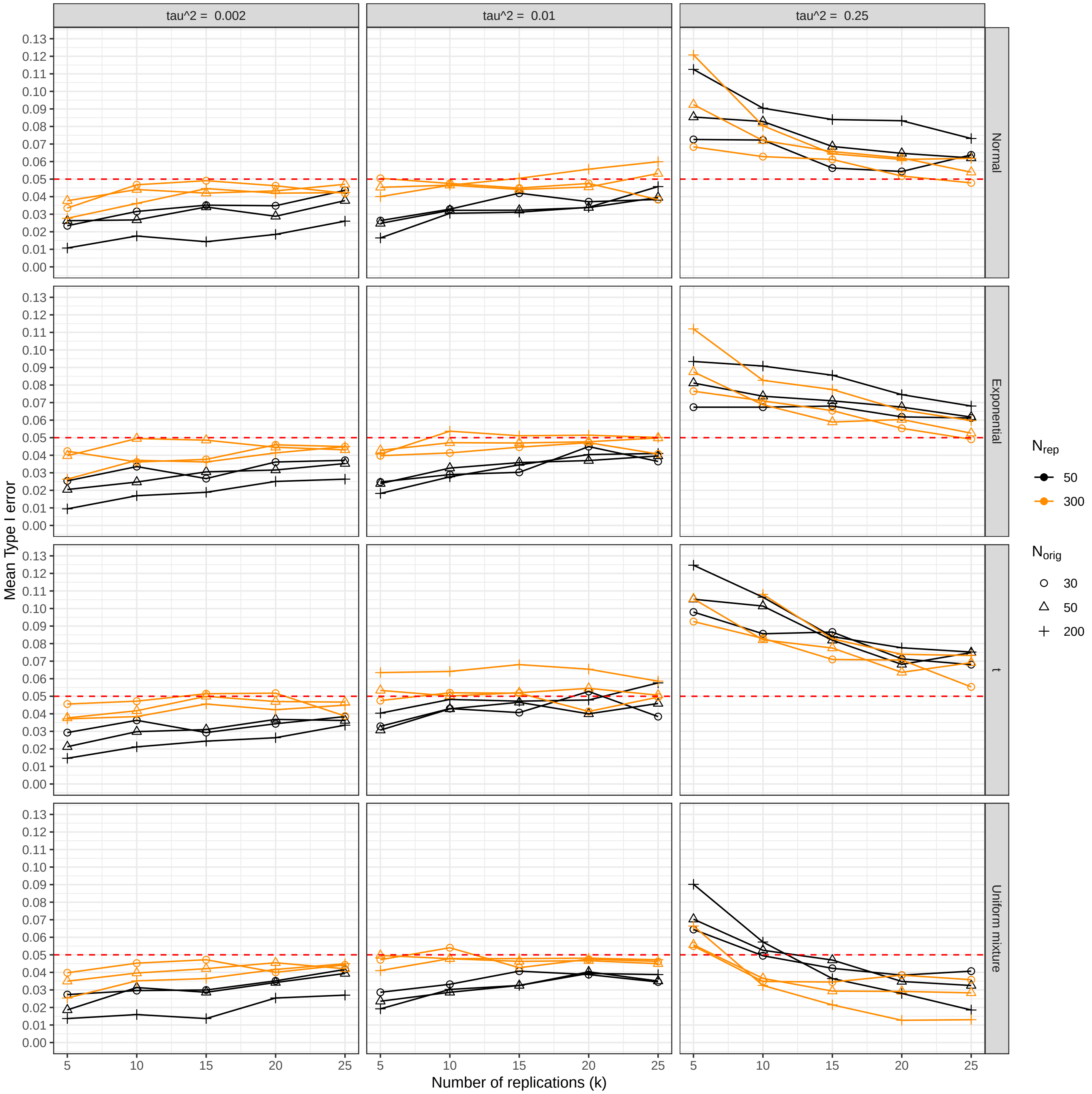

Figure S2: *Type I error of P_{orig} (scenarios with $\Delta = 0$)*

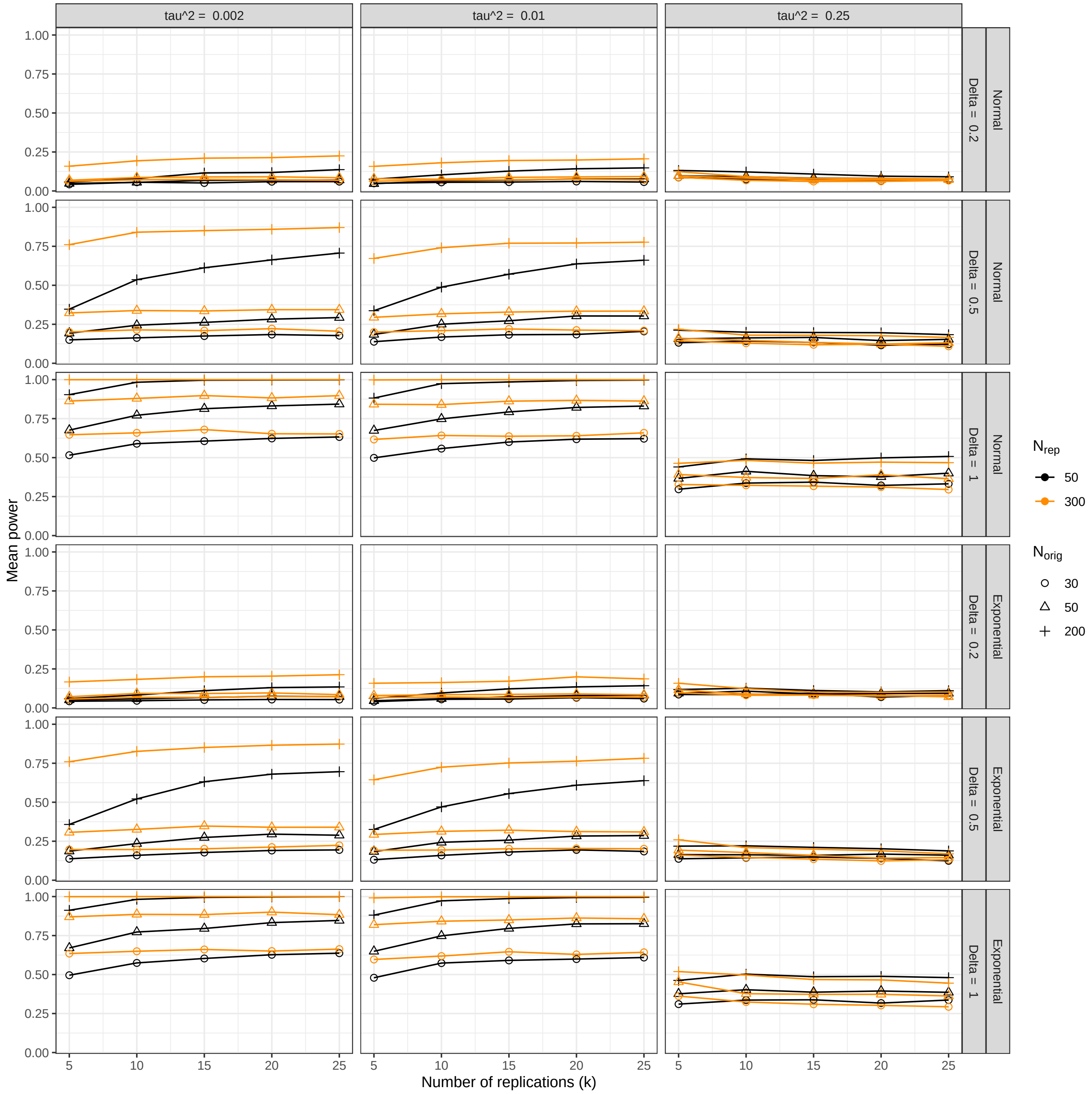

Figure S3: *Power of P_{orig} for normal and exponential distributions (scenarios with $\Delta > 0$)*

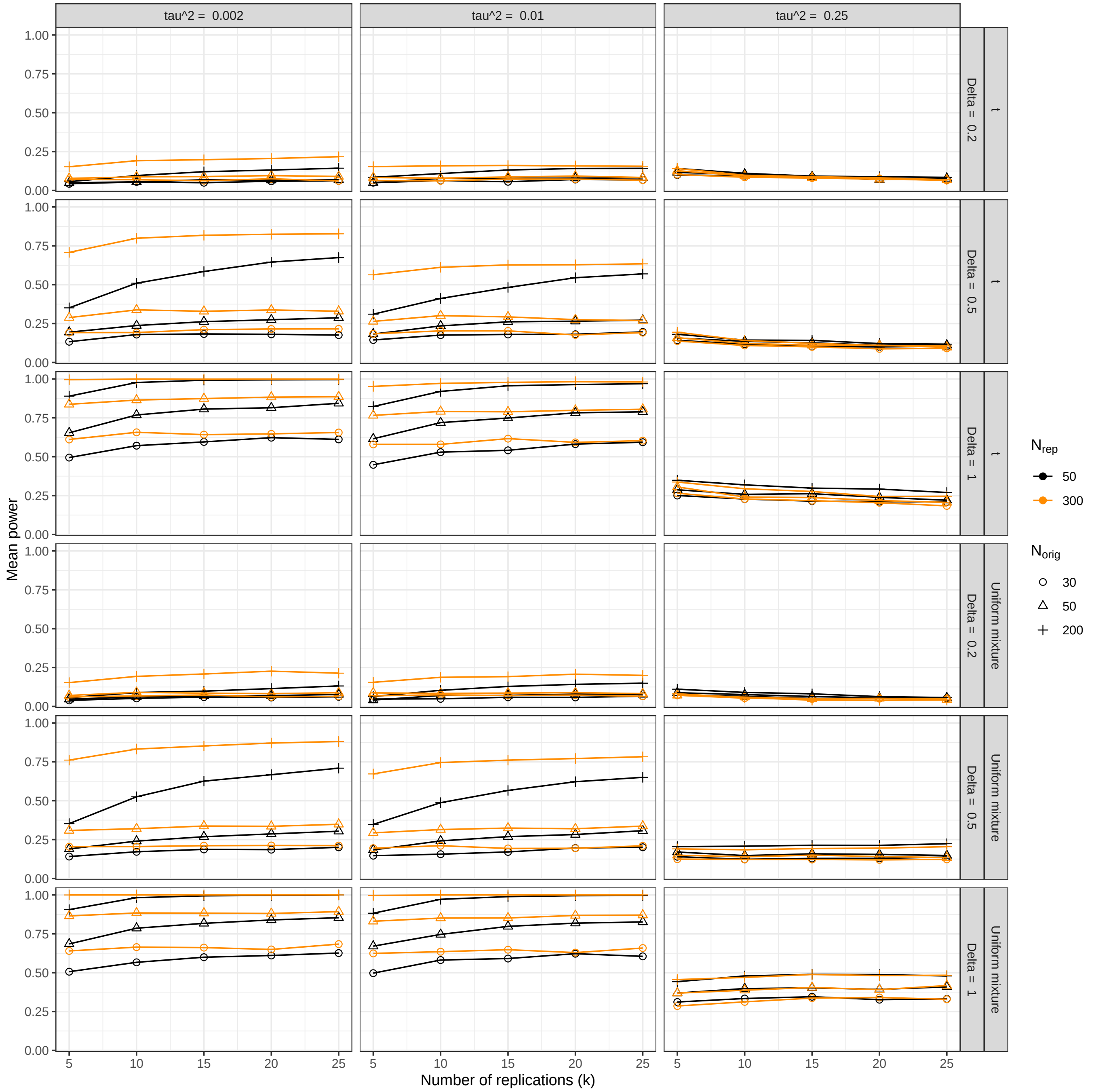

Figure S4: *Power of P_{orig} for t and uniform mixture distributions (scenarios with $\Delta > 0$)*

3.3. Bias and root mean square error of $\widehat{P}_{>q}$

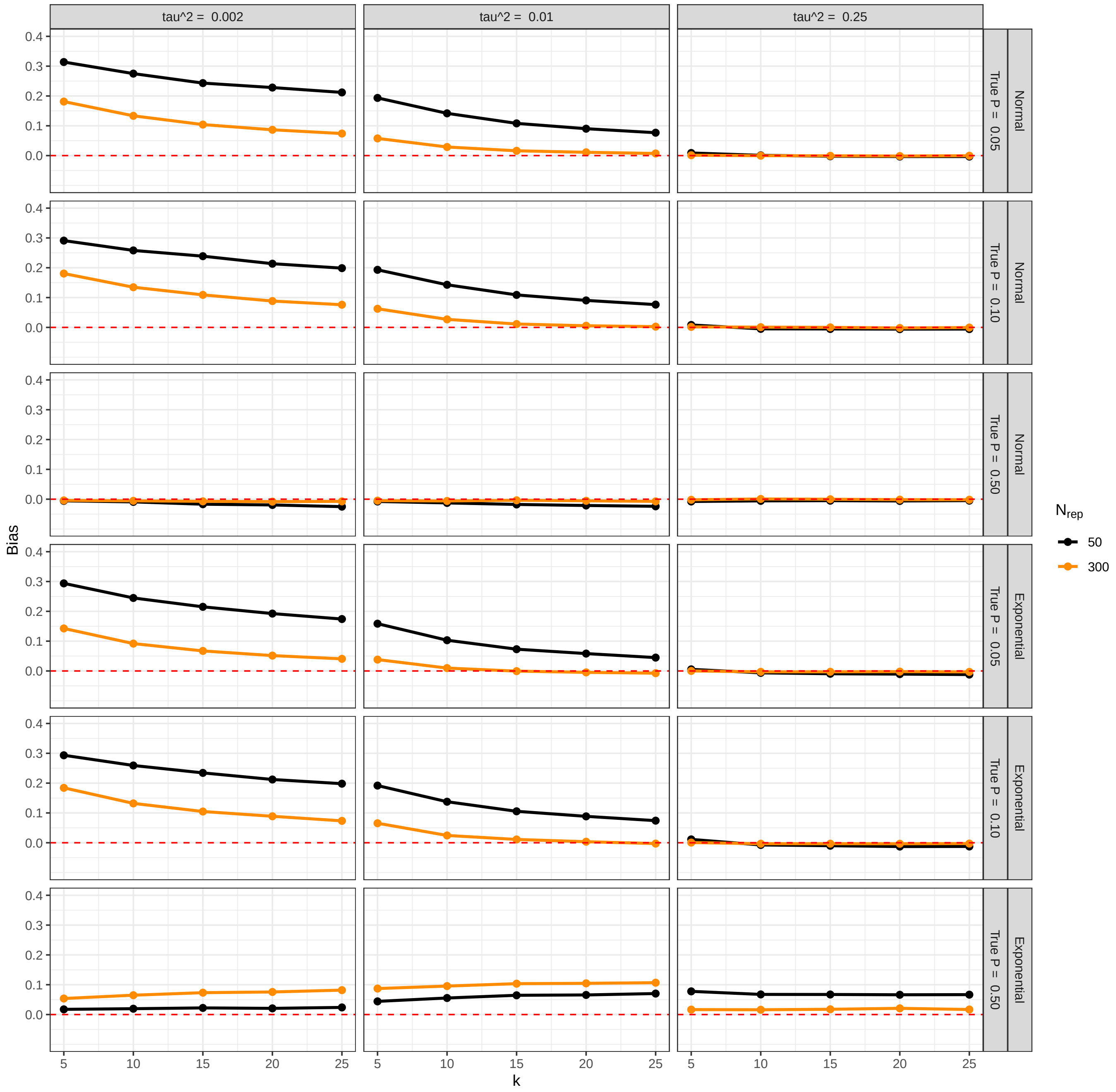

Figure S5: *Bias of $\widehat{P}_{>q}$ for normal and exponential distributions*

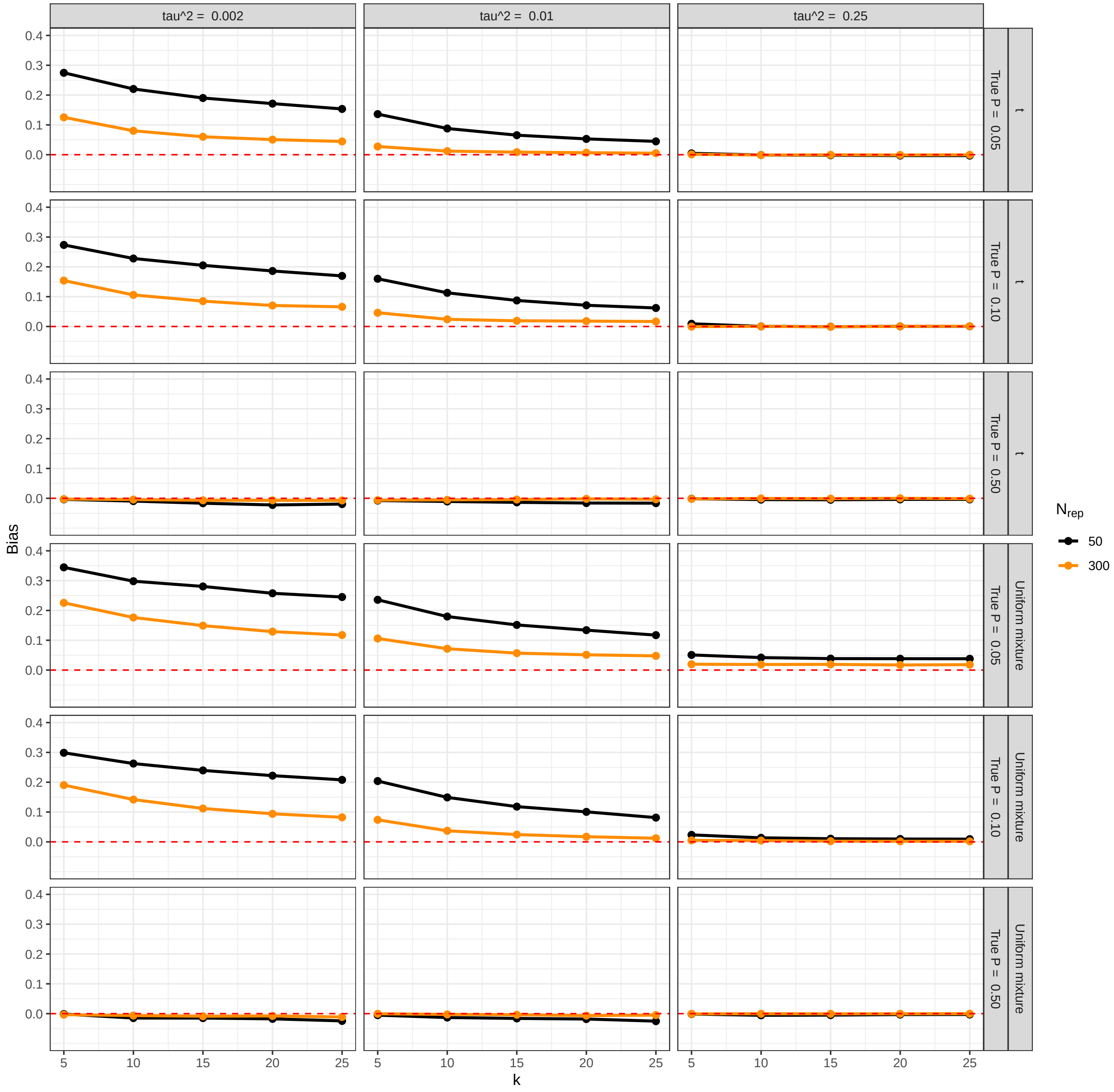

Figure S6: *Bias of $\widehat{P}_{>q}$ for t and uniform mixture distributions*

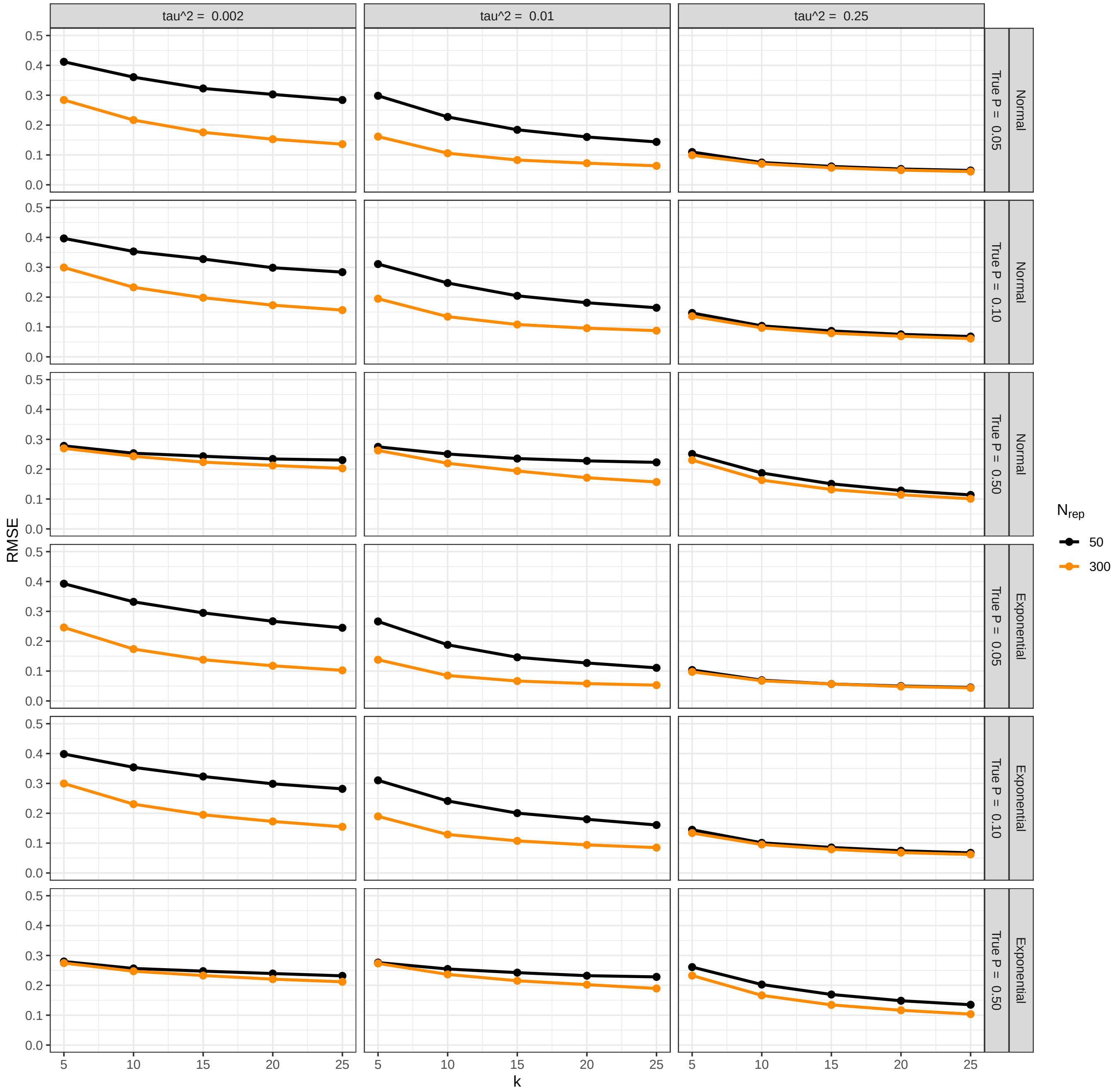

Figure S7: *RMSE of $\widehat{P}_{>q}$ for normal and exponential distributions*

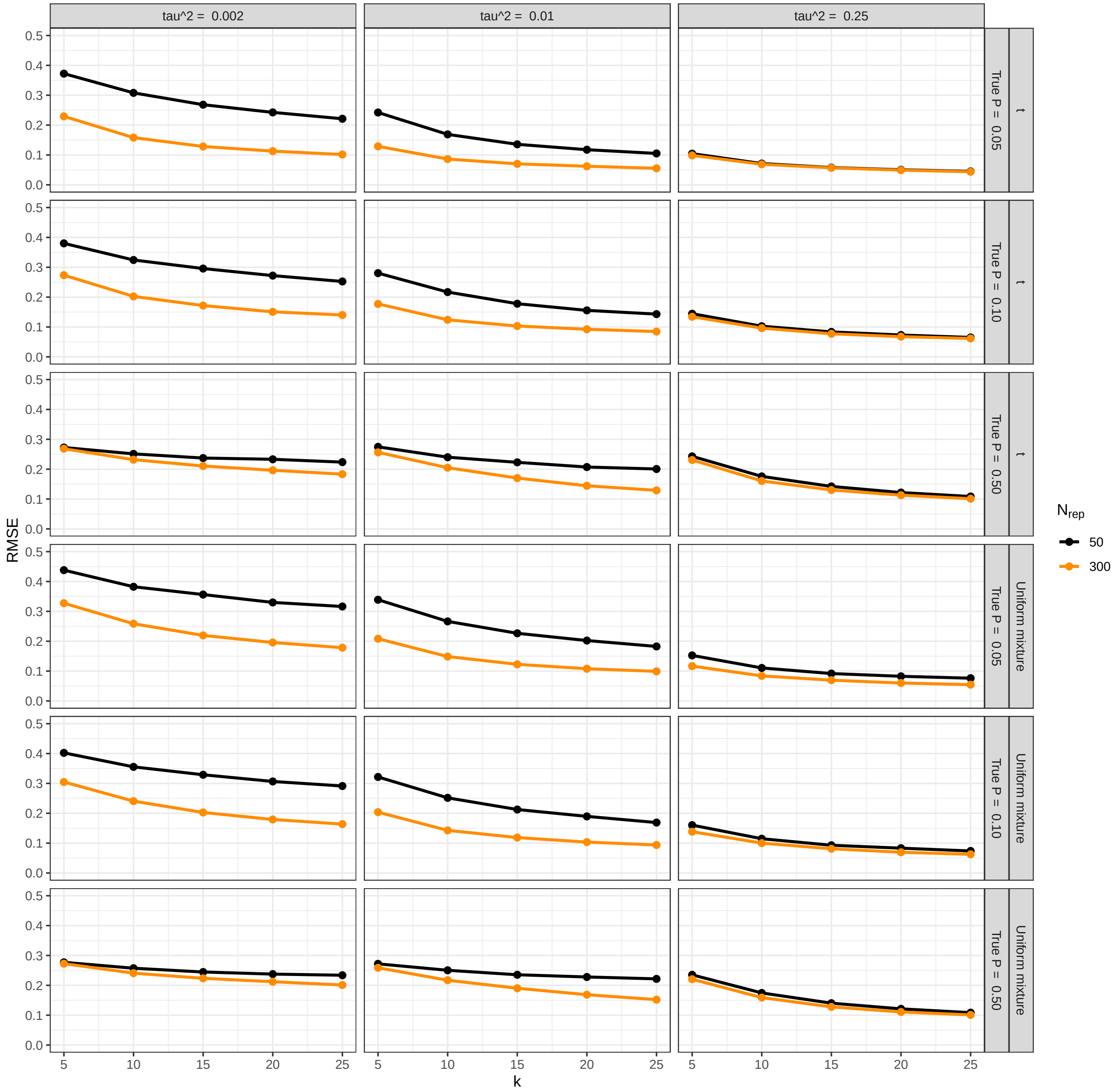

Figure S8: *RMSE of $\widehat{P}_{>q}$ for t and uniform mixture distributions*

4. Software

The R package `Replicate` (version 1.2.0) contains the following functions; details are available in the standard R documentation.

- `prob_signif_agree` computes the theoretical probability that a given replication study would agree in "statistical significance" and effect direction with the original study, if the population effect were indeed the same in the two studies.

- pred_int computes prediction interval limits and indicators for whether each replication estimate is within its corresponding prediction interval.
- p_orig computes P_{orig}, i.e., the probability of observing an original estimate as extreme as that actually observed (compared to the replication studies) if the original were indeed consistent with the estimated distribution of the replication studies.

The R package MetaUtility (version 2.1.0) contains the following function, again with details in the standard R documentation.

- prop_stronger estimates $\widehat{P}_{>0}$, $\widehat{P}_{>q}$, or $\widehat{P}_{<q^*}$, i.e., the probability of a population effect above or below a user-specified threshold for a meaningfully strong effect size using estimates of the population effect distribution (based on the replications).

References

Braithwaite, R. S., Meltzer, D. O., King Jr, J. T., Leslie, D., & Roberts, M. S. (2008). What does the value of modern medicine say about the $50,000 per quality-adjusted life-year decision rule? *Medical Care*, *46*(4), 349–356.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*(22), 3127–3131.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of clinical epidemiology*, *56*(5), 395–407.

Eichler, H.-G., Kong, S. X., Gerth, W. C., Mavros, P., & Jönsson, B. (2004). Use of cost-effectiveness analysis in health-care resource allocation decision-making: how are cost-effectiveness thresholds expected to emerge? *Value in Health*, *7*(5), 518–528.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, *10*(4), 407–415.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*(1), 33.

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical Care*, *41*(5), 582–592.

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, *11*(4), 539–544.

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? meta-analytic tests of three mediators. *European Journal of Social Psychology*, *38*(6), 922–934.

Redelmeier, D. A., Guyatt, G. H., & Goldstein, R. S. (1996). Assessing the minimal important difference in symptoms: a comparison of two techniques. *Journal of Clinical Epidemiology*, *49*(11), 1215–1219.

Rücker, G., & Schumacher, M. (2008). Simpson's Paradox visualized: the example of the rosiglitazone meta-analysis. *BMC Medical Research Methodology*, *8*(1), 34.

Stewart, G. B., Altman, D. G., Askie, L. M., Duley, L., Simmonds, M. C., & Stewart, L. A. (2012). Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. *PloS One*, *7*(10), e46042.

VanderWeele, T. J. (2017). On a square-root transformation of the odds ratio for a common outcome. *Epidemiology*, *28*(6), e58–e60.