

**Supplementary material: Quantifying replicability and consistency in
systematic reviews**

Iman Jaljuli

Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, Israel.

email: jaljuli.iman@gmail.com

and

Yoav Benjamini

Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, Israel.

email: ybenja@gmail.com

and

Liat Shenhav

Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA.

email: liashenhav@gmail.com

and

Orestis A. Panagiotou

Department of Health Services, Policy & Practice, Brown University, USA.

email: orestis_panagiotou@brown.edu

and

This paper has been submitted for consideration for publication in *Biometrics*

Ruth Heller

Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, Israel.

email: ruheller@gmail.com

1. Simulations

1.1 Simulation settings

For study $i \in \{1, \dots, n\}$, the estimated effect size, $\hat{\theta}_i$, is sampled from the normal distribution with mean θ_i and standard deviation $SE_i = \sqrt{1/n_{Ci} + 1/n_{Ti}}$, where n_{Ci} and n_{Ti} are the control and treatment group sizes, respectively. We examined a wide range of values for $(\theta_1, \dots, \theta_n)$, n , and $\{(n_{Ci}, n_{Ti}) : i = 1, \dots, n\}$.

In the paper, we displayed results for $n = 8$, with unequal group sizes as follows: $\{22, 210, 26, 192, 60, 38, 53, 15\}$ for the control groups and $\{22, 121, 24, 187, 31, 53, 49, 16\}$ for the treatment groups (these values are similar to those in the example detailed in Figure 8). Simulations for other variations of $n = 4, 8, 20$ with unequal samples sizes or equal ($\{n_{Ci} = n_{Ti} = 25 \forall i = 1, \dots, n\}$) are shown in figure ??

[Figure 1 about here.]

[Figure 2 about here.]

Figure ?? shows results for the random effects settings in figure 4 in the body of the paper. Here we show for $N = 4, 8, 20$ with unequal samples sizes (like in fig. ??), or with equal group sizes ($\{n_{Ci} = n_{Ti} = 25 \forall i = 1, \dots, n\}$).

[Figure 3 about here.]

2. Replicability-analysis: Assuming common effect

Proposition 6.1. *Let $Z = \hat{\theta}/\widehat{SE}$ be the fixed-effects meta-analysis test statistic, with meta-analysis p-value $p = 2 \min(p^R, p^L)$, where $p^L = \Phi(Z)$ and $p^R = 1 - \Phi(Z)$. Let $r(u), r^L(u), r^R(u)$ be the u out of N r -values for $u \in \{2, \dots, n\}$, as defined in this section. Then $p < r(u)$, and moreover, if $Z < 0$, $p^L < r^L(u)$, and if $Z > 0$, $p^R < r^R(u)$.*

Proof:

Let Z_v , $\hat{\theta}_v$, and SE_v be the fixed-effect meta-analysis test statistic, estimated effect, and SE, respectively, for the intersection hypotheses indexed by $v \in \Pi(n - u + 1)$. Since $\sum_{v \in \Pi(n-u+1)} \frac{z_v}{SE_v} = \binom{n-1}{n-u} \sum_{i=1}^n \frac{\hat{\theta}_i}{SE_i}$, the meta-analysis test statistic can be expressed in terms of (z_v, SE_v) , $v \in \Pi(n - u + 1)$:

$$Z = \frac{1}{\binom{n-1}{n-u}} \sum_{v \in \Pi(n-u+1)} \frac{z_v}{SE_v} SE \quad (1)$$

Let $v^* = \arg \max_{v \in \Pi(n-u+1)} Z_v$. By definition, $r^L = \Phi(Z_{v^*})$. We shall show that if $Z < 0$, $p^L < r^L$ and $\min(p^L, p^R) < \min(r^L, r^R)$. Clearly, $p^L < 0.5$ since $Z < 0$. Therefore, the result follows by showing that $p^L < r^L$ and $r^R > 0.5$.

We start by showing that $p^L < r^L$. If $Z_{v^*} > 0$, then by definition $r^L > 0.5$ and therefore it follows that $p^L < r^L$. If $Z_{v^*} < 0$ then

$$Z \leq \frac{Z_{v^*}}{\binom{n-1}{n-u}} \sum_{v \in \Pi(n-u+1)} \frac{SE}{SE_v} \leq \frac{Z_{v^*}}{\binom{n-1}{n-u}} \sum_{v \in \Pi(n-u+1)} \frac{SE^2}{SE_v^2} = Z_{v^*},$$

where the first inequality follows from (??) and the definition of v^* , the second inequality follows since $SE/SE_v < 1$ for all $v \in \Pi(n - u + 1)$, and the last equality follows since

$$\sum_{v \in \Pi(n-u+1)} \frac{1}{SE_v^2} = \binom{n-1}{n-u} \sum_{i=1}^n \frac{1}{SE_i^2} = \binom{n-1}{n-u} \frac{1}{SE^2}.$$

Since $Z \leq Z_{v^*}$ it thus follows that $p^L < r^L$.

Next, we show that $r^R > 0.5$. By definition, $r^R = 1 - \Phi(\min_{v \in \Pi(n-u+1)} Z_v)$. Since $Z < 0$ and

$$\frac{\min_{v \in \Pi(n-u+1)} Z_v}{\binom{n-1}{n-u}} \sum_{v \in \Pi(n-u+1)} \frac{SE}{SE_v} < Z,$$

it follows that $\min_{v \in \Pi(n-u+1)} Z_v < 0$ and therefore that $r^R > 0.5$.

Therefore, if $Z < 0$ we have $p^R < r^R$ and $\min(p^L, p^R) < \min(r^L, r^R)$. Similar arguments show that if $Z > 0$, $p^R < r^R$ and $\min(p^L, p^R) < \min(r^L, r^R)$. It thus follows that $p < r$.

remarkhe property that, with probability one, the global null p -value is smaller than the r -value, is not satisfied with popular combining functions such as Fisher, Simes, and Bonferroni.

For example, if $p_{(1)} \leq \dots \leq p_{(n)}$ are the ordered p -values, then the Bonferroni meta-analysis p -value is $n \times p_{(1)}$, its r -value for $u = 2$ is $(n - 1) \times p_{(2)}$, and $Pr(n \times p_{(1)} < (n - 1) \times p_{(2)}) > 0$.

Received XXX 201X. Revised XXX 201X. Accepted XX 201X.

Extensions of Fig. 2

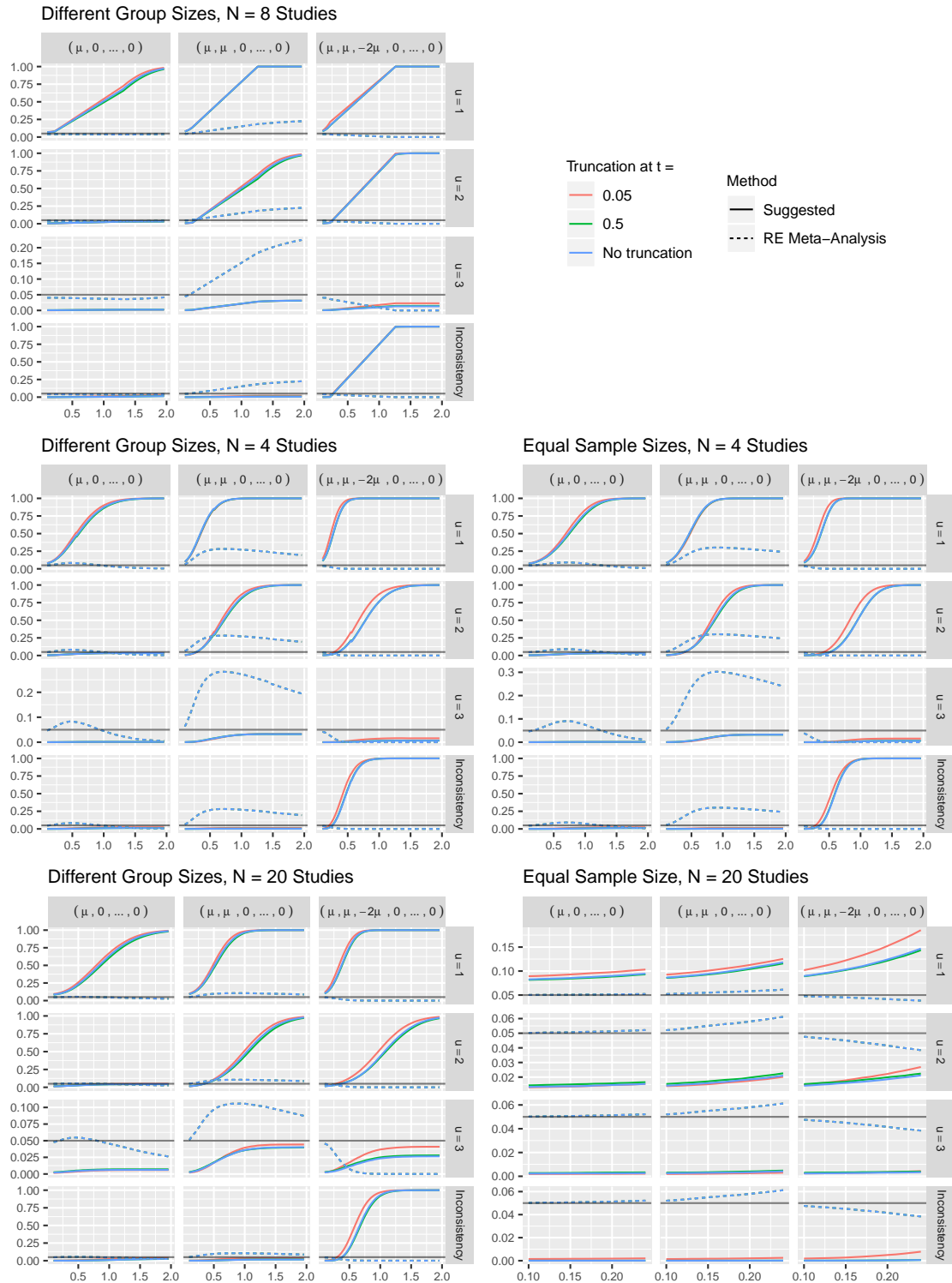


Figure 1. Each panel shows the results of a simulation similar to figure 2 in the body of the paper. The different panels differentiate by the number of studies N and whether the group sizes are equal or not (detailed in the panel title).

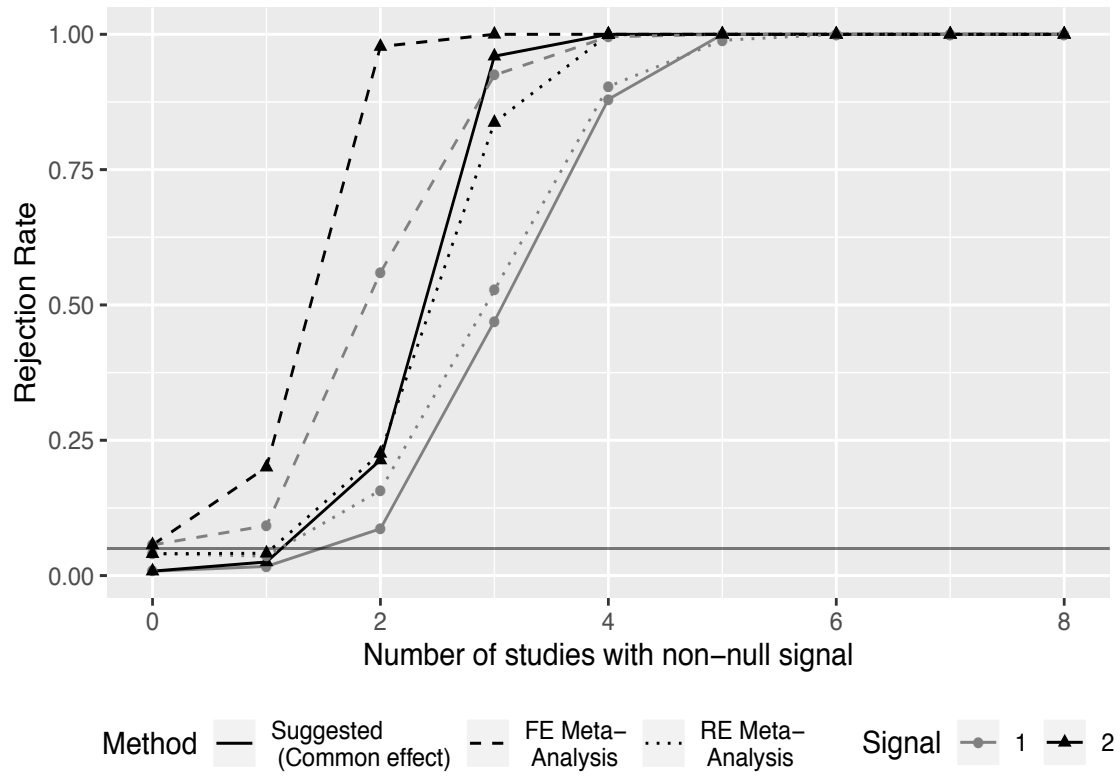


Figure 2. Each panel shows the results of a simulation similar to figure 3 in the body of the paper. The different panels differentiate by the number of studies N and whether the group sizes are equal or not (detailed in the panel title).

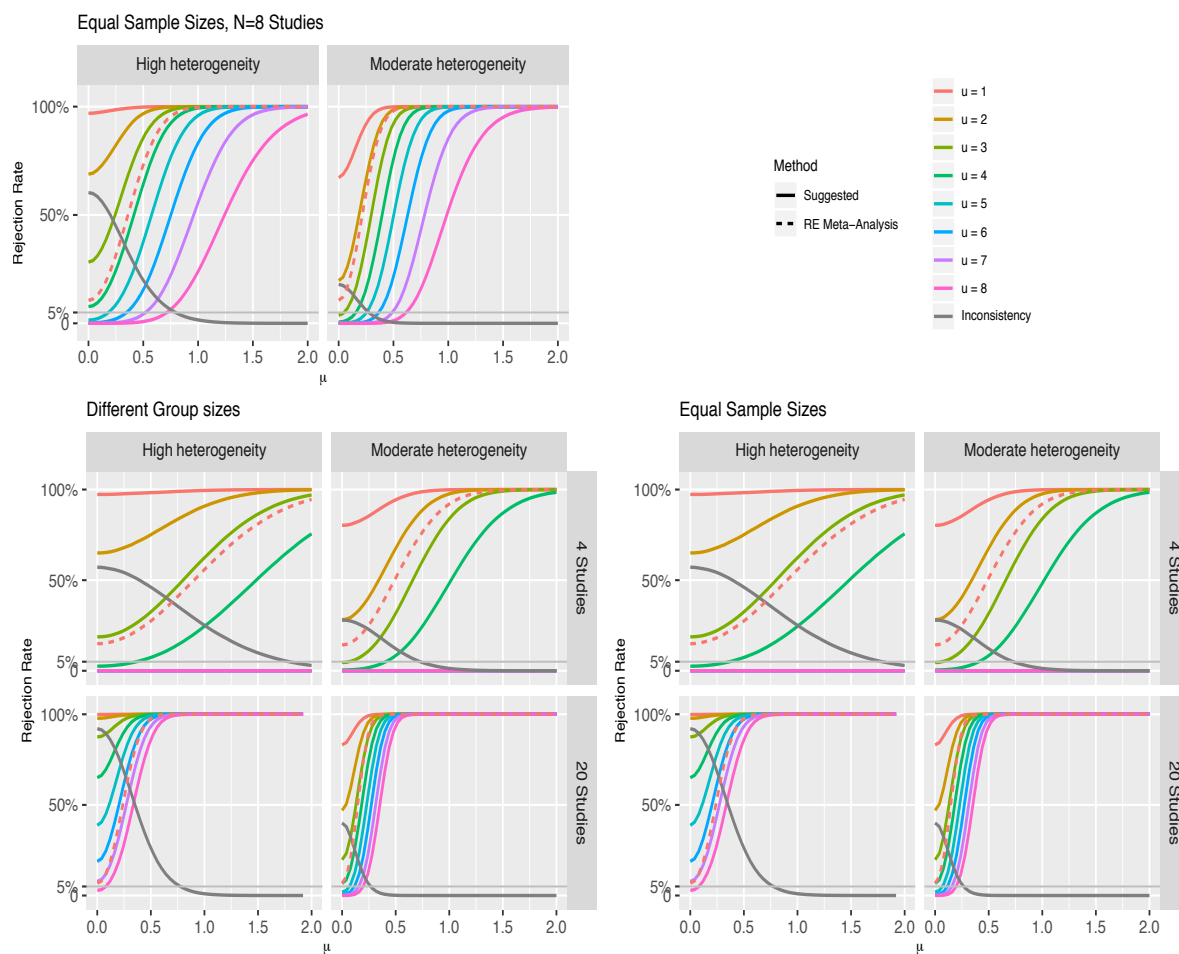


Figure 3. Each panel shows the results of a simulation similar to figure 3 in the body of the paper. The different panels differentiate by the number of studies N and whether the group sizes are equal or not (detailed in the panel title).