Supplement: Sensitivity Analysis for Publication Bias in Meta-Analyses

Contents

1. Theoretical results

1.1. Consistency of corrected point estimates

Here, we show that $\widehat{\mu}_\eta$ is consistent for μ; the proof is structured as follows. We first describe notation and assumptions. We establish a supporting Lemma 1.1, which states that the inverse-probability weights can be constructed using only the relative publication probability ratio, η, without specification of the absolute probability of publication for affirmative studies. In a second supporting Lemma 1.2 and Corollary 1.1, we find the expectations of terms that will appear in the main theorem and establish a limiting result. We then use these results to prove the main theorem (Theorem 1.1).

Notation and assumptions For the i^{th} underlying study, define the inverse-probability weight $\pi_i^* = \eta \mathbb{1}\{A_i^* = 0\} + \mathbb{1}\{A_i^* = 1\}$. As in the main text, let w_i^* denote an additional unstandardized, common-effects or random-effects inverse-variance weight; for example, for common-effects meta-analysis, $w_i^* = (\sigma_i^*)^{-2}$. We consider publication bias that operates based on a study's affirmative status (via the indicator D_i^* as defined in the main text) and also potentially based on studies' standard errors, σ_i^*. To the latter end, let F_i^* be an indicator variable whose success probability is an arbitrary function of σ_i^*, subject to the constraints given in the assumptions below. For example, if studies with smaller σ_i^* are more likely to be published, above and beyond their affirmative statuses, then selection might take a form similar to $F_i^* \sim \text{Bern}\left(\frac{1}{1+\exp \sigma_i^*}\right)$. (This functional form is purely illustrative; as we will show, selection via F_i^* can be simply be ignored entirely in estimation without specifying a functional form.) Then, study i is published if and only if $D_i^* = F_i^* = 1$. In the main text, we had focused on the special case in which publication bias operates only on affirmative status; this arises simply by setting $F_i^* = 1$ for all studies and taking study i as published if and only if $D_i^* = 1$.

The bias-corrected estimator given in the main text, which weights each published study by its inverse-probability weight π_i^* and its usual meta-analytic weight w_i^* but ignores selection via F_i^* (i.e., it does not incorporate weights related to selection on σ_i^*), can therefore be written as:

$$\widehat{\mu}_\eta := \sum_{i=1}^{k^*} D_i^* F_i^* \; \frac{\pi_i^* w_i^*}{\sum_{i=1}^{k^*} D_i^* F_i^* \pi_i^* w_i^*} \widehat{\theta}_i^*$$

We assume that:

$$E\left[\widehat{\theta}_i^* \mid \sigma_i^*\right] = E\left[\widehat{\theta}_i^*\right] \tag{A1}$$

$$E\left[D_i^* w_i^* \mid A_i^*, F_i^*\right] = E\left[D_i^* \mid A_i^*, F_i^*\right] E\left[w_i^* \mid A_i^*, F_i^*\right] \tag{A2}$$

$$E\left[D_i^* w_i^* \widehat{\theta}_i^* \mid A_i^*, F_i^*\right] = E\left[D_i^* \mid A_i^*, F_i^*\right] E\left[w_i^* \widehat{\theta}_i^* \mid A_i^*, F_i^*\right] \tag{A3}$$

$$F_i^* \perp\!\!\!\perp A_i^* \mid \sigma_i^* \tag{A4}$$

$$E\left[\widehat{\theta}_i^* \mid F_i^*, A_i^*, \sigma_i^*\right] = E\left[\widehat{\theta}_i^* \mid A_i^*, \sigma_i^*\right] \tag{A5}$$

$$\frac{1}{k^*} \sum_{i=1}^{k^*} D_i^* F_i^* P\left(D_i^* = 1 \mid A_i^*\right)^{-1} w_i^* = E\left[D_i^* F_i^* P\left(D_i^* = 1 \mid A_i^*\right)^{-1} w_i^*\right] + \mathcal{O}_p(1/\sqrt{k^*}) \tag{A6}$$

Assumption (A1) is a version of a standard assumption in meta-analysis and states that the point estimates are mean-independent from their standard errors. Assumptions (A2)-(A3) regarding uncorrelatedness essentially state that, conditional on a study's affirmative or nonaffirmative status and on whether it meets the selection criterion based on σ_i^*, selection on affirmative status does not select further based on the inverse-variance weights nor on the product of the point estimates with their inverse-variance weights. Assumptions (A4) and (A5) essentially state that any additional selection criterion based on studies' standard errors operates in the same way for affirmative and for nonaffirmative studies (A4) and that, conditional on a study's standard error and affirmative status, any selection criterion based on studies' standard errors does not also select based on the point estimate (A5). Assumption (A6) gives a limiting result that is often plausible by a Central Limit Theorem, such as the Lyapunov variant. Note that these assumptions are generalizations of those in the main text, which describe only selection based on D_i^*, and Assumptions (A4) and (A5) do not appear in the main text because they are relevant only when there is also selection on F_i^*.

We now establish the first of the two supporting lemmas.

Lemma 1.1 (Invariance to absolute probabilities)**.** *Weighting by π_i^* (that is, using the selection ratio η) is equivalent to weighting by the absolute probabilities $P\left(D_i^* = 1 \mid A_i^*\right)$, which differ from π_i^* only by an unknown scale factor corresponding to the probability for affirmative studies, $P\left(D^* = 1 \mid A^* = 1\right)$. (The study indices "i" are omitted from the term $P\left(D^* = 1 \mid A^* = 1\right)$ because under the assumed model of publication bias, this probability* *conditional on affirmative status is constant across studies.) That is:*

$$\widehat{\mu}_\eta = \sum_{i=1}^{k^*} D_i^* F_i^* \; \frac{P\left(D_i^* = 1 \mid A_i^*\right)^{-1} w_i^*}{\sum_{i=1}^{k*} D_i^* F_i^* P\left(D_i^* = 1 \mid A_i^*\right)^{-1} w_i^*} \; \widehat{\theta}_i^*$$

Proof. By the construction of π_i^*, we have:

$$\begin{aligned} P\left(D_i^* = 1 \mid A_i^*\right)^{-1} &= \begin{cases} P\left(D^* = 1 \mid A^* = 1\right)^{-1}, A_i^* = 1 \\ \eta P\left(D^* = 1 \mid A^* = 1\right)^{-1}, A_i^* = 0 \end{cases} \\ &= P\left(D^* = 1 \mid A^* = 1\right)^{-1} \pi_i^* \end{aligned}$$

Therefore, from the definition of $\widehat{\mu}_\eta$:

$$\begin{aligned} \widehat{\mu}_\eta &:= \sum_{i=1}^{k^*} D_i^* F_i^* \; \frac{\pi_i^* w_i^*}{\sum_{i=1}^{k^*} D_i^* F_i^* \pi_i^* w_i^*} \widehat{\theta}_i^* \\ &= \sum_{i=1}^{k^*} D_i^* F_i^* \; \frac{\frac{P\left(D_i^*=1 \mid A_i^*\right)^{-1}}{P(D^*=1 \mid A^*=1)^{-1}} w_i^*}{\sum_{i=1}^{k^*} D_i^* F_i^* \frac{P\left(D_i^*=1 \mid A_i^*\right)^{-1}}{P(D^*=1 \mid A^*=1)^{-1}} w_i^*} \; \widehat{\theta}_i^* \\ &= \sum_{i=1}^{k^*} D_i^* F_i^* \; \frac{P\left(D_i^* = 1 \mid A_i^*\right)^{-1} w_i^*}{\sum_{i=1}^{k*} D_i^* F_i^* P\left(D_i^* = 1 \mid A_i^*\right)^{-1} w_i^*} \; \widehat{\theta}_i^* \end{aligned}$$

as desired. We now establish the second supporting lemma. □

Lemma 1.2 (Expectations)**.** *We establish the expectations of two related terms that will appear in the proof of Theorem 1.1:*

$$E\left[D_i^* F_i^* P\left(D_i^* = 1 \mid A_i^*\right)^{-1} w_i^* \widehat{\theta}_i^*\right] = E\left[\widehat{\theta}_i^*\right] E_{\sigma_i^*}\left[P\left(F_i^* = 1 \mid \sigma_i^*\right) w_i^*\right] \tag{1.1}$$

$$E\left[D_i^* F_i^* P\left(D_i^* = 1 \mid A_i^*\right)^{-1} w_i^*\right] = E_{\sigma_i^*}\left[P\left(F_i^* = 1 \mid \sigma_i^*\right) w_i^*\right] \tag{1.2}$$

Proof. When helpful for clarity, we use subscripts on expectations to indicate the variable(s) with respect to which the expectation is taken. We use Φ to denote the cumulative distribution function of the standard normal distribution. Starting from the left-hand side of Equation 1.1 and taking iterated expectations first over (F_i^*, A_i^*) and then over σ_i^*:

$$E\left[D_i^* F_i^* P\left(D_i^* = 1 \mid A_i^*\right)^{-1} w_i^* \widehat{\theta}_i^*\right] = E_{\sigma_i^*}\left[E\left[E_{F_i^*, A_i^*}\left[E\left[D_i^* F_i^* P\left(D_i^* = 1 \mid A_i^*\right)^{-1} w_i^* \widehat{\theta}_i^* \mid F_i^*, A_i^*\right]\right] \mid \sigma_i^*\right]\right]$$

$$= E_{\sigma_i^*}\Bigg[E\Bigg[P\left(F_i^*=0, A_i^*=0\right)\times 0 + P\left(F_i^*=0, A_i^*=1\right)\times 0$$

$$+ P\left(F_i^*=1, A_i^*=0\right) E\Big[D_i^* F_i^* P\left(D_i^*=1 \mid A_i^*\right)^{-1} w_i^* \widehat{\theta}_i^* \mid F_i^*=1, A_i^*=0\Big]$$

$$+ P\left(F_i^*=1, A_i^*=1\right) E\Big[D_i^* F_i^* P\left(D_i^*=1 \mid A_i^*\right)^{-1} w_i^* \widehat{\theta}_i^* \mid F_i^*=1, A_i^*=1\Big]$$

$$\mid \sigma_i^*\Bigg]\Bigg]$$

Conditional on F_i^* and A_i^*, both F_i^* and $P\left(D_i^*=1 \mid A_i^*\right)^{-1}$ are fixed. By Assumption (A3) and the fact that D_i^* depends on A_i^* but not F_i^*, we have $E\big[D_i^* w_i^* \widehat{\theta}_i^* \mid F_i^*, A_i^*\big] = E\big[D_i^* \mid A_i^*\big] E\big[w_i^* \widehat{\theta}_i^* \mid F_i^*, A_i^*\big]$. Therefore:

$$= E_{\sigma_i^*}\Bigg[E\Bigg[P\left(F_i^*=1\right) P\left(A_i^*=0 \mid F_i^*=1\right) \cancel{E\Big[D_i^* \mid A_i^*=0\Big]} \cancel{P\left(D_i^*=1 \mid A_i^*=0\right)^{-1}} \times$$

$$E\Big[w_i^* \widehat{\theta}_i^* \mid F_i^*=1, A_i^*=0\Big] + P\left(F_i^*=1\right) P\left(A_i^*=1 \mid F_i^*=1\right) \cancel{E\Big[D_i^* \mid A_i^*=1\Big]} \cancel{P\left(D_i^*=1 \mid A_i^*=1\right)^{-1}}$$

$$E\Big[w_i^* \widehat{\theta}_i^* \mid F_i^*=1, A_i^*=1\Big] \mid \sigma_i^*\Bigg]\Bigg]$$

$$= \int_0^\infty P\left(F_i^*=1 \mid \widetilde{\sigma}_i^*\right) \Bigg\{P\left(A_i^*=0 \mid F_i^*=1, \widetilde{\sigma}_i^*\right) E\Big[w_i^* \widehat{\theta}_i^* \mid F_i^*=1, A_i^*=0, \widetilde{\sigma}_i^*\Big] +$$

$$P\left(A_i^*=1 \mid F_i^*=1, \widetilde{\sigma}_i^*\right) E\Big[w_i^* \widehat{\theta}_i^* \mid F_i^*=1, A_i^*=1, \widetilde{\sigma}_i^*\Big]\Bigg\} f_{\sigma_i^*}\left(\widetilde{\sigma}_i^*\right) \partial \widetilde{\sigma}_i^*$$

Assumption (A4) implies that $P\left(A_i^*=0 \mid F_i^*=1, \widetilde{\sigma}_i^*\right) = P\left(A_i^*=0 \mid \widetilde{\sigma}_i^*\right)$, and similarly for $A_i^*=1$. Additionally, conditional on $\widetilde{\sigma}_i^*$, the inverse-variance weights w_i^* are either exactly fixed (in the case of common-effect meta-analysis) or approximately fixed (in the case of random-effects meta-analysis with a relatively large number of studies). Therefore, letting $\widetilde{w}_i^*$ denote the inverse-variance weight calculated using $\widetilde{\sigma}_i^*$, we have $E\Big[\widetilde{w}_i^* \widehat{\theta}_i^* \mid F_i^*, A_i^*, \widetilde{\sigma}_i^*\Big] = \widetilde{w}_i^* E\Big[\widehat{\theta}_i^* \mid F_i^*, A_i^*, \widetilde{\sigma}_i^*\Big]$, so:

$$= \int_0^\infty P\left(F_i^*=1 \mid \widetilde{\sigma}_i^*\right) \Bigg\{P\left(A_i^*=0 \mid \widetilde{\sigma}_i^*\right) \widetilde{w}_i^* E\Big[\widehat{\theta}_i^* \mid F_i^*=1, A_i^*=0, \widetilde{\sigma}_i^*\Big] +$$

$$P\left(A_i^*=1 \mid \widetilde{\sigma}_i^*\right) \widetilde{w}_i^* E\Big[\widehat{\theta}_i^* \mid F_i^*=1, A_i^*=1, \widetilde{\sigma}_i^*\Big]\Bigg\} f_{\sigma_i^*}\left(\widetilde{\sigma}_i^*\right) \partial \widetilde{\sigma}_i^*$$

By Assumption (A5), $E\Big[\widehat{\theta}_i^* \mid F_i^*=1, A_i^*, \widetilde{\sigma}_i^*\Big] = E\Big[\widehat{\theta}_i^* \mid A_i^*, \widetilde{\sigma}_i^*\Big]$. Using this and also rewriting A_i^* in terms of its definition:

$$= \int_0^\infty P\left(F_i^*=1 \mid \widetilde{\sigma}_i^*\right) \Bigg\{ P\left(\widehat{\theta}_i^* \leq \Phi^{-1}(0.975)\, \widetilde{\sigma}_i^* \mid \widetilde{\sigma}_i^*\right) \widetilde{w}_i^* E\Big[\widehat{\theta}_i^* \mid \widehat{\theta}_i^* \leq \Phi^{-1}(0.975)\, \widetilde{\sigma}_i^*, \widetilde{\sigma}_i^*\Big] +$$

$$P\left(\widehat{\theta}_i^* > \Phi^{-1}(0.975)\, \widetilde{\sigma}_i^* \mid \widetilde{\sigma}_i^*\right) \widetilde{w}_i^* E\Big[\widehat{\theta}_i^* \mid \widehat{\theta}_i^* > \Phi^{-1}(0.975)\, \widetilde{\sigma}_i^*, \widetilde{\sigma}_i^*\Big]\Bigg\} f_{\sigma_i^*}\left(\widetilde{\sigma}_i^*\right) \partial \widetilde{\sigma}_i^*$$

Writing out the truncated conditional expectations:

$$= \int_0^\infty P\left(F_i^*=1 \mid \widetilde{\sigma}_i^*\right) \Bigg\{ \cancel{P\left(\widehat{\theta}_i^* \leq \Phi^{-1}(0.975)\, \widetilde{\sigma}_i^* \mid \widetilde{\sigma}_i^*\right)} \widetilde{w}_i^* \frac{1}{\cancel{P\left(\widehat{\theta}_i^* \leq \Phi^{-1}(0.975)\, \widetilde{\sigma}_i^* \mid \widetilde{\sigma}_i^*\right)}} \times$$

$$\int_{-\infty}^{\Phi^{-1}(0.975)\, \widetilde{\sigma}_i^*} q f_{\widehat{\theta}_i^* \mid \widetilde{\sigma}_i^*}(q) dq + \cancel{P\left(\widehat{\theta}_i^* > \Phi^{-1}(0.975)\, \widetilde{\sigma}_i^* \mid \widetilde{\sigma}_i^*\right)} \widetilde{w}_i^* \frac{1}{\cancel{P\left(\widehat{\theta}_i^* > \Phi^{-1}(0.975)\, \widetilde{\sigma}_i^* \mid \widetilde{\sigma}_i^*\right)}} \times$$

$$\int_{\Phi^{-1}(0.975)\, \widetilde{\sigma}_i^*}^{\infty} r f_{\widehat{\theta}_i^* \mid \widetilde{\sigma}_i^*}(r) dr \Bigg\} f_{\sigma_i^*}\left(\widetilde{\sigma}_i^*\right) \partial \widetilde{\sigma}_i^*$$

Combining the two integrals in the brackets:

$$= \int_0^\infty P\left(F_i^*=1 \mid \widetilde{\sigma}_i^*\right) \widetilde{w}_i^* \left\{ \int_{-\infty}^{\infty} t f_{\widehat{\theta}_i^* \mid \widetilde{\sigma}_i^*}(t) dt \right\} f_{\sigma_i^*}\left(\widetilde{\sigma}_i^*\right) \partial \widetilde{\sigma}_i^*$$

The bracketed term is now $E\big[\widehat{\theta}_i^* \mid \sigma_i^*\big]$, which is in fact equal to $E\big[\widehat{\theta}_i^*\big]$ by Assumption (A1). Therefore:

$$= E_{\sigma_i^*}\Big[P\left(F_i^*=1 \mid \sigma_i^*\right) w_i^* E\big[\widehat{\theta}_i^*\big]\Big]$$

$$= E\big[\widehat{\theta}_i^*\big] E_{\sigma_i^*}\Big[P\left(F_i^*=1 \mid \sigma_i^*\right) w_i^*\Big]$$

This proves Equation 1.1. The proof of Equation 1.2 follows nearly identical mechanics except that, instead of invoking Assumption (A3), we instead invoke Assumption (A2) to write $E\big[D_i^* w_i^* \mid F_i^*, A_i^*\big]$ as $E\big[D_i^* \mid A_i^*\big] E\big[w_i^* \mid F_i^*, A_i^*\big]$. □

Corollary 1.1 (Limiting result)**.**

$$\frac{1}{k^*}\sum_{i=1}^{k^*} D_i^* F_i^* P\left(D_i^*=1 \mid A_i^*\right)^{-1} w_i^* = E_{\sigma_i^*}\Big[P\left(F_i^*=1 \mid \sigma_i^*\right) w_i^*\Big] + \mathcal{O}_p(1/\sqrt{k^*})$$

Proof. This follows immediately from combining the limiting result of Assumption (A6) with the expectation of Lemma 1.2. □

Theorem 1.1 (Consistency of $\widehat{\mu}_\eta$)**.** $\widehat{\mu}_\eta$ *is consistent for the mean effect size in the underlying population:*

$$\widehat{\mu}_\eta := \sum_{i=1}^{k^*} D_i^* F_i^* \ \frac{\pi_i^* w_i^*}{\sum_{i=1}^{k^*} D_i^* F_i^* \pi_i^* w_i^*}\widehat{\theta}_i^* \ \xrightarrow[k^*\to\infty]{p} \ E\big[\widehat{\theta}_i^*\big] = \mu$$

Proof. Taking limits, rewriting $\widehat{\mu}_\eta$ as in Lemma 1.1, and introducing $\frac{k^*}{k^*}$ inside the summation:

$$\lim_{k^*\to\infty}\sum_{i=1}^{k^*} D_i^* F_i^* \ \frac{P\left(D_i^*=1 \mid A_i^*\right)^{-1} w_i^*}{\sum_{i=1}^{k^*} D_i^* F_i^* P\left(D_i^*=1 \mid A_i^*\right)^{-1} w_i^*} \ \widehat{\theta}_i^*$$

$$= \lim_{k^*\to\infty}\sum_{i=1}^{k^*}\left\{\frac{1}{\frac{1}{k^*}\sum_{i=1}^{k^*} D_i^* F_i^* P\left(D_i^*=1 \mid A_i^*\right)^{-1} w_i^*} \times \frac{1}{k^*} D_i^* F_i^* \ P\left(D_i^*=1 \mid A_i^*\right)^{-1} \ w_i^* \widehat{\theta}_i^*\right\}$$

Applying the limiting result of Corollary 1.1 to the denominator term:

$$= \lim_{k^*\to\infty}\sum_{i=1}^{k^*}\left\{\frac{1}{E_{\sigma_i^*}\Big[P\left(F_i^*=1 \mid \sigma_i^*\right) w_i^*\Big] + \mathcal{O}_p(1/\sqrt{k^*})} \times \frac{1}{k^*} D_i^* F_i^* \ P\left(D_i^*=1 \mid A_i^*\right)^{-1} \ w_i^* \widehat{\theta}_i^*\right\}$$

The term $E_{\sigma_i^*}\Big[P\left(F_i^*=1 \mid \sigma_i^*\right) w_i^*\Big] + \mathcal{O}_p(1/\sqrt{k^*})$ is the same for all i, yielding:

$$= \lim_{k^*\to\infty}\frac{1}{E_{\sigma_i^*}\Big[P\left(F_i^*=1 \mid \sigma_i^*\right) w_i^*\Big] + \mathcal{O}_p(1/\sqrt{k^*})} \times \lim_{k^*\to\infty}\sum_{i=1}^{k^*}\frac{1}{k^*} D_i^* F_i^* \ P\left(D_i^*=1 \mid A_i^*\right)^{-1} \ w_i^* \widehat{\theta}_i^*$$

$$= \frac{1}{E_{\sigma_i^*}\Big[P\left(F_i^*=1 \mid \sigma_i^*\right) w_i^*\Big]} E\Big[D_i^* F_i^* \ P\left(D_i^*=1 \mid A_i^*\right)^{-1} w_i^* \widehat{\theta}_i^*\Big]$$

$$= E\left[\widehat{\theta}_i^*\right]$$

The final equality follows from applying Equation 1.1.

□

Given Theorem 1.1, our subsequent theoretical developments assume without loss of generality that there is no selection based on the standard errors and describe as "published" all studies with $D_i^* = 1$.

1.2. Conditions for the assumption of one-tailed selection to be conservative

We now establish conditions under which, when conducting sensitivity analyses for $\widehat{\mu}$, assuming one-tailed selection is conservative compared to assuming two-tailed selection. To this end, we first establish a lemma establishing the conditions under which the corrected estimate under the assumption of one-tailed selection, $\widehat{\mu}_\eta$, is conservative compared to its counterpart under the assumption two-tailed selection (Lemma 1.3). Then, by assuming that the conditions in Lemma 1.3 hold, we establish a lemma showing that when the corrected point estimates are nondecreasing in η, this indicates that no amount of publication bias could shift the point estimate to q (or alternatively that the point estimate is already equal to q), which we term "complete robustness" (Lemma 1.4). Finally, in Theorem 1.2, we show the desired conservatism result regarding $S(\widehat{\mu}, q)$. We first consider the common-effect specifications, later arguing that results for both random-effects specifications follow essentially identical logic.

Denote the set of "significant" negative, published studies and the set of "nonsignificant", published studies respectively as $\mathcal{N}^- = \{i : \widehat{\theta}_i < 0, p_i < 0.05\}$ and $\mathcal{N}^0 = \{i : p_i \geq 0.05\}$, such that the set of published nonaffirmative studies can be expressed as $\mathcal{N} = \mathcal{N}^- \bigcup \mathcal{N}^0$. We can rewrite the common-effect $\widehat{\mu}_\eta$ under the assumption of one-tailed selection, as in the main text, as:

$$\widehat{\mu}_\eta = \left(\sum_{i \in \mathcal{N}^0} \frac{\eta}{\sigma_i^2} \widehat{\theta}_i + \sum_{j \in \mathcal{N}^-} \frac{\eta}{\sigma_j^2} \widehat{\theta}_j + \sum_{l \in \mathcal{A}} \frac{1}{\sigma_l^2} \widehat{\theta}_l \right) \left(\sum_{i \in \mathcal{N}^0} \frac{\eta}{\sigma_i^2} + \sum_{j \in \mathcal{N}^-} \frac{\eta}{\sigma_j^2} + \sum_{l \in \mathcal{A}} \frac{1}{\sigma_l^2} \right)^{-1}$$

An analog under the assumption of two-tailed selection, defined as $\widehat{\mu}^t_\eta$, simply removes the upweighting on studies in $\mathcal{N}^-$:

$$\widehat{\mu}^t_\eta = \left(\sum_{i \in \mathcal{N}^0} \frac{\eta}{\sigma_i^2} \widehat{\theta}_i + \sum_{j \in \mathcal{N}^-} \frac{1}{\sigma_j^2} \widehat{\theta}_j + \sum_{l \in \mathcal{A}} \frac{1}{\sigma_l^2} \widehat{\theta}_l \right) \left(\sum_{i \in \mathcal{N}^0} \frac{\eta}{\sigma_i^2} + \sum_{j \in \mathcal{N}^-} \frac{1}{\sigma_j^2} + \sum_{l \in \mathcal{A}} \frac{1}{\sigma_l^2} \right)^{-1}$$

$$= \left(\eta \bar{y}_{\mathcal{N}^0} + \bar{y}_{\mathcal{N}^-} + \bar{y}_{\mathcal{A}} \right) \left(\eta \nu_{\mathcal{N}^0} + \nu_{\mathcal{N}^-} + \nu_{\mathcal{A}} \right)^{-1}$$

We now establish the two lemmas and theorem regarding conservatism. Without loss of generality, we consider the case in which the naïve estimate $\widehat{\mu} > 0$, such that conservatism holds, by definition, when $\widehat{\mu}_\eta \leq \widehat{\mu}^t_\eta$ for all η.

Lemma 1.3 (Equivalent condition and sufficient condition for conservatism of $\widehat{\mu}_\eta$)**.** $\widehat{\mu}_\eta \leq \widehat{\mu}^t_\eta$ *for all* $\eta \geq 1$ *if and only if:*

$$\frac{\eta \bar{y}_{\mathcal{N}^0} + \bar{y}_{\mathcal{A}}}{\eta \nu_{\mathcal{N}^0} + \nu_{\mathcal{A}}} \geq \frac{\bar{y}_{\mathcal{N}^-}}{\nu_{\mathcal{N}^-}} \quad \text{for all} \;\; \eta \geq 1 \tag{1.3}$$

This condition states that the inverse-probability-weighted, common-effects mean among only the "nonsignificant" and affirmative studies must be at least as large as the common-effects mean among only the "significant" negative studies. Note that since $\bar{y}_{\mathcal{A}} \geq 0$ *and* $\bar{y}_{\mathcal{N}^-} \leq 0$, *a sufficient condition for Equation (1.3) to hold is that* $\bar{y}_{\mathcal{N}^0} \geq 0$.

Proof. Let $A = \eta \bar{y}_{\mathcal{N}^0} + \bar{y}_{\mathcal{A}}$ and $B = \eta \nu_{\mathcal{N}^0} + \nu_{\mathcal{A}} > 0$. Then, conservatism holds by definition when, for all $\eta \geq 1$:

$$\begin{aligned}
\widehat{\mu}^t_\eta &\geq \widehat{\mu}_\eta \\
\left(A + \bar{y}_{\mathcal{N}^-}\right)\left(B + \nu_{\mathcal{N}^-}\right)^{-1} &\geq \left(A + \eta \bar{y}_{\mathcal{N}^-}\right)\left(B + \eta \nu_{\mathcal{N}^-}\right)^{-1} \\
\left(A + \bar{y}_{\mathcal{N}^-}\right)\left(B + \eta \nu_{\mathcal{N}^-}\right) &\geq \left(A + \eta \bar{y}_{\mathcal{N}^-}\right)\left(B + \nu_{\mathcal{N}^-}\right) \\
\cancel{AB} + A\eta\nu_{\mathcal{N}^-} + B\bar{y}_{\mathcal{N}^-} + \cancel{\eta\nu_{\mathcal{N}^-}\bar{y}_{\mathcal{N}^-}} &\geq \cancel{AB} + A\nu_{\mathcal{N}^-} + B\eta\bar{y}_{\mathcal{N}^-} + \cancel{\eta\nu_{\mathcal{N}^-}\bar{y}_{\mathcal{N}^-}} \\
A\nu_{\mathcal{N}^-}\cancel{(\eta - 1)} &\geq B\bar{y}_{\mathcal{N}^-}\cancel{(\eta - 1)} \\
\frac{\eta \bar{y}_{\mathcal{N}^0} + \bar{y}_{\mathcal{A}}}{\eta \nu_{\mathcal{N}^0} + \nu_{\mathcal{A}}} &\geq \frac{\bar{y}_{\mathcal{N}^-}}{\nu_{\mathcal{N}^-}}
\end{aligned}$$

All steps are bidirectional, so the desired claim holds. □

Lemma 1.4 (Complete robustness)**.** *Let $\widehat{\mu}_\eta^{-1}(q)$ and $\left(\widehat{\mu}_\eta^t\right)^{-1}(q)$ be inverses with respect to η, taking q to be fixed. Let $S^t(\widehat{\mu}, q) := \left(\widehat{\mu}_\eta^t\right)^{-1}(q)$ denote a two-tailed counterpart to $S(\widehat{\mu}, q)$. For both the one-tailed and the two-tailed estimators, if the corrected point estimate is nondecreasing in η, then we have complete robustness. That is, $\frac{\partial \widehat{\mu}_\eta}{\partial \eta} \geq 0 \Rightarrow S(\widehat{\mu}, q) \leq 1$ and $\frac{\partial \widehat{\mu}_\eta^t}{\partial \eta} \geq 0 \Rightarrow S^t(\widehat{\mu}, q) \leq 1$.*

Proof. Trivially, we have $\widehat{\mu}_{\eta=1} = \widehat{\mu}_{\eta=1}^t = \widehat{\mu}$, where $\widehat{\mu}$ is the uncorrected point estimate. Since $\widehat{\mu}_\eta$ and $\widehat{\mu}_\eta^t$ are nondecreasing in η by assumption, we have for all $q < \widehat{\mu}$ that $S(\widehat{\mu}, q) := \widehat{\mu}_\eta^{-1}(q) \leq 1$ and $S^t(\widehat{\mu}, q) := \left(\widehat{\mu}_\eta^t\right)^{-1}(q) \leq 1$. □

Theorem 1.2 (Conservatism of $S(\widehat{\mu}, q)$)**.** *Assume Lemma 1.3 holds. Then the one-tailed $S(\widehat{\mu}, q)$ is conservative compared to its two-tailed counterpart, $S^t(\widehat{\mu}, q)$, in the sense that:*

$$
\begin{aligned}
&S(\widehat{\mu}, q) \leq S^t(\widehat{\mu}, q), \text{ for } S(\widehat{\mu}, q) > 1 \text{ and } S^t(\widehat{\mu}, q) > 1 \\
&S(\widehat{\mu}, q) \leq 1 \Rightarrow S^t(\widehat{\mu}, q) \leq 1
\end{aligned}
$$

The first line states that when both $S(\widehat{\mu}, q)$ and $S^t(\widehat{\mu}, q)$ indicate some sensitivity to publication bias rather than complete robustness, the former indicates at least as much sensitivity as the latter. Excluding the trivial case in which $S(\widehat{\mu}, q) = S^t(\widehat{\mu}, q) = 1$, the second line states that when $S(\widehat{\mu}, q)$ indicates complete robustness to publication bias, then so must $S^t(\widehat{\mu}, q)$. That is, there may be cases in which both $S(\widehat{\mu}, q)$ and $S^t(\widehat{\mu}, q)$ indicate complete robustness and in which $S(\widehat{\mu}, q)$ indicates some sensitivity while $S^t(\widehat{\mu}, q)$ indicates complete robustness, but there cannot be cases in which $S(\widehat{\mu}, q)$ indicates complete robustness while $S^t(\widehat{\mu}, q)$ indicates some sensitivity.

Proof. We ignore the trivial case in which $S(\widehat{\mu}, q) = S^t(\widehat{\mu}, q) = 1$, such that $\widehat{\mu} = q$ already. For the other cases, we first establish conditions under which $\widehat{\mu}_\eta$ and $\widehat{\mu}_\eta^t$ are monotonically decreasing or increasing in η. For $\widehat{\mu}_\eta$, we have:

$$
\frac{\partial \widehat{\mu}_\eta}{\partial \eta} = \frac{\bar{y}_{\mathcal{N}} \nu_{\mathcal{A}} - \bar{y}_{\mathcal{A}} \nu_{\mathcal{N}}}{\left(\eta \nu_{\mathcal{N}} + \nu_{\mathcal{A}}\right)^2} \tag{1.4}
$$

$$
= \begin{cases} < 0, \text{ for } \frac{\bar{y}_{\mathcal{N}}}{\nu_{\mathcal{N}}} < \frac{\bar{y}_{\mathcal{A}}}{\nu_{\mathcal{A}}} \\ 0, \text{ for } \frac{\bar{y}_{\mathcal{N}}}{\nu_{\mathcal{N}}} = \frac{\bar{y}_{\mathcal{A}}}{\nu_{\mathcal{A}}} \\ > 0, \text{ for } \frac{\bar{y}_{\mathcal{N}}}{\nu_{\mathcal{N}}} > \frac{\bar{y}_{\mathcal{A}}}{\nu_{\mathcal{A}}} \end{cases} \tag{1.5}
$$

For $\widehat{\mu}^t_\eta$, we have:

$$\frac{\partial \widehat{\mu}^t_\eta}{\partial \eta} = \frac{\bar{y}_{\mathcal{N}^0}\left(\nu_{\mathcal{N}^-} + \nu_{\mathcal{A}}\right) - \left(\bar{y}_{\mathcal{N}^-} + \bar{y}_{\mathcal{A}}\right)\nu_{\mathcal{N}^0}}{\left(\eta\nu_{\mathcal{N}^0} + \nu_{\mathcal{N}^-} + \nu_{\mathcal{A}}\right)^2} \tag{1.6}$$

$$\begin{cases} < 0, \text{ for } \frac{\bar{y}_{\mathcal{N}^0}}{\nu_{\mathcal{N}^0}} < \frac{\bar{y}_{\mathcal{N}^-}+\bar{y}_{\mathcal{A}}}{\nu_{\mathcal{N}^-}+\nu_{\mathcal{A}}} \\ 0, \text{ for } \frac{\bar{y}_{\mathcal{N}^0}}{\nu_{\mathcal{N}^0}} = \frac{\bar{y}_{\mathcal{N}^-}+\bar{y}_{\mathcal{A}}}{\nu_{\mathcal{N}^-}+\nu_{\mathcal{A}}} \\ > 0, \text{ for } \frac{\bar{y}_{\mathcal{N}^0}}{\nu_{\mathcal{N}^0}} > \frac{\bar{y}_{\mathcal{N}^-}+\bar{y}_{\mathcal{A}}}{\nu_{\mathcal{N}^-}+\nu_{\mathcal{A}}} \end{cases} \tag{1.7}$$

We therefore have four cases to consider:

Case 1: $\frac{\partial \widehat{\mu}_\eta}{\partial \eta} < 0$ and $\frac{\partial \widehat{\mu}^t_\eta}{\partial \eta} < 0$

By definition, $S(\widehat{\mu}, q) = \widehat{\mu}_\eta^{-1}(q)$ and $S^t(\widehat{\mu}, q) = \widehat{\mu}_\eta^{t\,-1}(q)$. Since both $\widehat{\mu}_\eta$ and $\widehat{\mu}^t_\eta$ are monotonically decreasing in η and $\widehat{\mu}_\eta \leq \widehat{\mu}^t_\eta$ by Lemma 1.3, we have $S(\widehat{\mu}, q) \leq S^t(\widehat{\mu}, q)$, so conservatism holds.

Case 2: $\frac{\partial \widehat{\mu}_\eta}{\partial \eta} \geq 0$ and $\frac{\partial \widehat{\mu}^t_\eta}{\partial \eta} \geq 0$

In this case, $S(\widehat{\mu}, q) \leq 1$ and $S^t(\widehat{\mu}, q) \leq 1$ by Lemma 1.4, so both indicate complete robustness, and the notion of conservatism is not meaningful.

Case 3: $\frac{\partial \widehat{\mu}_\eta}{\partial \eta} < 0$ and $\frac{\partial \widehat{\mu}^t_\eta}{\partial \eta} \geq 0$

In this case, $S^t(\widehat{\mu}, q) \leq 1$ by Lemma 1.4, indicating complete robustness to publication bias. If we also have $S(\widehat{\mu}, q) \leq 1$, then both estimators indicate complete robustness. If instead $S(\widehat{\mu}, q) > 1$, then conservatism holds.

Case 4: $\frac{\partial \widehat{\mu}_\eta}{\partial \eta} \geq 0$ and $\frac{\partial \widehat{\mu}^t_\eta}{\partial \eta} < 0$

Since $\widehat{\mu}_\eta \leq \widehat{\mu}^t_\eta$ for $\eta > 1$ by Lemma 1.3 and $\widehat{\mu}_\eta = \widehat{\mu}^t_\eta$ for $\eta = 1$, this case is not possible.

Thus, conservatism holds for all cases in which the notion is meaningful. □

For both random-effects specifications, the proof is identical upon replacing σ_i^2 with $\sigma_i^2 + \widehat{\tau}^2$

in the weights for $\bar{y}_{\mathcal{N}}$, $\nu_{\mathcal{N}}$, and their counterparts for the sets $\mathcal{A}$, $\mathcal{N}^0$, and $\mathcal{N}^-$. This works because $\widehat{\tau}^2$ is held constant between the one- and two-tailed specifications; as described in the main text, it is treated as a nuisance parameter that is estimated in a naïve initial model rather than estimated jointly with $\widehat{\mu}_\eta$. Note that $S(\widehat{\mu}^{lb}, q)$, unlike $S(\widehat{\mu}, q)$, is not necessarily conservative compared to its two-tailed counterpart, $S^t(\widehat{\mu}^{lb}, q)$. This is because $\widehat{\mu}^t_\eta$ upweights a smaller number of studies than $\widehat{\mu}_\eta$, so especially for large η, $\widehat{\mu}^t_\eta$ will typically have a smaller effective sample size and hence a wider confidence interval than $\widehat{\mu}_\eta$. Thus, even if $\widehat{\mu}_\eta < \widehat{\mu}^t_\eta$, we may have $\widehat{\mu}^{lb}_\eta > \widehat{\mu}^{t,lb}_\eta$, such that $S^t(\widehat{\mu}^{lb}, q)$ is in fact more conservative with respect to the confidence interval limit.

1.3. A "fail-safe" number

Lemma 1.5. *Let $|\mathcal{N}|$ denote the number of published nonaffirmative studies and $|\mathcal{N}^*|$ denote the total number of nonaffirmative studies in the underlying population, such that $(|\mathcal{N}^*| - |\mathcal{N}|)$ represents the number of unpublished nonaffirmative studies. Then, an approximate lower bound on the number of unpublished nonaffirmative studies is:*

$$(|\mathcal{N}^*| - |\mathcal{N}|) \gtrapprox |\mathcal{N}| \times (S(t,q) - 1)$$

Proof. Using the same notation introduced in Section 1.1 above, we can express the probability of publication for each nonaffirmative study in the underlying population via Bayes' Rule:

$$P(D^*_i = 1 \mid A^*_i = 0) = \frac{P(A^*_i = 0 \mid D^*_i = 1)\, P(D^*_i = 1)}{P(A^*_i = 0)}$$

The left-hand side can be rewritten using the definition of $S(t,q)$ as a ratio of publication probabilities, such that $P(D^*_i = 1 \mid A^*_i = 1) / P(D^*_i = 1 \mid A^*_i = 0) = S(t,q)$. For the right-hand side, note that $P(A^*_i = 0 \mid D^*_i = 1) = P(A_i = 0 \mid D_i = 1)$ because all underlying results with $D^*_i = 1$ are by definition also in the published sample. In turn, $P(A_i = 0 \mid D_i = 1) \approx |\mathcal{N}|/k$, its sample estimate. Similar sample estimates or proportions in the underlying population can be substituted for the other terms on the right-hand side. Thus:

$$\frac{P(D^*_i = 1 \mid A^*_i = 1)}{S(t,q)} \approx \frac{(|\mathcal{N}|/k)\,(k/k^*)}{(|\mathcal{N}^*|/k^*)}$$

$$|\mathcal{N}^*| \approx \frac{|\mathcal{N}| \times S(t,q)}{P(D_i^* = 1 \mid A_i^* = 1)}$$

Minimizing the right hand side over $P(D_i^* = 1 \mid A_i^* = 1)$ by setting $P(D_i^* = 1 \mid A_i^* = 1) = 1$ yields $|\mathcal{N}|^* \gtrapprox |\mathcal{N}| \times S(t,q)$, which immediately yields the desired result.

□

For example, if applying the proposed sensitivity analyses yields $S(t,q) = 10$, and we observe $|\mathcal{N}| = 5$ nonaffirmative studies, then we estimate that there would need to be at least $5 \times (10 - 1) = 45$ unpublished nonaffirmative studies in order to shift the estimate t to q. Under our assumed model of publication bias, these unpublished nonaffirmative studies are assumed to be comparable to the published nonaffirmative studies as in the assumptions formalized in Section 1.1. Like a very large value of $S(t,q)$, a very large fail-safe number provides some reassurance that the meta-analysis results are robust to even severe publication bias. Our proposed fail-safe number is conceptually related to previous methods (e.g., Orwin (1983); Rosenthal (1979)), but relaxes those methods' assumption of homogeneous population effects. Additionally, by treating the published nonaffirmative studies as representative of the underlying population of nonaffirmative studies, the present fail-safe number does not require specifying the mean of the unpublished studies.

The fail-safe number is an approximate lower bound, reflecting the fact that the minimum number of unobserved nonaffirmative studies for any given relative probability of publication, $S(t,q)$, is attained when the affirmative studies' *absolute* probability of publication is maximized. If affirmative results have a publication probability less than 1, then $(|\mathcal{N}^*| - |\mathcal{N}|)$ would increase yet further. When interpreting the fail-safe number as a metric of robustness, it is important to recall that the underlying population technically comprises all conducted hypothesis tests that would, if published, have been included in the meta-analysis. Thus, $(|\mathcal{N}^*| - |\mathcal{N}|)$ counts not only papers written but never accepted for publication, but also potentially multiple hypothesis tests on independent samples conducted for any given paper.

1.4. A parametric specification

As an alternative to the robust independent specification presented in the main text, it would be possible to conduct maximum-likelihood sensitivity analyses under the standard parametric, independent random-effects model, invoking the additional assumptions that, in the *published* studies, $\gamma_i \sim_{iid} N(0, \tau^2)$ and $\epsilon_i \sim_{iid} N(0, \sigma_i^2)$ (e.g., Brockwell & Gordon (2001); Viechtbauer (2005)). We considered this approach for several reasons. First, when correctly specified, the parametric score approach would likely be more efficient than the robust independent specification. Second, unlike the robust independent specification, the parametric approach enables direct estimation of τ^2; this estimate is both informative in its own right and could in principle be used to construct more efficient weights for the robust specifications. In direct analog to inverse-probability weighting for survey sampling or missing data for general M-estimators (Wooldridge, 2007), the approach we consider here weights the score contributions of each observation. Under the parametric random-effects specification, we have $\widehat{\theta}_i \sim N(\mu, \tau^2 + \sigma_i^2)$, leading to the following log-likelihood (Brockwell & Gordon, 2001; Veroniki et al., 2015):

$$\log \mathcal{L}(\mu, \tau^2) = -\frac{1}{2} \sum_{i=1}^{k} \log\left(2\pi\left(\tau^2 + \sigma_i^2\right)\right) - \frac{1}{2} \sum_{i=1}^{k} \frac{\left(\widehat{\theta}_i - \mu\right)^2}{\tau^2 + \sigma_i^2} \ , \ \tau^2 \geq 0$$

Letting $\mathcal{L}_i$ denote the contribution of the i^{th} study to the likelihood, the score contributions are:

$$\begin{aligned}
\frac{\partial \log \mathcal{L}_i}{\partial \mu} &= -\frac{1}{2\left(\tau^2 + \sigma_i^2\right)} \times 2\left(\widehat{\theta}_i - \mu\right) \times (-1) \\
&= \frac{\widehat{\theta}_i - \mu}{\tau^2 + \sigma_i^2} \\
\frac{\partial \log \mathcal{L}_i}{\partial \tau^2} &= -\frac{1}{2} \frac{2\pi}{2\pi\left(\tau^2 + \sigma_i^2\right)} - \left(\frac{1}{2}\left(\widehat{\theta}_i - \mu\right)^2 \times \left[-\left(\tau^2 + \sigma_i^2\right)^{-2}\right]\right) \\
&= -\frac{1}{2\left(\tau^2 + \sigma_i^2\right)} + \frac{\left(\widehat{\theta}_i - \mu\right)^2}{2\left(\tau^2 + \sigma_i^2\right)^2} \\
&= \frac{\left(\widehat{\theta}_i - \mu\right)^2 - \left(\tau^2 + \sigma_i^2\right)}{2\left(\tau^2 + \sigma_i^2\right)^2}
\end{aligned}$$

The usual maximum likelihood estimators without correction for publication bias are therefore

(Brockwell & Gordon, 2001; Viechtbauer, 2005):

$$\widehat{\mu} = \frac{\sum_{i=1}^{k} \frac{1}{\widehat{\tau}^2 + \sigma_i^2} \widehat{\theta}_i}{\sum_{i=1}^{k} \frac{1}{\widehat{\tau}^2 + \sigma_i^2}}$$

$$\widehat{\tau}^2 = \max\left\{ 0, \frac{\sum_{i=1}^{k} \left(\frac{1}{\widehat{\tau}^2 + \sigma_i^2}\right)^2 \left(\left(\widehat{\theta}_i - \widehat{\mu}\right)^2 - \sigma_i^2\right)}{\sum_{i=1}^{k} \left(\frac{1}{\widehat{\tau}^2 + \sigma_i^2}\right)^2} \right\}$$

The publication bias-corrected score contributions are:

$$\frac{\partial \log \mathcal{L}_i}{\partial \mu} = \frac{\pi_i \left(\widehat{\theta}_i - \mu\right)}{\tau^2 + \sigma_i^2}$$

$$\frac{\partial \log \mathcal{L}_i}{\partial \tau^2} = \frac{\pi_i \Big[\left(\widehat{\theta}_i - \mu\right)^2 - (\tau^2 + \sigma_i^2) \Big]}{2\left(\tau^2 + \sigma_i^2\right)^2}$$

Upon setting the summed bias-corrected score contributions equal to 0, the maximum likelihood estimates can be obtained in the usual iterative manner, and their asymptotic variances can be estimated as a function of the unweighted Hessian and bias-corrected score contributions per Wooldridge (2007)'s Equation (3.10). Our code is publicly available (`https://osf.io/7wc2t/`). We next describe the empirical behavior of this estimation approach.

We assessed the bias and efficiency of the bias-corrected score specification using a similar simulation study as that described in the main text, considering only scenarios with normal population effects and no selection on the standard error. We were primarily interested in assessing the method's performance for the scenarios without clustering (i.e., $\text{Var}(\zeta) = 0$), for which the bias-corrected score specification is correctly specified. Additionally, for scenarios with clustering ($\text{Var}(\zeta) = 0.5$), we investigated the impact on efficiency of weighting the robust clustered model using an estimate $\widehat{\tau}^2$ from the bias-corrected score model instead of from the naïve parametric model. As expected, Figure S1 shows that, when correctly specified, the bias-corrected score model had nominal coverage when $\eta = 1$ regardless of sample size. However, its coverage sharply declined with increasing η unless the number of studies was large (bottom row). Considering all scenarios clustering, weighting the robust clustered model by the bias-corrected $\widehat{\tau}^2$ versus the naïve estimate made little difference in coverage or efficiency. We speculate that the latter finding regarding efficiency reflects our observation that the bias-corrected $\widehat{\tau}^2$ was in fact quite biased except with very small η or unrealistically large k. Given the overall poor performance of the bias-corrected score model in realistic scenarios, we did not pursue this approach and do not recommend its use in practice.

Figure S1: *Mean coverage in scenarios without clustering. "Robust (score)": Robust independent model in which $\widehat{\tau}^2$ is chosen by first fitting the weighted score model. "Robust independent": Robust independent model as in the main text, in which $\widehat{\tau}^2$ is chosen by first fitting the naïve parametric model. "Wtd. score": Weighted score model.*

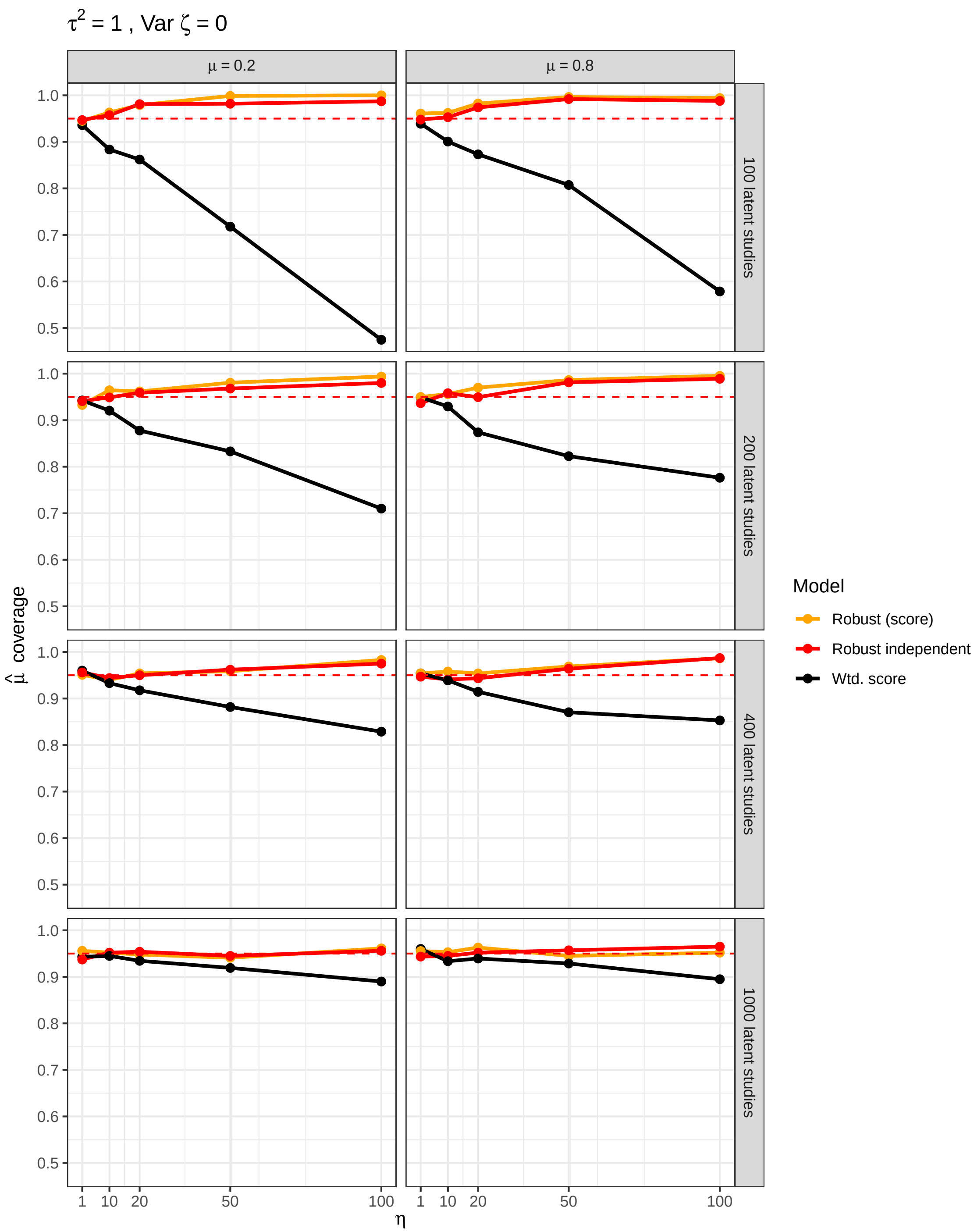


Figure S2: *Median width of confidence interval for $\widehat{\mu}_\eta$ in scenarios without clustering. "Robust (score)": Robust independent model in which $\widehat{\tau}^2$ is chosen by first fitting the weighted score model. "Robust independent": Robust independent model as in the main text, in which $\widehat{\tau}^2$ is chosen by first fitting the naïve parametric model. "Wtd. score": Weighted score model. The y-axis is presented on the log-10 scale with numerical labels on the untransformed scale.*

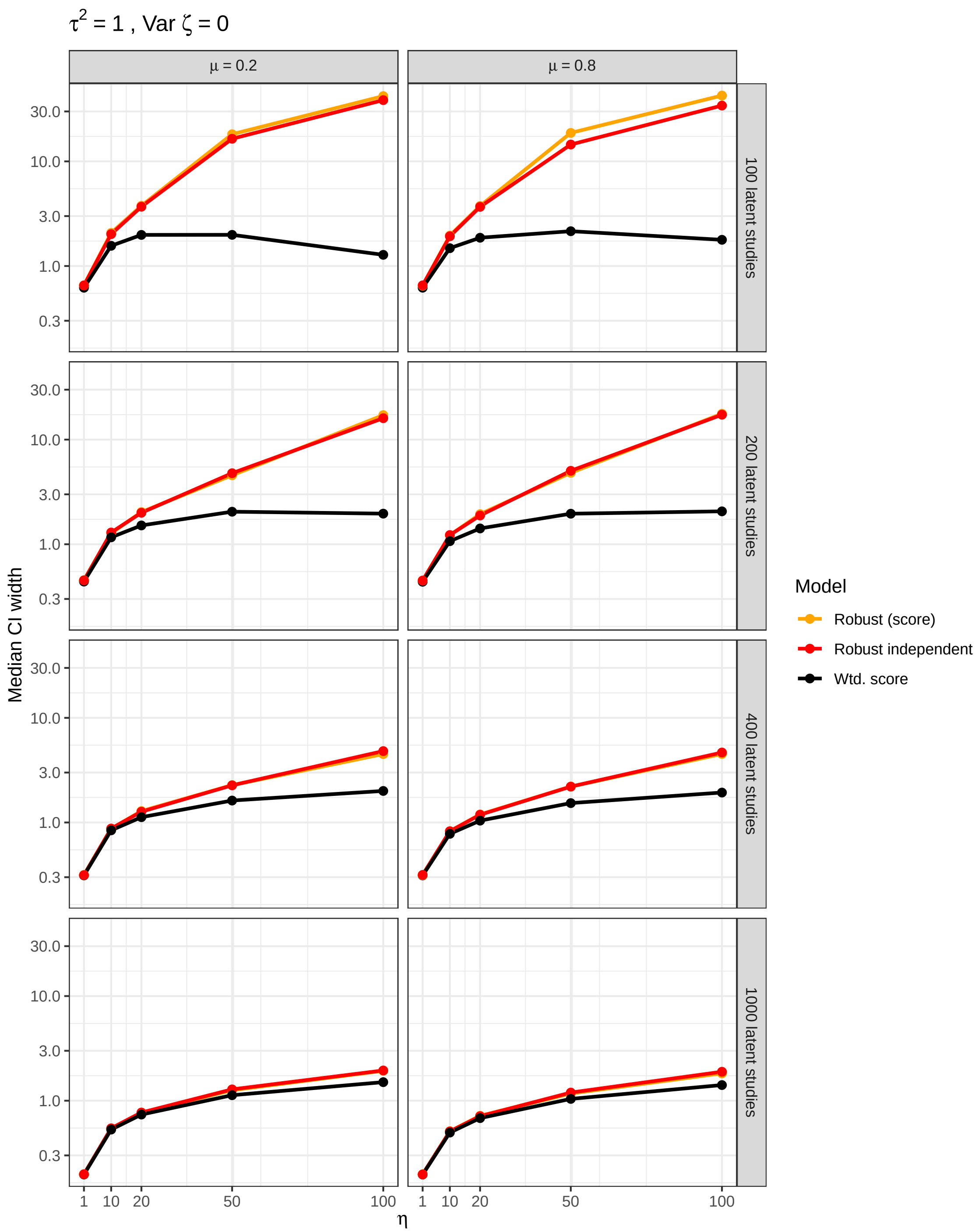


Figure S3: *Mean coverage in scenarios with clustering. "Robust (score)": Robust clustered model in which $\hat{\tau}^2$ is chosen by first fitting the weighted score model. "Robust clustered": Robust clustered model as in the main text, in which $\hat{\tau}^2$ is chosen by first fitting the naïve parametric model. "Wtd. score": Weighted score model.*

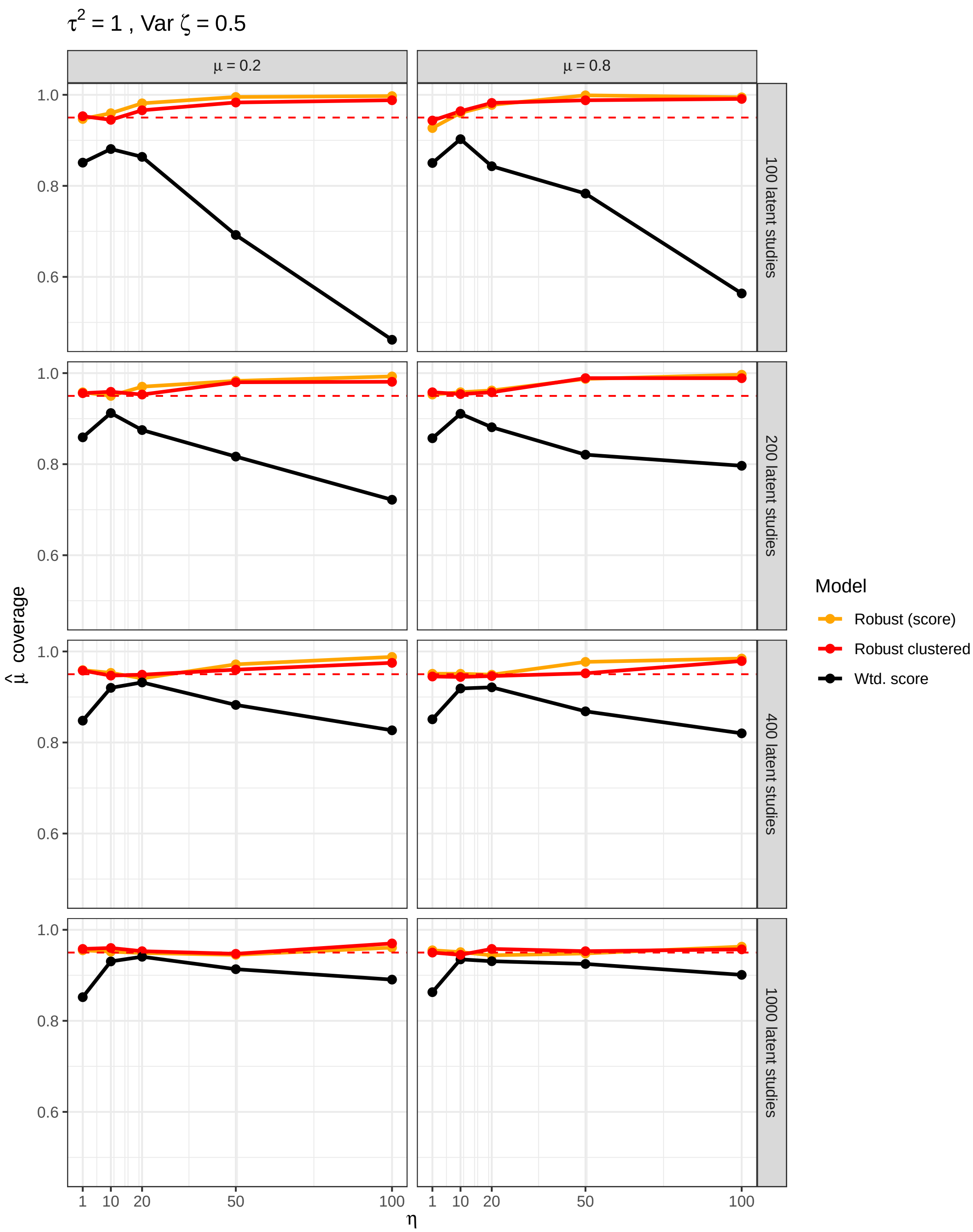
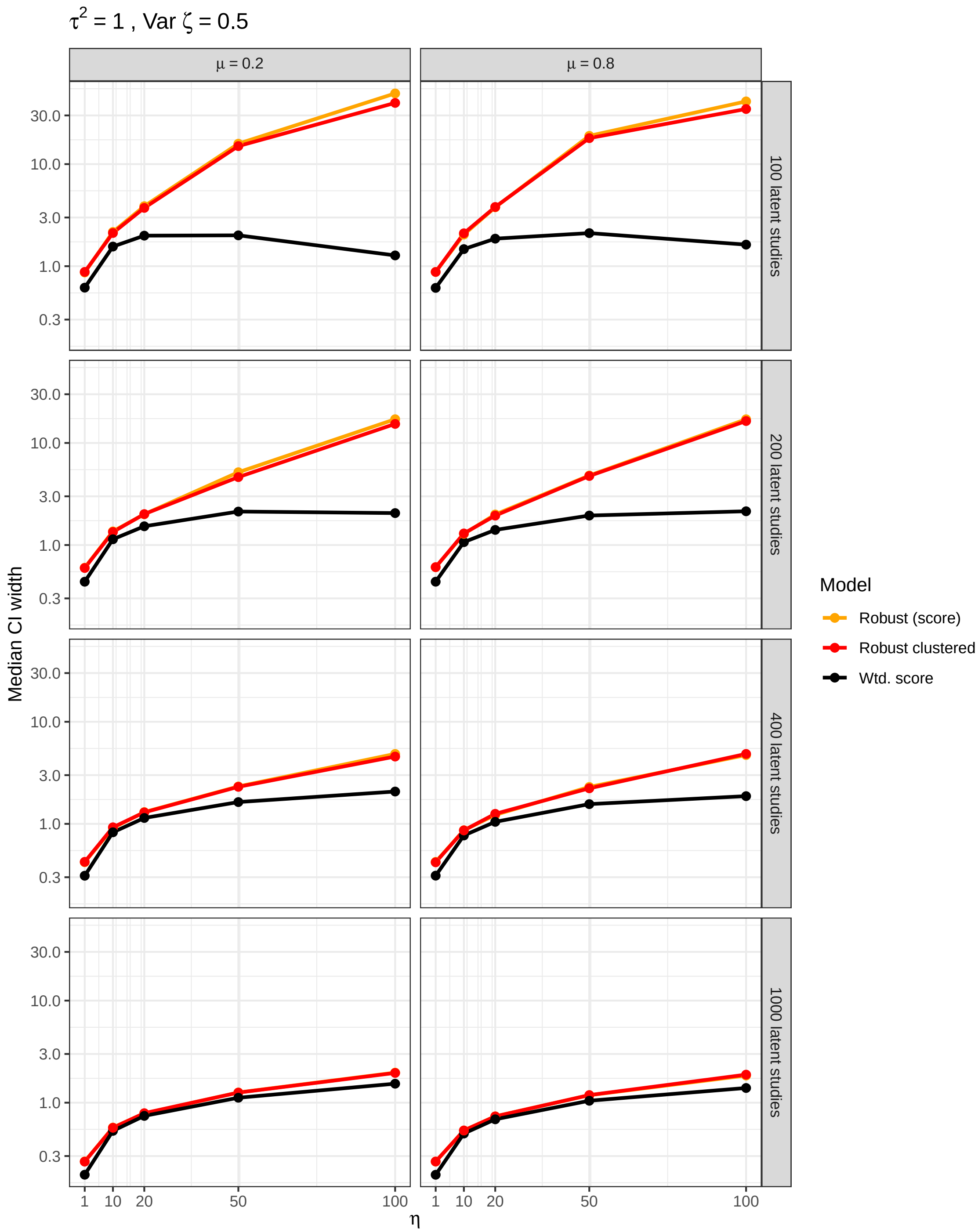


Figure S4: *Median width of confidence interval for $\widehat{\mu}_\eta$ in scenarios with clustering. "Robust (score)": Robust clustered model in which $\widehat{\tau}^2$ is chosen by first fitting the weighted score model. "Robust clustered": Robust clustered model as in the main text, in which $\widehat{\tau}^2$ is chosen by first fitting the naïve parametric model. "Wtd. score": Weighted score model. The y-axis is presented on the log-10 scale with numerical labels on the untransformed scale.*



2. Introduction to the R package PublicationBias

Here we briefly summarize the functions contained in the package `PublicationBias`; details and examples are available in the standard `R` documentation. For a fixed selection ratio η, the function `corrected_meta` estimates a publication bias-corrected pooled point estimate and confidence interval for the common-effect, robust independent, or robust clustered specifications. The function `svalue` estimates $S(t, q)$ for the point estimate and confidence interval limit for a chosen threshold q; it uses analytical results for the common-effect specification and a grid search for the robust specifications. The function `significance_funnel` creates a significance funnel plot. The function `pval_plot` plots studies' one-tailed p-values to help verify assumptions as described in the main text.

References

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, *20*(6), 825–840.

Orwin, R. G. (1983). A fail-safe n for effect size in meta-analysis. *Journal of Educational Statistics*, *8*(2), 157–159.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., … Salanti, G. (2015). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, *141*(2), 1281–1301.