Sensitivity Analysis for Publication Bias in Meta-Analyses

Maya B. Mathur[1] & Tyler J. VanderWeele[2]

[1]Quantitative Sciences Unit, Stanford University, Palo Alto, CA, USA

[2]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA

Citation: Mathur MB & VanderWeele TJ (in press). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C.* Preprint retrieved from: https://osf.io/s9dp6/.

Abstract

We propose sensitivity analyses for publication bias in meta-analyses. We consider a publication process such that "statistically significant" results are more likely to be published than negative or "nonsignificant" results by an unknown ratio, η. Our proposed methods also accommodate some plausible forms of selection based on a study's standard error. Using inverse-probability weighting and robust estimation that accommodates non-normal population effects, small meta-analyses, and clustering, we develop sensitivity analyses that enable statements such as: "For publication bias to shift the observed point estimate to the null, 'significant' results would need to be at least 30-fold more likely to be published than negative or 'nonsignificant' results." Comparable statements can be made regarding shifting to a chosen non-null value or shifting the confidence interval. To aid interpretation, we describe empirical benchmarks for plausible values of η across disciplines. We show that a worst-case meta-analytic point estimate for maximal publication bias under the selection model can be obtained simply by conducting a standard meta-analysis of only the negative and "nonsignificant" studies; this method sometimes indicates that no amount of such publication bias could "explain away" the results. We illustrate the proposed methods using real-life meta-analyses and provide an R package, `PublicationBias`.

Key words: Publication bias; file drawer; meta-analysis; sensitivity analysis

1. Introduction

Publication bias can distort meta-analytic results, sometimes justifying considerable skepticism toward meta-analyses reporting positive findings. Numerous statistical methods, mostly falling into two broad categories, help assess or correct for these biases. First, classical methods arising from the funnel plot (Duval & Tweedie, 2000; Egger et al., 1997) assess whether small studies have systematically larger point estimates than larger studies (see Jin et al. (2015) for a review). These methods effectively assume that publication bias does not operate on very large studies and operates based on the size of the point estimates rather than their p-values. A second class of methods, called selection models, instead assumes that publication bias selects for small p-values rather than large point estimates and allows for publication bias that operates on all studies, not only small ones (see Jin et al. (2015) and McShane et al. (2016) for reviews). These models specify a parametric form for the population effect distribution as well as for the dependence of a study's publication probability on its p-value. The latter weight function may, for example, be specified as a step function such that "affirmative" results (i.e., positive point estimates with a two-tailed $p < 0.05$) are published with higher probability than "nonaffirmative" results (i.e., negative point estimates or those with $p \geq 0.05$) (e.g., Dear & Begg (1992); Hedges (1992); Vevea & Hedges (1995)). Then, after weighting each study's contribution to the likelihood by its inverse-probability of publication per the weight function, the meta-analytic parameters of interest and the parameters of the weight function can be jointly estimated by maximum likelihood. Some relatively recent methods can be viewed as hybrids between classical funnel plot methods and selection models (Bom & Rachinger, 2019; Stanley & Doucouliagos, 2014).

Existing selection models focus on thus estimating bias-corrected meta-analytic estimates as well as the severity of publication bias itself. These models, while quite valuable and informative for large meta-analyses, often yield unstable estimates in meta-analyses of typical sizes (Field & Gillett, 2010; Vevea & Woods, 2005), particularly when the number of upweighted studies (e.g., "affirmative" studies as defined above) is also small. In practice, most meta-analyses are far too small to apply selection models; for example, the percentage of meta-analyses comprising fewer

than 10 studies likely exceeds 50% in the Cochrane database (Ioannidis & Trikalinos, 2007) and in medical journals (Sterne et al., 2000). Selection models may in fact require considerably more than 10 studies to achieve their asymptotic properties (Field & Gillett, 2010). Vevea & Woods (2005) therefore proposed repurposing selection models to conduct sensitivity analyses across a fixed range of parameters that govern the severity of publication bias, rather than attempting to jointly estimate these parameters. If the results of a meta-analysis are only mildly attenuated even under severe publication bias, then the results might be considered fairly robust to publication bias (Vevea & Woods, 2005), while if the results are severely attenuated under mild publication bias, the results might be considered sensitive to publication bias. These methods, like selection models in general, require specifying a parametric form on the population effect distribution, but related work suggests that results can be highly sensitive to this specification (Johnson et al., 2017). Empirical assessment of distributional assumptions may be particularly challenging when the distribution of point estimates is already distorted due to publication bias. Additionally, the method seems to model the sensitivity parameters as if they were observed data, an apparent vestige of the model's original purpose of estimating these parameters rather than conducting sensitivity analysis across fixed parameters (Hedges, 1992). Therefore, when the sensitivity parameters reflect more severe publication bias than actually exists, the corrected point estimate can sometimes be overcorrected. A different approach to sensitivity analysis considers the "fail-safe number", defined as the minimum number of unpublished studies with a mean point estimate of 0 (or another fixed value) that would need to exist in order for their inclusion in the meta-analysis to reduce the pooled point estimate to "statistical nonsignificance" (Rosenthal, 1979) or to a given effect size threshold (Orwin, 1983). These methods typically assume homogenenous population effects (Schmidt & Hunter, 2014) and require specification of the mean of the unpublished studies' point estimates.

Here, we develop sensitivity analyses that advance methodologically upon these existing approaches and also enable particularly simple, intuitive statements regarding sensitivity to publication bias. Methodologically, the present methods relax the distributional and asymptotic assumptions used in existing selection models. That is, our methods require specification of a

simple weight function but do not require any distributional assumptions on the population effects, and they accommodate dependence among the point estimates that can arise when some papers contribute multiple point estimates. Our proposed methods also provide correct inference in meta-analyses with a realistically small total number of studies, as well as with few observed nonaffirmative studies. We will show that, for common-effect meta-analysis (also called "fixed-effects meta-analysis"; Rice et al. (2018)), sensitivity analyses can be conducted in closed form, and for random-effects meta-analysis, they can be conducted through a numerical grid search.

Additionally, our proposed methods enable conclusions that are particularly straightforward to interpret and report, a key consideration for sensitivity analyses to gain widespread use (e.g., VanderWeele & Ding (2017); VanderWeele et al. (2019)). That is, analogously to recent work on sensitivity analysis for unmeasured confounding (VanderWeele & Ding, 2017), the present methods allow statements such as: "In order for publication bias to completely 'explain away' the meta-analytic pooled point estimate (i.e., to attenuate the population effect to the null), affirmative results would need to be at least 30-fold more likely to be published than nonaffirmative results. In order for publication bias to attenuate the confidence interval to include the null, affirmative results would need to be at least 16-fold more likely to be published than nonaffirmative results." Large ratios of publication probabilities would therefore indicate that the meta-analysis is relatively robust to publication bias, whereas small ratios would indicate that the meta-analysis is relatively sensitive to publication bias. We discuss empirical benchmarks for plausible values of these ratios based on a systematic review of existing meta-analyses across several empirical disciplines. We also extend these sensitivity analyses to assess the amount of publication bias required to attenuate the point estimate or its confidence interval to any non-null value, an approach that has been used in existing sensitivity analyses for other forms of bias (e.g., Ding & VanderWeele (2016); VanderWeele & Ding (2017)). These sensitivity analyses apply when the publication process is assumed to favor "statistically significant" and positive results over "nonsignificant" or negative results; the analyses also accommodate some forms of additional selection on studies' standard errors without requiring any modifications.

This paper is structured as follows. We first describe three standard meta-analytic specifications for which we will develop sensitivity analyses (Section 2). We describe our assumed model of publication bias (Section 2.2) and use it to incorporate bias corrections into the three meta-analytic specifications and thus conduct sensitivity analyses (Section 3). We discuss how to interpret the results in practice, providing evidence-based guidelines for plausible values of the sensitivity parameters and suggesting a simple graphical aid to interpretation (Section 6). We illustrate the methods with three applied examples (Section 7) and present a simulation study demonstrating the methods' robustness in challenging scenarios (Section 8). Last, we discuss the plausibility of our assumed model of publication bias and describe how our proposed methods could be extended to accommodate other models (Section 9).

2. Setting and notation

2.1. Common-effect and robust random-effects meta-analysis

Throughout, we consider a meta-analysis of k studies, in which $\widehat{\theta}_i$ and σ_i^2 respectively denote the point estimate and squared standard error of the i^{th} meta-analyzed study. We assume the mean model $\widehat{\theta}_i = \mu + \gamma_i + \epsilon_i$, where $\text{Var}(\epsilon_i) = \sigma_i^2$ (treated as known, as usual in meta-analysis) and $\text{Var}(\gamma_i) = \tau^2$. As usual in meta-analysis, we assume that the point estimates and their standard errors are uncorrelated in the underlying population of studies prior to selection based on publication bias. The **common-effect** specification, arising under the additional assumption that $\tau^2 = 0$ and that the errors, ϵ_i, are independent, estimates μ and its variance as the weighted average of the point estimates:

$$\widehat{\mu} = \left(\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\widehat{\theta}_i\right)\left(\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\right)^{-1} \qquad \widehat{\text{Var}}(\widehat{\mu}) = \left(\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\right)^{-1} \tag{2.1}$$

(If $\tau^2 > 0$, but interest lies in drawing inference only to the sample of meta-analyzed studies rather than to a broader population from which the meta-analyzed studies were drawn, the common-effect estimate and variance in fact remain unbiased with correct nominal coverage

(Rice et al., 2018).) Alternatively, if τ^2 may be greater than 0 and we intend to draw inference to a broader population of studies, then μ and its variance can be estimated under a random-effects model. For this case, we consider a robust estimation approach similar to generalized least squares which, unlike standard parametric random-effects meta-analysis, yields consistent estimates of μ without requiring the usual distributional and independence assumptions on ϵ_i or γ_i (Hedges et al., 2010). Additionally, whereas standard asymptotic inference for parametric random-effects meta-analysis can perform poorly for small k, simple corrections allow the robust specification to perform quite well in small samples (Tipton, 2015); this will become especially important when we consider sensitivity analyses for which the effective sample size is further reduced through inverse-probability weighting.

Specifically, following Hedges et al. (2010), suppose there are M clusters with k_m estimates in the m^{th} cluster. For example, each cluster might represent a paper that potentially contributes to the meta-analysis multiple, statistically independent point estimates arising from a hierarchical structure in which different papers have different mean effect sizes due, for example, to the use of similar subject populations. Alternatively, each cluster might represent a paper for which there is a single population effect size, but in which point estimates are statistically dependent because they are estimated in overlapping groups of subjects (Hedges et al., 2010). Arbitrary other correlation structures are also possible, and importantly, as in generalized estimating equations with robust inference, correct prespecification of the correlation structure allows optimal efficiency but is not required to achieve correct inference asymptotically or in finite samples with a small-sample correction (Hedges et al., 2010). The usual meta-analytic assumption of independent point estimates represents the special case with $M = k$, $k_m = 1$ for all m, and an inter-cluster correlation of 0; we term this the **"robust independent"** specification.

Let $\left(\widehat{\theta}_{m1}, \cdots, \widehat{\theta}_{mk_m}\right)'$ be the vector of point estimates for studies in cluster m. For the general case, which we will term the **"robust clustered"** specification, the point estimates may be arbitrarily dependent within clusters, but are assumed to be independent across clusters. That is, let $\Sigma_m \in \mathbb{R}^{k_m \times k_m}$ denote the within-cluster covariance matrix of the study-level error terms, $(\epsilon_{1m}, \cdots, \epsilon_{mk_m})$, and let $\Sigma = \text{diag}\left(\Sigma_1, \cdots, \Sigma_M\right)$ denote the overall covariance matrix. Let

$\mathbf{W}_m \in \mathbb{R}^{k_m \times k_m}$ be a diagonal matrix of arbitrary positive weights for studies in cluster m, such that the i^{th} study in cluster m has weight w_{mi}. Let $\mathbf{1}_{k_m}$ denote the 1-vector of length k_m. Then, as a direct special case of Hedges et al. (2010)'s Equations (3) and (6), a consistent estimate of μ and its exact variance are:

$$\widehat{\mu} = \left(\sum_{m=1}^{M}\sum_{i=1}^{k_m} w_{mk_m}\widehat{\theta}_{mk_m}\right)\left(\sum_{m=1}^{M}\sum_{i=1}^{k_m} w_{mk_m}\right)^{-1} \tag{2.2}$$

$$\text{Var}\left(\widehat{\mu}\right) = \left(\sum_{m=1}^{M}\mathbf{1}'_{k_m}\mathbf{W}_m\Sigma_m\mathbf{W}_m\mathbf{1}_{k_m}\right)\left(\sum_{m=1}^{M}\sum_{i=1}^{k_m} w_{mk_m}\right)^{-2}$$

Letting $\mathbf{e}_m = \left(\widehat{\theta}_{m1} - \widehat{\mu}, \cdots, \widehat{\theta}_{mk_m} - \widehat{\mu}\right)'$ denote the vector of residuals for studies in cluster m, an asymptotic plug-in estimate of the variance is:

$$\widehat{\text{Var}}\left(\widehat{\mu}\right) = \frac{M}{M-1}\left(\sum_{m=1}^{M}\mathbf{1}'_{k_m}\mathbf{W}_m\mathbf{e}_m\mathbf{e}'_m\mathbf{W}_m\mathbf{1}_{k_m}\right)\left(\sum_{m=1}^{M}\sum_{i=1}^{k_m} w_{mk_m}\right)^{-2} \tag{2.3}$$

(This model and the corresponding sensitivity analyses developed below also extend readily to meta-regression; see Hedges et al. (2010) for details. We focus on the intercept-only model for brevity.) In all subsequent work, we will use Tipton (2015)'s finite-sample correction to this variance estimator, which can be easily applied in R by fitting the model with the `robumeta` package (Fisher & Tipton, 2015) with the argument `small=TRUE`. We next describe our assumed model of publication bias and develop sensitivity analyses for the common-effect, robust independent, and robust clustered specifications.

2.2. Assumed model of publication bias

We consider a mechanism of publication bias in which studies are selected for publication from among an underlying population of all published and unpublished studies, and the probability of selection is higher for "affirmative" (defined by $\widehat{\theta} > 0$ and $p < 0.05$) versus "nonaffirmative" studies (defined by $\widehat{\theta} \leq 0$ or $p \geq 0.05$). (Throughout, we assume that publication favors point estimates in the positive direction and that the uncorrected $\widehat{\mu}$ is positive. As described in

Section 4, if publication instead favors results with negative point estimates and $\widehat{\mu} < 0$, one can simply reverse the sign of all point estimates and of the meta-analytic summary statistics prior to conducting our proposed analyses. Additionally, if there is a strong reason to believe that publication bias operates based on an α level other than 0.05 for a given meta-analysis, for example because of disciplinary conventions, the definition of affirmative status can simply be generalized to require $\widehat{\theta} > 0$ and $p < \alpha$ with no further changes to the following results. This model of publication bias is common in existing work (see McShane et al. (2016) for a review), and we discuss in detail its plausibility in Section 9 and discuss extensions to other models of publication bias. Suppose the underlying population is of size $k^* \geq k$, with the point estimate and standard error in the i^{th} underlying study denoted $\widehat{\theta}_i^*$ and σ_i^*, respectively. Let p_i^* denote the p-value of underlying study i and $A_i^* = \mathbb{1}\{\widehat{\theta}_i^* > 0 \text{ and } p_i^* < 0.05\}$ be an indicator for whether underlying study i is affirmative. Let w_i^* denote an additional unstandardized, common-effect or random-effects inverse-variance weight; for example, for common-effect meta-analysis, $w_i^* = (\sigma_i^*)^{-2}$. Let D_i^* be an indicator for whether underlying study i is in fact published. Then, we assume the publication process arises as:

$$P\left(D_i^* = 1 \mid A_i^*\right) \propto \eta^{-1}\mathbb{1}\{A_i^* = 0\} + \mathbb{1}\{A_i^* = 1\} \qquad \text{where } \eta \geq 1 \tag{2.4}$$

$$E\left[D_i^* w_i^* \mid A_i^*\right] = E\left[D_i^* \mid A_i^*\right] E\left[w_i^* \mid A_i^*\right] \tag{2.5}$$

$$E\left[D_i^* w_i^* \widehat{\theta}_i^* \mid A_i^*\right] = E\left[D_i^* \mid A_i^*\right] E\left[w_i^* \widehat{\theta}_i^* \mid A_i^*\right] \tag{2.6}$$

That is, we assume that the publication probability is η times higher for affirmative studies than for nonaffirmative studies, where we will treat η as an unknown sensitivity parameter. The assumptions regarding uncorrelatedness state that, conditional on a study's affirmative or nonaffirmative status, publication bias does not select further based on the inverse-variance weights nor on the product of the point estimates with their inverse-variance weights. For example, this assumption excludes the possibility that the publication process would favor studies with larger point estimates, smaller standard errors, or smaller p-values above and beyond its

favoring of affirmative results. However, in the Supplement (Section 1.1), we show that these assumptions can in fact be relaxed to accommodate a publication process that additionally favors studies with smaller standard errors (for example) as long as this form of selection operates in the same way for affirmative and for nonaffirmative studies. As we show in the Supplement, this additional form of selection can simply be ignored, requiring no modification to the sensitivity analyses we present below.

Additionally, this model is agnostic to the reasons for preferential publication of affirmative studies, which likely reflects a complex combination of authors' selective reporting of results (Chan et al., 2004; Coursol & Wagner, 1986) and selective submission of papers (Franco et al., 2014; Greenwald, 1975), as well as editors' and reviewers' biases (although some empirical work suggests that the latter may be a weaker influence than author decision-making; see Lee et al. (2006); Olson et al. (2002)). That is, we technically conceptualize the underlying population as the population of all conducted hypothesis tests that would, if published, be included in the meta-analysis. For example, suppose various researchers conduct and publish 10 studies on the topic of the meta-analysis, each involving two separate experiments on independent samples (all of which would be included in the meta-analysis if published). If each paper includes or omits the results of these two experiments based on the selection process defined by Equations (2.4)-(2.6), the underlying population would comprise the 20 total hypothesis tests. For brevity, though, we refer to a underlying population of "studies" and the aggregation of all selection effects conforming to the assumptions in Equations (2.4)-(2.6) as "publication bias".

Last, note that our definition of A_i^* assumes that the publication process preferentially selects studies with point estimates that are both "statistically significant" and positive in direction. Studies with "nonsignificant" point estimates and those with "statistically significant" negative point estimates are equally disfavored. As in existing literature (Vevea & Hedges, 1995), we refer to this as "one-tailed selection" in contrast to an alternative "two-tailed selection" model in which publication favors all studies with "significant" point estimates, regardless of direction. Note that the term "one-tailed" does not imply that studies report one-tailed hypothesis tests, but rather that publication selects in a one-tailed manner based on reported statistical results.

3. Main results

3.1. Sensitivity analysis under the common-effect specification

3.1.1 Publication bias required to attenuate the point estimate or its lower confidence interval limit to a chosen value

Under publication bias as described above, the naïve common-effect estimate $\widehat{\mu}$ will usually be biased for μ if $\eta > 1$. In this section, we first present consistent estimators $\widehat{\mu}_\eta$ and $\widehat{\text{Var}}\left(\widehat{\mu}_\eta\right)$ under a fixed ratio, η, which we then use to derive sensitivity analyses characterizing the value of η required to attenuate $\widehat{\mu}$ or its lower confidence limit, $\widehat{\mu}^{lb}$, to a chosen smaller value, q. When referring to the published and meta-analyzed, rather than underlying, studies, we use notation as in Section 2.2 above but without the "*" superscript. For example, $A_i = \mathbb{1}\{\widehat{\theta}_i > 0 \text{ and } p_i < 0.05\}$ is an indicator for whether observed study i is affirmative. Let $\mathcal{A} = \{i : A_i = 1\}$ and $\mathcal{N} = \{i : A_i = 0\}$ respectively be the sets of published, meta-analyzed affirmative and nonaffirmative studies. For an arbitrary subset of studies $\mathcal{S}$, define $\bar{y}_\mathcal{S} = \sum_{i\in\mathcal{S}} \frac{1}{\sigma_i^2}\widehat{\theta}_i$ and $\nu_\mathcal{S} = \sum_{i\in\mathcal{S}} \frac{1}{\sigma_i^2}$, such that $\bar{y}_\mathcal{S}/\nu_\mathcal{S}$ is the usual common-effect estimate for the studies in $\mathcal{S}$. Then, for a fixed η, consistent estimates of μ and its variance can be obtained under mild regularity conditions by weighting each study inversely to its publication probability (Supplement, Section 1.1, Theorem 1.1):

$$\widehat{\mu}_\eta = \left(\sum_{i=1}^{k} \frac{\eta^{\mathbb{1}\{A_i=0\}}}{\sigma_i^2}\widehat{\theta}_i\right)\left(\sum_{i=1}^{k} \frac{\eta^{\mathbb{1}\{A_i=0\}}}{\sigma_i^2}\right)^{-1} = \frac{\eta\bar{y}_\mathcal{N} + \bar{y}_\mathcal{A}}{\eta\nu_\mathcal{N} + \nu_\mathcal{A}} \tag{3.1a}$$

$$\widehat{\text{Var}}\left(\widehat{\mu}_\eta\right) = \frac{1}{\left(\eta\nu_\mathcal{N} + \nu_\mathcal{A}\right)^2}\widehat{\text{Var}}\left(\eta\bar{y}_\mathcal{N} + \bar{y}_\mathcal{A}\right) = \frac{\eta^2\nu_\mathcal{N} + \nu_\mathcal{A}}{\left(\eta\nu_\mathcal{N} + \nu_\mathcal{A}\right)^2} \tag{3.1b}$$

Because η is not known in practice, we develop these estimators only as means to the end of deriving sensitivity analyses as follows. For a meta-analytic effect estimate t (either $\widehat{\mu}$ or $\widehat{\mu}^{lb}$), define $S\left(t, q\right)$ as the value of η that would attenuate t to q, where $q < t$. For example, $S\left(\widehat{\mu}, q\right)$ is the value of η that would attenuate the point estimate to $q < \widehat{\mu}$, and its derivation

for common-effect meta-analysis follows directly from Equation (3.1a):

$$S\left(\widehat{\mu}, q\right) = \frac{\nu_{\mathcal{A}} q - \bar{y}_{\mathcal{A}}}{\bar{y}_{\mathcal{N}} - \nu_{\mathcal{N}} q}$$

If, as usual, the point estimates are meta-analyzed on a scale for which the null is 0, then for the special case of attenuating the point estimate to the null, we have $S\left(\widehat{\mu}, 0\right) = -\bar{y}_{\mathcal{A}}/\bar{y}_{\mathcal{N}}$. That is, to attenuate the point estimate to the null, affirmative studies would need to be more likely to be published than non-affirmative studies by the same ratio by which the magnitude of $\bar{y}_{\mathcal{A}}$ exceeds its counterpart in the non-affirmative studies, $\bar{y}_{\mathcal{N}}$. For example, if $\bar{y}_{\mathcal{A}}$ is 5-fold larger in magnitude and in the opposite direction from $\bar{y}_{\mathcal{N}}$, then to attenuate $\widehat{\mu}$ to the null, affirmative studies would need to be more likely to be published than non-affirmative studies by 5-fold.

To consider the severity of publication bias required to attenuate the lower 95% confidence limit of $\widehat{\mu}$ to q, we set q equal to the corrected confidence limit estimate using the variance estimate in Equation (3.1b). Letting c_{crit} denote the two-sided critical value for the t distribution on $k-1$ degrees of freedom, we thus have:

$$S\left(\widehat{\mu}^{lb}, q\right) = \frac{\pm c_{\text{crit}}\sqrt{B} - \bar{y}_{\mathcal{N}}\bar{y}_{\mathcal{A}} + \bar{y}_{\mathcal{N}}\nu_{\mathcal{A}} q + \bar{y}_{\mathcal{A}}\nu_{\mathcal{N}} q - \nu_{\mathcal{N}}\nu_{\mathcal{A}} q^2}{\bar{y}_{\mathcal{N}}^2 - 2\bar{y}_{\mathcal{N}}\nu_{\mathcal{N}}\ q + \nu_{\mathcal{N}}^2 q^2 - c_{\text{crit}}^2\nu_{\mathcal{N}}}$$

where $B = \bar{y}_{\mathcal{N}}^2\nu_{\mathcal{A}} - \left(2\nu_{\mathcal{N}}\nu_{\mathcal{A}} q\right)\left(\bar{y}_{\mathcal{N}} + \bar{y}_{\mathcal{A}}\right) + \bar{y}_{\mathcal{A}}^2\nu_{\mathcal{N}} + q^2\left(\nu_{\mathcal{N}}^2\nu_{\mathcal{A}} + \nu_{\mathcal{A}}^2\nu_{\mathcal{N}}\right) - c_{\text{crit}}^2\nu_{\mathcal{N}}\nu_{\mathcal{A}}$. When $q = 0$, this simplifies to:

$$S\left(\widehat{\mu}^{lb}, 0\right) = \frac{\pm c_{\text{crit}}\sqrt{\bar{y}_{\mathcal{N}}^2\nu_{\mathcal{A}} + \bar{y}_{\mathcal{A}}^2\nu_{\mathcal{N}} - c_{\text{crit}}^2\nu_{\mathcal{N}}\nu_{\mathcal{A}}} - \bar{y}_{\mathcal{N}}\bar{y}_{\mathcal{A}}}{\bar{y}_{\mathcal{N}}^2 - c_{\text{crit}}^2\nu_{\mathcal{N}}}$$

Note that, in general, it is possible to obtain $S\left(t, q\right) < 1$, which means that no amount of publication bias under the model described above, no matter how severe, could attenuate t to q. Additionally, $S\left(t, q\right)$ can be reparameterized to provide an alternative metric in terms of the number of unpublished nonaffirmative studies. Specifically, $S\left(t, q\right)$ can help provide an approximate lower bound on a form of "fail-safe number", here defined as the number of unpublished nonaffirmative results that would be required to shift the statistic t to q (see

Supplement, Section 1.3).

3.1.2 Simple estimates under worst-case publication bias

In addition to solving for the value of η required to attenuate $\widehat{\mu}$ and $\widehat{\mu}^{lb}$ to specific values, it is also straightforward to estimate $\widehat{\mu}_\eta$ and $\widehat{\mu}^{lb}_\eta$ for worst-case publication bias under the model we have assumed. That is, letting $\eta \to \infty$ in Equations (3.1a) and (3.1b), we have:

$$\lim_{\eta\to\infty} \widehat{\mu}_\eta = \lim_{\eta\to\infty} \frac{\bar{y}_{\mathcal{N}} + \eta^{-1}\bar{y}_{\mathcal{A}}}{\nu_{\mathcal{N}} + \eta^{-1}\nu_{\mathcal{A}}} = \bar{y}_{\mathcal{N}}/\nu_{\mathcal{N}}$$

$$\lim_{\eta\to\infty} \widehat{\mathrm{Var}}\left(\widehat{\mu}_\eta\right) = \lim_{\eta\to\infty} \frac{\nu_{\mathcal{N}} + \eta^{-2}\nu_{\mathcal{A}}}{\nu_{\mathcal{N}}^2 + 2\eta^{-1}\nu_{\mathcal{A}} + \eta^{-2}\nu_{\mathcal{A}}^2} = \nu_{\mathcal{N}}^{-1}$$

Since these expressions coincide with the usual common-effect estimates within only the non-affirmative studies (c.f. Equation (2.1)), worst-case estimates under the assumed model can be obtained simply by meta-analyzing only the nonaffirmative studies. Of course, the presence of at least one nonaffirmative study in the meta-analysis implies that $\eta < \infty$ in practice, but these estimates may nevertheless serve as useful heuristics to approximate the effect of extreme publication bias.

3.2. Sensitivity under the robust random-effects specifications

3.2.1 Publication bias required to attenuate the point estimate or its lower confidence interval limit to a chosen value

We first consider the robust clustered specification and then describe the robust independent specification as a special case. As in Section 2, suppose there are M known clusters of point estimates. For the i^{th} study in the m^{th} cluster, assign the weight $w_{mi} = \eta^{\mathbb{1}\{A_{mi}=0\}}\left(\sigma_i^2 + \widehat{\tau}^2\right)^{-1}$, so that the cluster-$m$ weight matrix $\mathbf{W}_m$ is diagonal with entries equal to the product of the usual random-effects inverse-variance weights with the inverse-probabilities of publication for each study. For the purpose of defining these weights, we recommend simply obtaining a naïve parametric estimate of $\widehat{\tau}^2$ in an initial meta-analysis under the standard parametric

random-effects specification (e.g., Brockwell & Gordon (2001)) without correction for publication bias. Although this estimate may be biased due to publication bias, extreme clustering, or non-normality, this bias does not compromise point estimation or inference on μ, but rather may only somewhat reduce efficiency[1] because the robust specification provides unbiased estimation and valid inference regardless of the choice of $\mathbf{W}$ (Hedges et al., 2010).

Then, from Equations (2.2) and (2.3), we can consistently estimate μ and $\text{Var}(\mu)$ for the **robust clustered** specification as (Supplement, Section 1.1, Theorem 1.1):

$$\widehat{\mu}_\eta = \sum_{i=1}^{k} \eta^{\mathbb{1}\{A_{mi}=0\}} \left(\sigma_i^2 + \widehat{\tau}^2\right)^{-1} \widehat{\theta}_i \left(\sum_{i=1}^{k} \eta^{\mathbb{1}\{A_{mi}=0\}} \left(\sigma_i^2 + \widehat{\tau}^2\right)^{-1}\right)^{-1} \tag{3.2a}$$

$$\widehat{\text{Var}}\left(\widehat{\mu}_\eta\right) = \frac{M}{M-1} \sum_{m=1}^{M} \mathbf{1}_{k_m}' \mathbf{W}_m \mathbf{e}_m \mathbf{e}_m' \mathbf{W}_m \mathbf{1}_{k_m} \left(\sum_{i=1}^{k} \eta^{\mathbb{1}_{A_{mi}=0}} \left(\sigma_i^2 + \widehat{\tau}^2\right)^{-1}\right)^{-2} \tag{3.2b}$$

(Recall that we have assumed $\mathbf{W}_m$ is chosen to be diagonal, so the double summation over clusters and individual studies in Equation (2.2) reduces to a single summation over individual studies in Equation (3.2a), and similarly for Equations (2.3) and (3.2b). Potential correlation of estimates within clusters is accommodated through the non-diagonal sandwich matrix, $\mathbf{e}_m \mathbf{e}_m'$, in Equation 3.2b.) For the **robust independent** specification, $\widehat{\mu}_\eta$ is identical, and $\widehat{\text{Var}}\left(\widehat{\mu}_\eta\right)$ simplifies to:

$$\widehat{\text{Var}}\left(\widehat{\mu}_\eta\right) = \frac{k}{k-1} \sum_{i=1}^{k} \left[\left(\widehat{\theta}_i - \widehat{\mu}_\eta\right) \eta^{\mathbb{1}\{A_{mi}=0\}} \left(\sigma_i^2 + \widehat{\tau}^2\right)^{-1}\right]^2 \left(\sum_{i=1}^{k} \eta^{\mathbb{1}\{A_{mi}=0\}} \left(\sigma_i^2 + \widehat{\tau}^2\right)^{-1}\right)^{-2} \tag{3.3}$$

As described in Section 2, these are asymptotic variance estimates, and in practice we recommend using Tipton (2015)'s small-sample correction. To approximate $S\left(\widehat{\mu}, q\right)$ or $S\left(\widehat{\mu}^{lb}, q\right)$, one can simply evaluate $\widehat{\mu}_\eta$ and $\widehat{\text{Var}}\left(\widehat{\mu}_\eta\right)$ over a grid of values of η, for example by passing user-specified

[1] To explore whether improving the initial estimate $\widehat{\tau}^2$ could improve efficiency in practice, we derived a parametric weighted score approach to jointly estimate $\widehat{\mu}_\eta$ and a bias-corrected $\widehat{\tau}_\eta^2$ under independence. We assessed whether using this improved $\widehat{\tau}_\eta^2$ in the weights would improve performance of the robust specification (Supplement, Section 1.4). However, the resulting $\widehat{\tau}_\eta^2$ was quite biased except in very large samples, so its use did not noticeably affect efficiency. We therefore recommend simply using a naïve $\widehat{\tau}^2$ estimate.

weights to the existing R package `robumeta` (Fisher & Tipton, 2015). Then, $S(t, q)$ can be set to the smallest value of η such that $t \leq q$. Our R package `PublicationBias` automates this approach, and we illustrate further in the applied examples of Section 7.

As an alternative to the robust independent specification, it would be possible to conduct maximum-likelihood sensitivity analyses under the standard parametric random-effects model, invoking the additional assumptions that, among 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 because it should, in principle, be more efficient than the robust specification, and it would also enable direct estimation of τ^2. In the Supplement (Section 1.4), we derive a parametric specification that is directly analogous to inverse-probability weighting for survey sampling or missing data for general M-estimators (Wooldridge, 2007). However, simulation results indicated that this model performed fairly poorly for moderate and large values of η (Supplement, Section 1.4).

3.2.2 Simple estimates under worst-case publication bias

As seen for the common-effect specification, corrected estimates for worst-case publication bias under our assumed model can be obtained by conducting a standard, uncorrected meta-analysis of only the nonaffirmative studies. For the robust clustered specification, letting $\eta \to \infty$ in Equations (3.2a) and (3.2b) yields:

$$\lim_{\eta \to \infty} \widehat{\mu}_\eta = \sum_{i \in \mathcal{N}} \left(\sigma_i^2 + \widehat{\tau}^2\right)^{-1} \widehat{\theta}_i \left(\sum_{i \in \mathcal{N}} \left(\sigma_i^2 + \widehat{\tau}^2\right)^{-1}\right)^{-1}$$

$$\lim_{\eta \to \infty} \widehat{\text{Var}}\left(\widehat{\mu}_\eta\right) = \frac{M}{M-1} \sum_{m=1}^{M} \mathbf{1}'_{|\mathcal{N}_m|} \widetilde{\mathbf{W}}_{\mathcal{N}_m} \mathbf{e}_{\mathcal{N}_m} \mathbf{e}'_{\mathcal{N}_m} \widetilde{\mathbf{W}}_{\mathcal{N}_m} \mathbf{1}_{|\mathcal{N}_m|} \left(\sum_{i \in \mathcal{N}} \left(\sigma_i^2 + \widehat{\tau}^2\right)^{-1}\right)^{-2}$$

where $\mathcal{N}_m$ denotes the set of nonaffirmative studies in cluster m, $|\mathcal{N}_m|$ denotes the number of studies in that set, and $\widetilde{\mathbf{W}}_{\mathcal{N}_m}$ denotes a modified diagonal weight matrix for cluster m in which $w_{mi} = (\sigma_i^2 + \widehat{\tau}^2)^{-1}$. Again, these expressions correspond to robustly meta-analyzing only the nonaffirmative studies (c.f. Equations (2.2) and (2.3)).

4. Additional practical considerations

4.1. Preparing data for analysis

For all three specifications above, the point estimates $\widehat{\theta}_i$ should be analyzed on a scale such that, conditional on their potentially non-normal true population effects γ_i, the point estimates are asymptotically approximately normal with variances σ_i^2. This is standard practice in meta-analysis. For example, estimates on the hazard ratio scale can be transformed to the log-hazard ratio scale for analysis, in which case the threshold q would also be chosen on the log scale. Additionally, our definition of affirmative studies assumes that estimates' signs are coded such that positive estimates are favored in the publication process. For meta-analyses in which negative, rather than positive, estimates are assumed to be favored (e.g., because negative estimates represent a protective effect of a candidate treatment), one can simply reverse the signs of the point estimates $\widehat{\theta}_i$ before conducting our sensitivity analyses and then reverse the sign of the bias-corrected $\widehat{\mu}_\eta$ once more to recover the original sign convention. Note that $S\left(\widehat{\mu}^{lb}, q\right)$, with $\widehat{\mu}^{lb}$ estimated after reversing signs for analysis, would be interpreted according the original sign convention as the severity of publication bias required to shift the *upper*, rather than the lower, confidence interval limit to the null. For example, if the meta-analytic estimate with the original sign convention is a hazard ratio of 0.79 (95% CI: [0.71, 0.88]), and we expect publication bias to favor protective effects, we would first transform the hazard ratios to the log scale and reverse their signs to estimate $S\left(\widehat{\mu}^{lb}, 0\right)$. When interpreted according to the original sign convention and on the hazard ratio scale, $S\left(\widehat{\mu}^{lb}, 0\right)$ would represent the severity of publication bias required to shift the upper confidence interval limit of 0.88 to 1. We illustrate conducting sensitivity analyses with effect-size transformations and sign reversals in the applied examples of Section 7.

4.2. Diagnostics regarding statistical assumptions

The possibility of certain violations of statistical assumptions described in Section 2.2 can be assessed as follows. To investigate the possibility that publication bias might favor "significant"

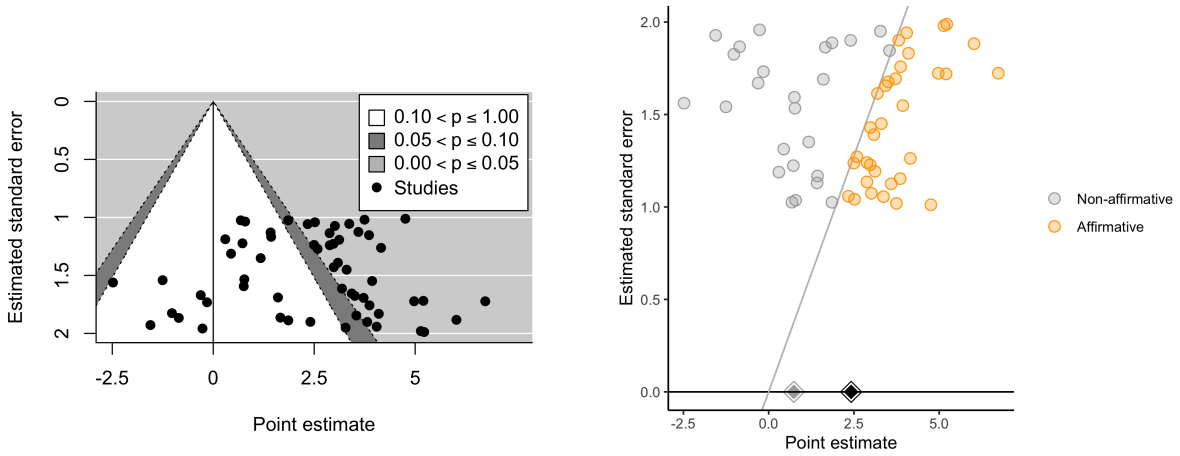
results regardless of the sign of the point estimate, rather than only affirmative results, one could calculate one-tailed p-values for all studies and examine a histogram or density plot of these one-tailed p-values (e.g., using the R function `PublicationBias::pval_plot`). If publication bias favors "significant" results regardless of sign, one would expect to see an increased density of one-tailed p-values not only below 0.025 but also above 0.975 (because the latter corresponds to two-tailed p-values less than 0.05 with negative point estimates). To investigate whether publication bias might select based on multiple α-levels (e.g., $\alpha = 0.10$ as well as $\alpha = 0.05$), one could examine a similar density plot of two-tailed p-values for evidence of clear discontinuities at p-values other than 0.05. The assumption that point estimates are not correlated with their standard errors in the underlying population is harder to assess because sufficiently strong publication bias that favors affirmative results will itself induce this correlation among the *published* studies, even if the assumption regarding the underlying population does hold. Thus, this assumption will generally need to be evaluated in terms of *a priori* plausibility than in terms of empirical evidence.

5. The significance funnel plot

As a visual supplement to the proposed sensitivity analyses, we suggest presenting a modified funnel plot, the "significance funnel", which shares some features with other modified funnel plots, such as the contour-enhanced funnel plot (Andrews & Kasy, 2019; Peters et al., 2008; Vevea et al., 1993). Like a standard funnel plot with or without contour enhancement, the significance funnel plot displays $\widehat{\theta}_i$ versus σ_i^2 or σ_i (e.g., Figure 1b). Whereas a standard funnel helps detect correlation between $\widehat{\theta}_i$ and σ_i, the significance funnel helps detect the extent to which the nonaffirmative studies' point estimates are systematically smaller than the entire set of point estimates, which is the more relevant consideration under our assumed model of publication bias. The significance funnel distinguishes visually between affirmative studies (orange points) and nonaffirmative studies (gray points) and also displays the point estimates within all studies (black diamond) and within only the nonaffirmative studies (gray diamond). As discussed above,

the latter represents the corrected estimate for worst-case publication bias under the assumed model. Thus, as a simple heuristic, when the diamonds are close to one another, our quantitative sensitivity analyses will typically indicate that the meta-analysis is fairly robust to publication bias. When the diamonds are distant or if the gray diamond represents a negligible effect size, then our sensitivity analyses may indicate that the meta-analysis is not robust. Of course, conducting formal sensitivity analyses and reporting $\eta(\widehat{\mu}, q)$ and $\eta(\widehat{\mu}^{lb}, q)$ for $q = 0$ and possibly another non-null value provides more precise information than presenting the significance funnel alone.

Even though the significance funnel and the standard funnel display the same data, the former better complements sensitivity analyses for publication bias. Indeed, the standard funnel can be quite misleading when publication bias operates on p-values. For example, the standard contour-enhanced funnel in Figure 1a shows right-skewed point estimates generated with publication bias ($\eta = 10$), but suggests little correlation between the estimates and standard errors, giving an impression of robustness to publication bias under standard criteria. Yet the significance funnel (Figure 1b) shows that the nonaffirmative studies in fact have much smaller point estimates than the affirmative studies, correctly suggesting that results may in fact be sensitive to publication bias.



(a) **(b)**

Figure 1: *Standard contour-enhanced funnel plot (left column; Peters et al. (2008); Viechtbauer (2010)) versus significance funnel plot (right column) for data generated with publication bias and with right-skewed population effect sizes (bottom row; $\eta = 10$). Studies lying on the diagonal line have exactly $p = 0.05$. Black diamond: robust independent point estimate within all studies; gray diamond: robust independent point estimate within only the nonaffirmative studies.*

6. Empirical benchmarks for interpreting $S(t, q)$

Interpreting our proposed sensitivity analyses ultimately involves assessing whether $S(t, q)$ is small enough that it represents a plausible amount of publication bias, in which case the meta-analysis may be considered relatively sensitive to publication bias; or conversely whether it represents an implausibly large amount of publication bias, in which case the meta-analysis may be considered relatively robust. To help empirically ground such assessments, we conducted a preregistered meta-meta-analysis to estimate the actual value of η in an objectively chosen sample of meta-analyses across several scientific disciplines. Detailed methods and results are provided in Mathur & VanderWeele (2020a).

We systematically searched for meta-analyses from four sources: (1) *PLOS One*; (2) four top medical journals; (3) three top experimental psychology journals; and (4) Metalab, an online, unpublished repository of meta-analyses on developmental psychology. Metalab is a database of meta-analyses on developmental psychology whose datasets are made publicly available and are continuously updated; these meta-analyses are often released online prior to publication in

peer-reviewed journals (Bergmann et al., 2018; Lewis et al., 2016). We selected these sources in order to represent a range of disciplines, particularly via the inclusion of *PLOS One* meta-analyses. Our inclusion criteria were: (1) the meta-analysis comprised at least 40 studies to enable reasonable power and asymptotic properties to estimate publication bias (Hedges, 1992); (2) the meta-analysis included at least 3 affirmative studies and 3 nonaffirmative studies to minimize problems of statistical instability (Hedges, 1992); (2) the meta-analyzed studies tested a hypothesis (e.g., they were not purely descriptive); (3) we could obtain study-level point estimates and standard errors. This search yielded a total of 58 analyzed meta-analyses (30 from *PLoS One*, 6 from top medical journals, 17 from top psychology journals, and 5 from Metalab).

For each included meta-analysis, we fit Vevea & Hedges (1995)'s selection model under one-tailed selection, thus estimating a parameter equivalent to $\widehat{\eta}^{-1}$ and its standard error. This model assumes normally distributed population effects in the underlying population, prior to selection due to publication bias. As a primary analysis, we robustly meta-analyzed the log-transformed estimates, $\log\widehat{\eta}$ (Hedges et al., 2010), approximating their variances via the delta method. Combining all 58 meta-analyses, we thus estimated that affirmative results were on average $\widehat{\eta} = 1.17$ times more likely to be published than nonaffirmative results (95% CI: [0.93, 1.47]). Estimates within each of the four sources of meta-analyses were $\widehat{\eta} = 0.83$ (95% CI: [0.62, 1.11]) for meta-analyses in *PLoS One*, $\widehat{\eta} = 1.02$ (95% CI: [0.52, 1.98]) for those in top medical journals, $\widehat{\eta} = 1.54$ (95% CI: [1.02, 2.34]) for those in top psychology journals, and $\widehat{\eta} = 4.70$ (95% CI: [1.94, 11.34]) for those in Metalab (Mathur & VanderWeele (2020a), Table 1). Thus, except for Metalab, estimates of publication bias were fairly close to the null and with confidence intervals all bounded below $\eta = 3$. We conducted a number of sensitivity analyses to assess the impacts of possible violations of modeling assumptions, all of which yielded similar results (Mathur & VanderWeele, 2020a).

For the purpose of informing our proposed sensitivity analyses for publication bias, the upper tail of the distribution of true η values is particularly relevant as an indicator of the most severe publication bias that can considered plausible in a meta-analysis similar to those included in our sampling frame. To this end, we additionally estimated the 95^{th} quantile of the true

selection ratios using a nonparametric shrinkage method that accounts for sampling error (Wang & Lee, 2019). In contrast to simply considering the empirical 95^{th} quantile of the estimates $\widehat{\eta}$, this approach accounts for statistical error in estimating each $\widehat{\eta}$. The estimated 95^{th} quantiles of the distributions of true η values were 3.51 for all meta-analyses combined, 1.70 for *PLoS One*, 1.62 for top medical journals, 4.84 for top psychology journals, and 9.94 for Metalab. These results may serve as useful approximate benchmarks for the severity of publication bias.

These estimates of publication bias severity were lower than we had expected. We speculated that publication bias might operate primarily on individual studies published in higher-tier journals, or alternatively on the chronologically first few studies published on a topic. If so, publication bias might have been relatively mild in meta-analyses because, in principle, high-quality meta-analyses include all studies published in any journal and at any time, so even if elite journals and nascent fields induce severe publication bias by excluding nearly all nonaffirmative results, it is possible that these nonaffirmative results are still eventually published, perhaps in a lower-tier journals, and hence are still included in the meta-analysis. However, additional analyses did not support these hypotheses (Mathur & VanderWeele, 2020a). Instead, preliminary evidence suggested that the key alleviator of publication bias in the meta-analyses may have been their inclusion of "non-headline" results; that is, results that are reported in published papers but that are de-emphasized (e.g., those reported only in secondary or supplemental analyses) and those that meta-analysts obtain through manual calculation or by contacting authors (Mathur & VanderWeele, 2020a).

7. Applied examples

We now use the proposed methods to conduct sensitivity analyses for three existing meta-analyses for which the effects of publication bias have been controversial or difficult to assess using existing methods. Our re-analyses will suggest that, by shifting the focus from estimating publication bias severity to conducting sensitivity analyses and by relaxing asymptotic and distributional assumptions, our proposed methods can sometimes lead to clearer conclusions

than existing methods.

7.1. Video games and aggressive behavior

First, Anderson et al. (2010)'s meta-analysis assessed, among several outcomes, the association of playing violent video games with aggressive behavior. For brevity, we restrict attention here to analyses of 75 studies (27 experimental, 12 longitudinal, and 36 cross-sectional) satisfying Anderson et al. (2010)'s best-practice criteria for internal validity and that adjusted for sex, a suspected confounder. The 75 studies were contributed by 40 papers. (Throughout, we use "studies" to refer to point estimates and "papers" to refer to articles that potentially contribute multiple studies.) The experimental studies measured aggressive behavior in laboratory tasks that incentivized subjects to administer various aversive stimuli to other subjects, whereas the observational studies typically used standardized self- or peer-report measures. Anderson et al. (2010) estimated[2] a common-effect pooled point estimate of Pearson's $r = 0.15$ (95% CI: [0.14, 0.17]), such that playing violent video games was associated with a small increase in aggressive behavior. Debate ensued regarding whether the results could be explained away by publication bias; other authors suggested that publication bias might largely explain these results (Ferguson & Kilburn, 2010; Hilgard et al., 2017), while the original authors argued that the results were in fact robust to publication bias (Kepes et al., 2017)).

We performed our proposed sensitivity analyses for the common-effect specification as reported in Anderson et al. (2010) as well both random-effects specifications. We conducted analyses on Fisher's z scale but present results transformed to Pearson's r, except where otherwise stated. Throughout the applied examples, we use prime superscripts to denote estimates (e.g., $\widehat{\mu}'$) and effect size thresholds (q') that have been transformed back to the original scale, such as Pearson's r. For the robust clustered specification, we defined clusters as point estimates extracted from a single paper, resulting in 40 clusters (23 of these contained a single point estimate, while the remaining 17 contained between 2 and 9 point estimates). Table 1, rows

[2]Throughout the applied examples, our re-analyses sometimes yielded point estimates and confidence intervals that differed negligibly from those reported in the meta-analyses, often reflecting our use of restricted maximum likelihood estimation and Knapp-Hartung adjusted standard errors.

1-4 show uncorrected meta-analytic point estimates and confidence intervals along with the standard parametric random-effects specification for comparison. Rows 5-7 show the worst-case estimates from meta-analyses of only the 27 nonaffirmative studies.

Figure 2 shows a significance funnel plot, which suggests a positive correlation between the point estimates and their standard errors. Our methods apply under the assumption that such correlation arises from selection due to publication bias rather than to correlation between the point estimates and standard errors in the underlying population. To estimate the severity of publication bias required to attenuate $\widehat{\mu}'$ or $\widehat{\mu}^{lb\prime}$ to the null and to a non-null correlation of 0.10, we used the analytic results in Section 3.1 for the common-effect specification. For the two random-effects specifications, we conducted a grid search across values of η between 1 and 200 (Figure 3). Table 2, columns 2-3 indicate that for all three model specifications, no amount of publication bias under the assumed model could attenuate the observed point estimate or even its lower confidence interval limit to the null. Columns 4-5 indicate that when considering the non-null value $q' = 0.10$, extreme publication bias ($\eta = 11$ or $\eta = 28$ depending on whether common-effect or robust meta-analysis is used) or substantial publication bias ($\eta = 3$ to $\eta = 5$) would be capable of attenuating the estimate $\widehat{\mu}'$ or the lower bound of its confidence interval, $\widehat{\mu}^{lb\prime}$, respectively, to a correlation of 0.10. Thus, overall, we might conclude that regardless of the severity of publication bias, this meta-analysis provides strong evidence for an average effect in the observed direction, albeit possibly of small size.

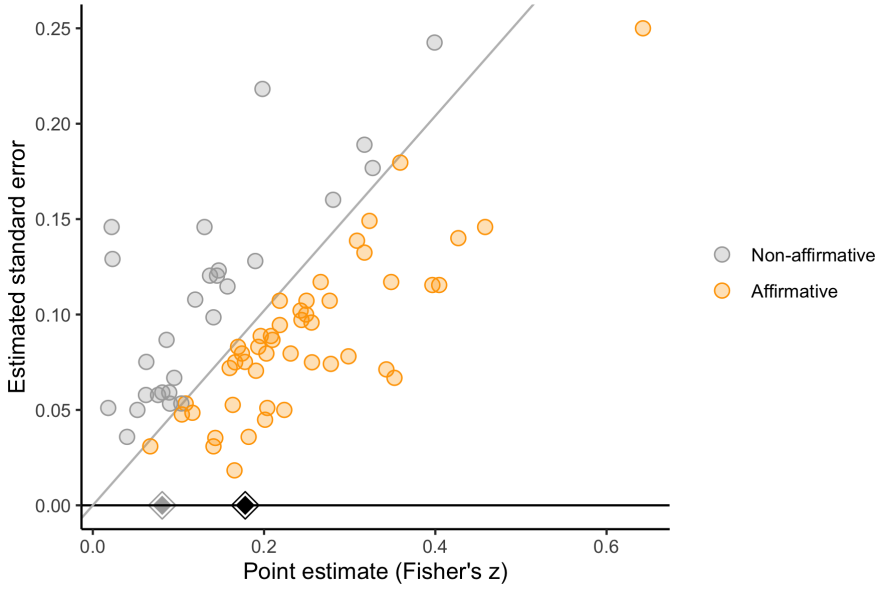


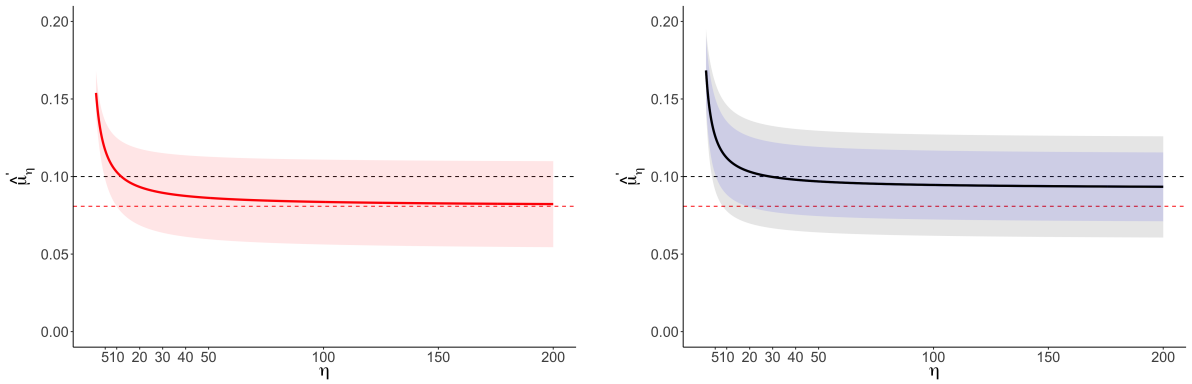
Figure 2: *Significance funnel plot for Anderson et al. (2010)'s video games meta-analysis. Point estimates are on Fisher's z scale, the scale on which p-values were computed. Studies lying on the diagonal line have p = 0.05. Gray diamond: robust clustered estimate in nonaffirmative studies only; black diamond: robust clustered estimate in all studies.*

Table 1: *Uncorrected and worst-case point estimates (Pearson's r) for Anderson et al. (2010)'s video games meta-analysis. Standard: Standard parametric random-effects meta-analysis assuming independence, included for comparison. $\widehat{\tau}$ and its CI are presented on the Fisher's z scale, the scale on which data were analyzed.*

Uncorrected	$\widehat{\mu}'$	CI for $\widehat{\mu}'$	$\widehat{\tau}$	CI for $\widehat{\tau}$
Common-effect	0.15	[0.14, 0.17]	–	–
Standard	0.17	[0.15, 0.19]	0.05	[0.02, 0.07]
Robust (independent)	0.17	[0.15, 0.19]	–	–
Robust (clustered)	0.18	[0.15, 0.20]	–	–
Worst-case				
Common-effect	0.08	[0.05, 0.11]	–	–
Robust (independent)	0.08	[0.06, 0.10]	–	–
Robust (clustered)	0.08	[0.05, 0.12]	–	–

Table 2: *Severity of publication bias (S) in Anderson et al. (2010) required to attenuate $\widehat{\mu}'$ or $\widehat{\mu}^{lb\prime}$ to null or to $q' = 0.10$ on the Pearson's r scale. "Not possible" indicates that no value of η could sufficiently attenuate the statistic. Values are conservatively rounded down to the nearest integer.*

Model	$S(\widehat{\mu}', 0)$	$S(\widehat{\mu}^{lb\prime}, 0)$	$S(\widehat{\mu}', 0.10)$	$S(\widehat{\mu}^{lb\prime}, 0.10)$
Common-effect	Not possible	Not possible	11	4
Robust (independent)	Not possible	Not possible	28	5
Robust (clustered)	Not possible	Not possible	28	3



(a) *Common-effect specification.*

(b) *Robust specifications. Blue band: CI assuming independence. Gray band: CI allowing for clustering.*

Figure 3: *Corrected point estimates and confidence intervals (CIs) for Anderson et al. (2010)'s video games meta-analysis as a function of η. Black dashed lines: non-null value q'. Red dashed lines: worst-case estimates.*

7.2. PI3K/AKT/mTOR inhibitors and progression-free cancer survival

A second meta-analysis of assessed the effect of PI3K/AKT/mTOR inhibitors on progression-free survival from advanced solid cancers (Li et al., 2018). The meta-analysis comprised 50 randomized controlled studies contributed by 39 papers and found that PI3K/AKT/mTOR inhibitors improved progression-free survival compared to various control therapies (hazard ratio $[HR] = 0.79$; 95% CI: [0.71, 0.88]). To conduct our sensitivity analyses, we assumed that publication bias would favor studies showing a protective effect of PI3K/AKT/mTOR inhibitors, so as described in Section 4, we first reversed the signs of all point estimates on the log-hazard scale. We transformed the results back to the hazard ratio scale and took inverses so that, in

all results we report, $HR < 1$ indicates a protective effect of PI3K/AKT/mTOR inhibitors as in the original meta-analysis. When conducting sensitivity analyses with a non-null effect-size threshold, we set $q = -\log(0.90) \approx \log(1.1)$ in analysis in order to consider attenuating the point estimate on the original scale ($\widehat{\mu}' = 0.79$) to a hazard ratio of $q' = 0.90$.

Figure 4 shows a significance funnel plot, Table 3 shows uncorrected and worst-case point estimates, Table 4 shows $S(\widehat{\mu}', q)$ and $S(\widehat{\mu}^{ub\prime}, q)$ for two choices of q, and Figure 5 shows $\widehat{\mu}_\eta$ as a function of η. The worst-case point estimates were close to the null and in the opposite direction from the original estimate for all three model specifications (e.g., 1.03 with 95% CI: [0.94, 1.12] for the robust clustered specification). Considerable publication bias would be required to shift the point estimate of 0.79 to the null ($S(\widehat{\mu}', 1) = 8$ for the robust clustered specification). However, only moderate publication bias would be required to shift the upper confidence interval limit to the null ($S(\widehat{\mu}^{ub\prime}, 1) = 2$) or to shift the point estimate to a non-null value of 0.90 ($S(\widehat{\mu}', 0.90) = 2$).

Li et al. (2018) had concluded that there was "no significant publication bias" ($p = 0.23$) based on Egger's test, which assumes that publication bias operates on point estimate size rather than "statistical significance" and does not affect the largest studies, and that effects are not heterogeneous. However, our sensitivity analyses suggest that the conclusions may be sensitive to plausible degrees of publication bias, such as publication bias in which affirmative studies are twice as likely to be published as nonaffirmative studies. The significance funnel plot 4 helps clarify the discrepancy: although the point estimates do not appear to be correlated with their standard errors (as Egger's test assesses), the estimates in the nonaffirmative studies are typically close to the null or even in the unexpected direction, so publication bias that favors affirmative results could potentially attenuate the meta-analytic estimate considerably.

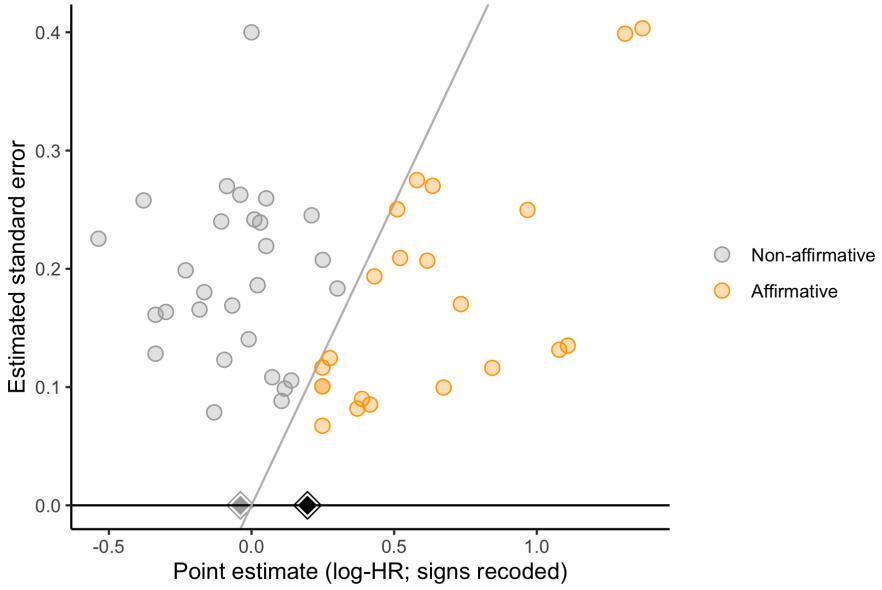


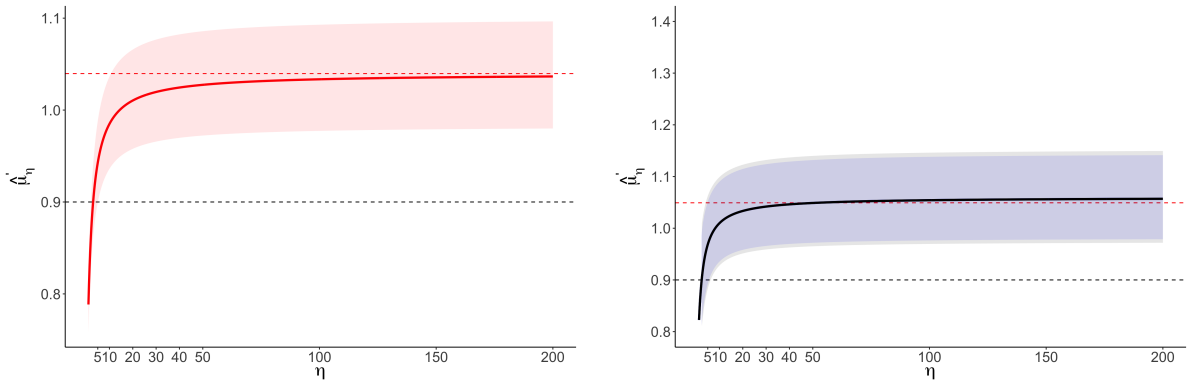
Figure 4: *Significance funnel plot for Li et al. (2018)'s cancer meta-analysis. Studies lying on the diagonal line have $p = 0.05$. Gray diamond: robust clustered estimate in nonaffirmative studies only; black diamond: robust clustered estimate in all studies.*

Table 3: *Uncorrected and worst-case point estimates (HR) for Li et al. (2018)'s cancer meta-analysis. Standard: Standard parametric random-effects meta-analysis assuming independence, included for comparison. $\widehat{\tau}$ and its CI are presented on the log-HR scale, the scale on which data were analyzed.*

Model	$\widehat{\mu}'$	CI for $\widehat{\mu}'$	$\widehat{\tau}$	CI for $\widehat{\tau}$
Common-effect	0.79	[0.76, 0.82]	–	–
Standard	0.79	[0.71, 0.89]	0.36	[0.26, 0.44]
Robust (independent)	0.80	[0.71, 0.89]	–	–
Robust (clustered)	0.82	[0.74, 0.91]	–	–
Worst-case				
Common-effect	1.04	[0.98, 1.10]		
Robust (independent)	1.05	[0.97, 1.14]		
Robust (clustered)	1.03	[0.94, 1.12]		

Table 4: *Severity of publication bias (S) in Li et al. (2018) required to attenuate $\widehat{\mu}'$ or $\widehat{\mu}^{ub\prime}$ (the upper limit of the CI for $\widehat{\mu}'$) to null or to $q' = 0.90$. "1" indicates that the statistic is already $\geq q'$. Values are conservatively rounded down to the nearest integer.*

Model	$S(\widehat{\mu}', 1)$	$S(\widehat{\mu}^{ub\prime}, 1)$	$S(\widehat{\mu}', 0.90)$	$S(\widehat{\mu}^{ub\prime}, 0.90)$
Common-effect	14	5	3	2
Robust (independent)	8	2	2	1
Robust (clustered)	8	2	2	1



(a) *Common-effect specification.*

(b) *Robust specifications. Blue band: CI assuming independence. Gray band: CI allowing for clustering.*

Figure 5: *Corrected point estimates (HR) and confidence intervals (CIs) for Li et al. (2018)'s cancer meta-analysis as a function of η. Black dashed lines: non-null value q'. Red dashed lines: worst-case estimates.*

7.3. Optimism and dietary quality

A third meta-analysis assessed the association between optimism and several health behaviors (Boehm et al., 2018). We focus here on the meta-analysis for dietary quality, which included 15 studies (8 cross-sectional and 7 longitudinal) contributed by 13 papers and found that optimism was associated with a small improvement in dietary quality ($r = 0.12$; 95% CI: [0.08, 0.16]). The authors reported that a standard Trim & Fill sensitivity analysis left the estimate and its confidence interval unchanged, yet applying a selection model (Andrews & Kasy, 2019) reversed its direction (Pearson's $r = -0.11$, 95% CI: [-0.30, 0.09]). This meta-analysis also serves as an interesting test case because its small size (and in particular, its inclusion of only 2 nonaffirmative studies) warrants considerable circumspection about the results of both Trim & Fill and standard selection models.

We conducted our sensitivity analyses as for Anderson et al. (2010)'s meta-analysis on video games. Figure 7 shows a significance funnel plot. Like Boehm et al. (2018)'s results using a standard selection model, our worst-case estimates (-0.12 for all specifications; Table 5) are in the opposite direction from the observed point estimate. However, because these worst-case

estimates are based on meta-analyzing only 2 nonaffirmative studies, the robust confidence intervals are, quite reasonably, almost completely uninformative (e.g., [-0.95, 0.92] for both random-effect specifications); this suggests that the narrower asymptotic confidence interval from the selection model (i.e., 95% CI: [-0.30, 0.09]) may be anticonservative for this small meta-analysis. Sensitivity analyses across multiple values of η (Table 6 and Figure 6) suggested that, under the two random-effects specifications, fairly considerable publication bias ($\eta = 3$ or $\eta = 4$) could attenuate $\widehat{\mu}'$ and its lower confidence interval limit to the null or to a correlation of 0.10.

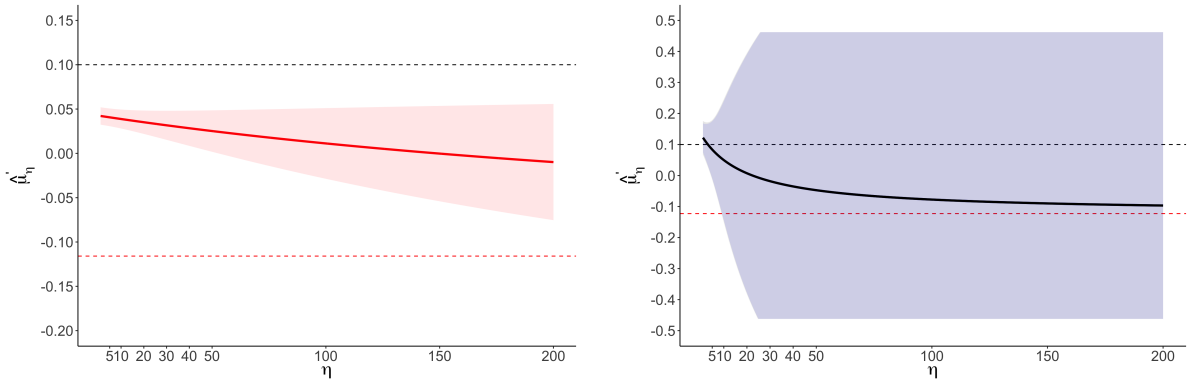
These analyses therefore indicate that while we can draw some conclusions regarding the sensitivity of these results to mild publication bias, considerable uncertainty remains regarding the impact of severe publication bias. However, unlike inference for standard selection models, our proposed methods of inference perform nominally even for small meta-analyses with few nonaffirmative studies (see Section 8 for simulation results). Thus, the wide confidence intervals for large values η may nevertheless be informative; they indicate that there simply is not enough information in this meta-analysis to provide any reasonable assurance that the results are robust to moderate or severe publication bias.

Table 5: *Uncorrected and worst-case point estimates (Pearson's r) for Boehm et al. (2018)'s optimism meta-analysis. Standard: Standard parametric random-effects meta-analysis assuming independence, included for comparison. $\widehat{\tau}$ and its CI are presented on the Fisher's z scale, the scale on which data were analyzed.*

Model	$\widehat{\mu}'$	CI for $\widehat{\mu}'$	$\widehat{\tau}$	CI for $\widehat{\tau}$
Common-effect	0.04	[0.03, 0.05]	–	–
Standard	0.12	[0.06, 0.18]	0.08	[0.00, 0.11]
Robust (independent)	0.12	[0.06, 0.17]	–	–
Robust (clustered)	0.13	[0.07, 0.19]	–	–
Worst-case				
Common-effect	-0.12	[-0.30, 0.08]		
Robust (independent)	-0.12	[-0.95, 0.92]		
Robust (clustered)	-0.12	[-0.95, 0.92]		

Table 6: *Severity of publication bias (S) in Boehm et al. (2018) required to attenuate $\widehat{\mu}'$ or $\widehat{\mu}^{lb\prime}$ to null or to $q' = 0.10$ on the Pearson's r scale. "1" indicates that the statistic is already $\leq q'$. Values are conservatively rounded down to the nearest integer.*

Model	$S(\widehat{\mu}', 0)$	$S(\widehat{\mu}^{lb\prime}, 0)$	$S(\widehat{\mu}', 0.10)$	$S(\widehat{\mu}^{lb\prime}, 0.10)$
Common-effect	148	49	1	1
Robust (independent)	22	4	3	1
Robust (clustered)	22	4	3	1



(a) *Common-effect specification.*

(b) *Robust specifications. Blue band: CI assuming independence. Gray band: CI allowing for clustering.*

Figure 6: *Corrected point estimates (Pearson's r) and confidence intervals (CIs) for optimism meta-analysis as a function of η. Black dashed lines: non-null value q'. Red dashed lines: worst-case estimates.*

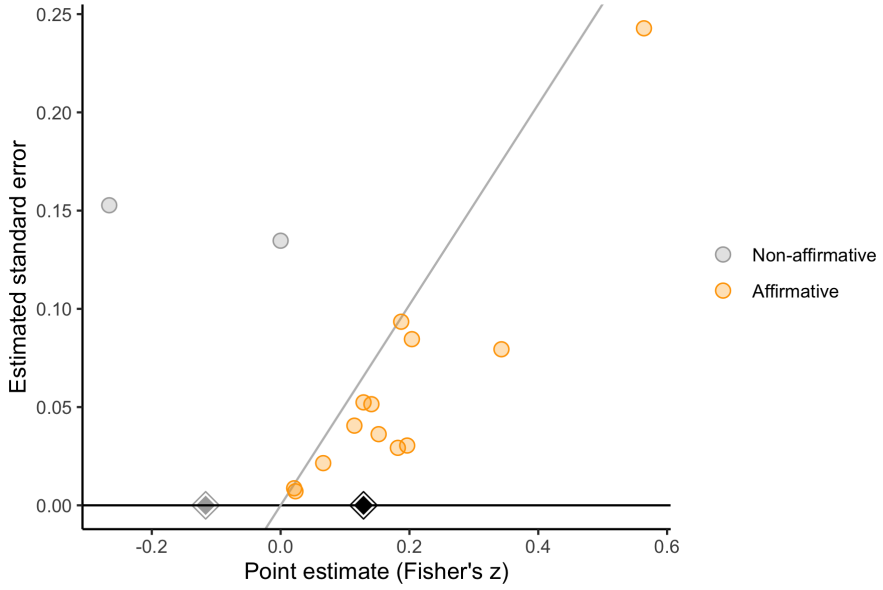


Figure 7: *Significance funnel plot for optimism meta-analysis. Studies lying on the diagonal line have* $p = 0.05$. *Gray diamond: robust clustered estimate in nonaffirmative studies only; black diamond: robust clustered estimate in all studies.*

8. Simulation study

We assessed the performance of our proposed sensitivity analyses under the three model specifications and in a variety of realistic and extreme scenarios, including those with quite small sample sizes. We considered three categories of scenarios: (1) those with homogeneous population effects, for which we applied the common-effect specification; (2) those with heterogeneous independent population effects, for which we used the robust independent specification; and (3) those with heterogeneous clustered population effects, for which we used the robust clustered specification. For each of 1,000 to 1,500 simulation iterates per scenario, we first generated an underlying population of studies before introducing publication bias. The underlying population comprised M^* clusters (potentially with no inter-cluster heterogeneity, depending on the scenario) of 5 studies each; note that $5 \times M^*$ represents the number of studies in the *underlying* population, but the number of published studies that were actually meta-analyzed was often much smaller after the introduction of publication bias, as described below. We generated studies' point estimates according to either a normal or an exponential random-intercepts specification such

that the total heterogeneity across studies was τ^2, comprising an inter-cluster heterogeneity of $\text{Var}(\zeta)$ and an intra-cluster heterogeneity of $\tau^2 - \text{Var}(\zeta)$. Continuing our convention of using asterisks to denote study-specific variables and parameters in the underlying population:

$$
\begin{aligned}
&\widehat{\theta}^*_{mi} = \mu + \zeta_m + \gamma^*_{mi} + \epsilon^*_{mi} \\
&\zeta_m \sim N(0, \text{Var}(\zeta)) && \text{(cluster-level random effects)} \\
&\gamma^*_{mi} \sim N(0, \tau^2 - \text{Var}(\zeta)) \text{ or } \gamma_{mi} \sim \text{Exp}(\left(\tau^2 - \text{Var}(\zeta)\right)^{-1/2}) && \text{(study-level random effects)} \\
&\epsilon^*_{mi} \sim N(0, (\sigma^*_{mi})^2) && \text{(study-level error)} \\
&\sigma^*_{mi} \sim \text{Unif}\,(1, 1.5)
\end{aligned}
$$

In a full-factorial design, we varied the parameters M^*, μ, τ^2, and $\text{Var}(\zeta)$, as well as the distribution of the study-level random effects (γ^*_{mi}), across the values in Table 7. We simulated a scenario for each combination of values in Table 7 except those with both $\tau^2 = 0$ and $\text{Var}(\zeta) = 0.5$, which would have had $\tau^2 - \text{Var}(\zeta) < 0$. Thus, scenarios with $\tau^2 = 0$ and $\text{Var}(\zeta) = 0$ were common-effect specifications, scenarios with $\tau^2 = 1$ and $\text{Var}(\zeta) = 0$ were independent random-effects specifications, and scenarios with $\tau^2 = 1$ and $\text{Var}(\zeta) = 0.5$ were clustered random-effects specifications.

We then introduced publication bias as in Section 2.2, varying η from 1 (no publication bias) to 100 (extreme publication bias). The effective sample size on which the precision of our methods depends most strongly is the number of published nonaffirmative studies, so we ensured that our choices of M^* and η resulted in numerous scenarios in which the median number of published nonaffirmative studies was less than 10 and sometimes as small as 1. Based on τ^2 and $\text{Var}(\zeta)$, we fit the correctly-specified fixed- or random-effects model using the true η to estimate $\widehat{\mu}_\eta$ and its 95% confidence interval. Additionally, as described in Section 2.2, our proposed sensitivity analyses can be applied without modification in certain situations in which the publication process not only favors affirmative studies, but also favors both affirmative and nonaffirmative studies with small standard errors. To confirm this result, we also ran all of the above simulation scenarios with this more complicated publication process.

Table 7: *Possible values of simulation parameters.*

η	M^*	μ	τ^2	Var (ζ)	γ distribution	Additional selection on σ_i^*
1	20	0.20	0	0	Normal	No
10	40	0.80	1	0.50	Exponential	Yes
20	80					
50	200					
100						

Figures 8 and 9, respectively, show the median numbers of published studies and of published nonaffirmative studies for all 480 simulation scenarios. Figure 10 shows that $\widehat{\mu}_\eta$ was approximately unbiased for scenarios with $\eta \leq 20$. Bias increased somewhat under extreme publication bias (e.g., $\eta = 100$), though coverage remained nominal for all scenarios (mean 96% and minimum 93% across all scenarios). Table 8 shows the median width of 95% confidence intervals for $\widehat{\mu}_\eta$, showing the expected patterns of dependence on the number of published nonaffirmative studies and on the severity of publication bias. Also as expected theoretically, results from scenarios in which the publication process also favored small standard errors yielded comparable results to scenarios in which the publication process favored only affirmative studies, except insofar as the former publication process resulted in smaller numbers of published studies.

Figure 8: *Median number of published studies across all simulation iterates. Rows: number of studies in underlying population prior to publication bias ($5 \times M^*$). Columns: distribution of study-level random effects and true mean, μ.*

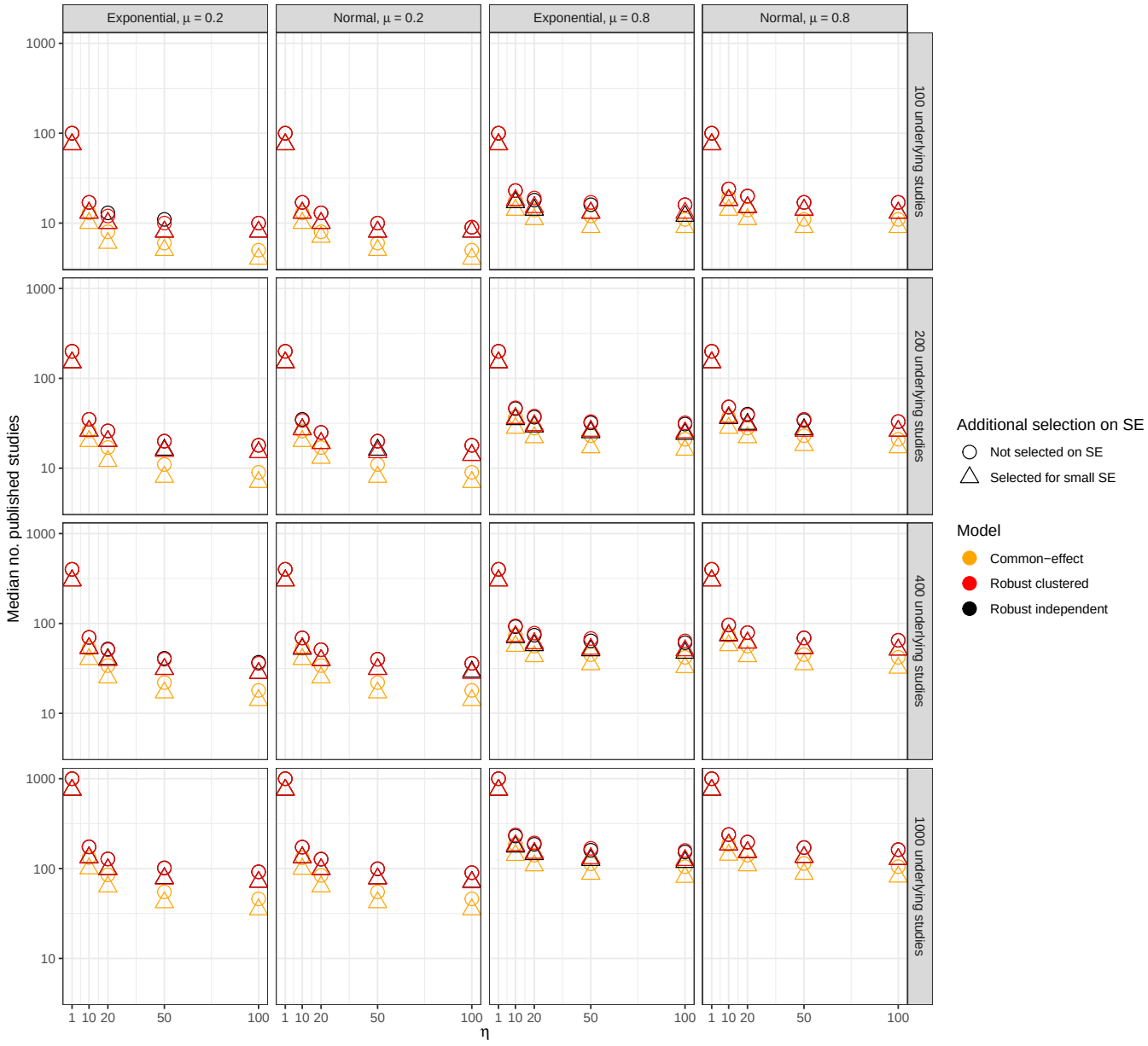


Figure 9: *Median number of published* ***nonaffirmative*** *studies across all simulation iterates. Rows: number of studies in underlying population prior to publication bias ($5 \times M^*$). Columns: distribution of study-level random effects and true mean, μ.*

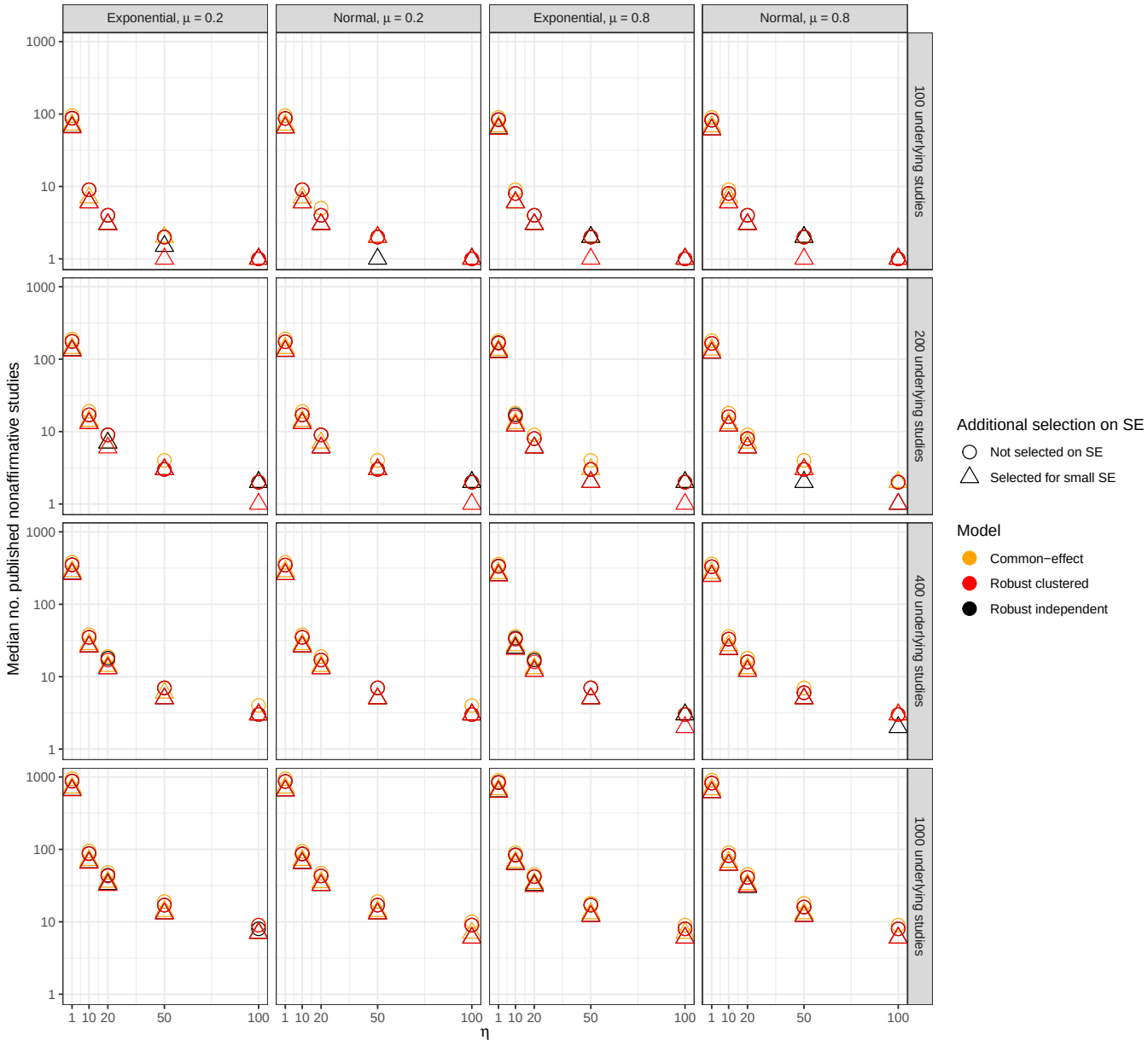


Figure 10: *Mean point estimate, $\widehat{\mu}_\eta$, across simulation iterates. Rows: number of studies in underlying population prior to publication bias ($5 \times M^*$). Columns: distribution of study-level random effects and true mean, μ. Red dashed line: μ.*

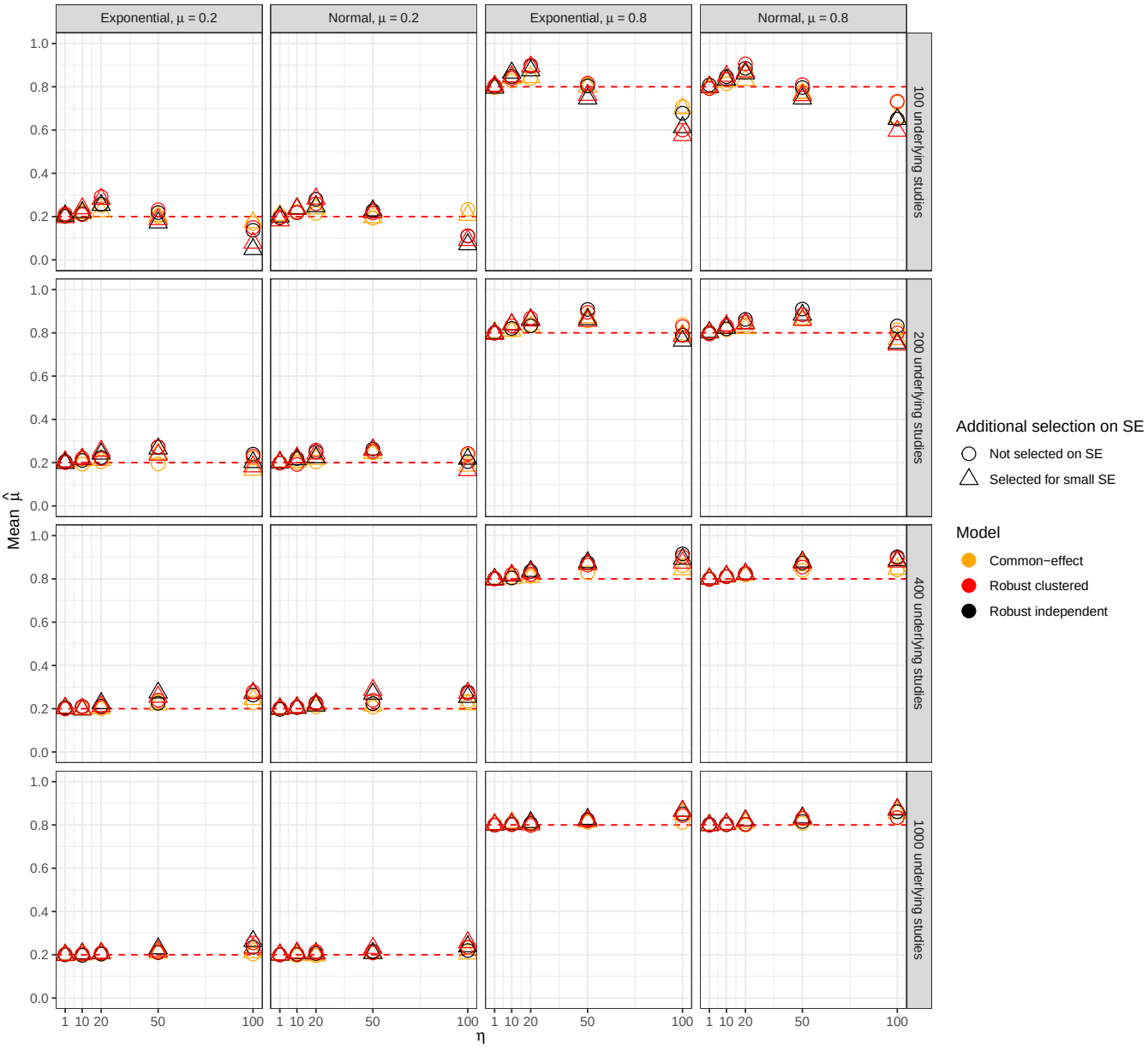


Table 8: *Median width of confidence interval (CI) for $\widehat{\mu}_\eta$ by the median number of published nonaffirmative studies in the simulation scenario ($|\mathcal{N}|$) and η. (Among scenarios with $\eta = 1$, all had a median of $\geq$ 10 published nonaffirmative studies.)*

Median $\lvert\mathcal{N}\rvert$	η	Median CI width
≥ 10	1	0.40
	10	0.86
	20	0.96
	50	1.19
	100	1.49
< 10	10	1.97
	20	2.30
	50	3.33
	100	4.23

9. Alternative models of publication bias

Throughout, we have considered a one-tailed model of publication bias in which "significant" results with positive point estimates are favored, while "significant" results with negative point estimates and "nonsignificant" results are equally disfavored. We have also assumed that results are selected for publication based on a single p-value cutoff at $p = 0.05$. This section discusses our rationale for these modeling choices and describes how our results could be easily extended to accommodate other models of publication bias.

Whereas we assumed one-tailed selection, it is possible instead that "significant" results are favored regardless of direction, while only "nonsignificant" results are disfavored (called "two-tailed" selection). Although our methods can be trivially modified for two-tailed selection, as described below, we speculate that one-tailed selection is more realistic in many scientific contexts. For many research questions, only effects in the positive direction are scientifically marketable, while both negative and null results are interpreted as a failure to support the marketable hypothesis (e.g., Vevea et al. (1993)). Even in controversial realms, in which some investigators try to prove a given hypothesis while others try to disprove it, "nonsignificant" and negative results may be comparably interpreted as failing to support the hypothesis at stake. For example, we suspect that this may be the case for the literature on violent video

games, which has tended to interpret results suggesting *beneficial* effects of violent video games primarily as failures to support hypothesized detrimental effects, rather than as support for *a priori* less plausible benefits. Indeed, the distributions of p-values in the meta-analyses included in the empirical study of Section 6 usually appeared to conform well to one-tailed selection (Mathur & VanderWeele, 2020a).

Some areas of research may, however, exhibit two-tailed selection. For example, if there are two scientifically marketable hypotheses at stake, each predicting results in a different direction, then perhaps publication bias would equally favor "significant" positive results and "significant" negative results. In such cases, one could conduct our proposed sensitivity analyses under two-tailed selection simply by redefining $A_i = \mathbb{1}\{p_i < 0.05\}$, such that all "significant" studies, regardless of direction, receive weights of 1, while only "nonsignificant" studies receive weights of η. However, we nevertheless recommend by default conducting sensitivity analyses under a one-tailed selection model, even when there is reason to suspect some degree of two-tailed selection, because assuming one-tailed selection is often (though not always) conservative in the sense that it leads to smaller $\eta(\widehat{\mu}, q)$ than assuming two-tailed selection. Specifically, conservatism holds when the inverse-probability-weighted, common-effect mean among only the "nonsignificant" and affirmative studies is at least as large as the common-effect mean among only the "significant" negative studies, for which a sufficient condition is that the common-effect mean in the "nonsignificant" studies is positive (Supplement, Section 1.2). (However, note that this conservatism does not necessarily hold for the analogous metric pertaining to lower bound of the confidence interval for the pooled estimate, $\eta(\widehat{\mu}^{lb}, q)$.)

Note also that we modeled selection using a single cutoff at $p = 0.05$ both for ease of interpretation and because it conforms well to empirical findings on how applied researchers and statisticians interpret and report p-values (Head et al., 2015; Masicampo & Lalande, 2012; McShane & Gal, 2017). In principle, other cutoffs might also be relevant; for example, "marginally significant" findings with $0.05 < p < 0.10$ might have an intermediate publication probability. However, in practice, experimental evidence suggests that researchers do not distinguish much between two different p-values both falling above or below the major 0.05 cutoff (McShane et al.,

2016). Our proposed sensitivity analyses could in fact be modified for multiple cutoffs simply by defining more than two groups of studies, each with a distinct publication probability, and again weighting each study by its inverse publication probability (as in, for example, Hedges (1992)). However, by introducing multiple free sensitivity parameters, this approach would less readily yield straightforward single-number summaries of the severity of publication bias required to "explain away" the results.

10. Discussion

This paper has proposed sensitivity analyses for bias due to selective publication and reporting in meta-analyses. These sensitivity analyses shift the focus from estimating the severity of publication bias to quantifying the amount of publication bias that would be required to attenuate the observed point estimate, or its lower confidence interval limit, to the null or to a chosen non-null value. This shift in focus enables particularly simple statements regarding sensitivity to publication bias that would be easy to report in meta-analyses. Our metric $S(t, q)$ describes the amount of publication bias required to attenuate the meta-analytic statistic t (i.e., either the pooled point estimate or its lower confidence interval limit) to a smaller value q. If this sensitivity parameter is small enough that it represents a plausible amount of publication bias (perhaps informed by the empirical benchmarks we have provided), then the meta-analysis may be considered relatively sensitive to publication bias. In contrast, if $S(t, q)$ represents an implausibly large amount of publication bias, then one might consider the meta-analysis to be relatively robust. The proposed methods can sometimes indicate that no amount of publication bias under the assumed model could "explain away" the results of a meta-analysis, providing a compelling argument for robustness. In contrast to existing methods for sensitivity analysis, the present methods can accommodate non-normal population effects, small meta-analyses, and non-independent point estimates. All methods are implemented in the R package `PublicationBias` (described in the Supplement, Section 3).

These methods have limitations. Although they relax distributional assumptions on the

population effects, they do assume a particular model of publication bias chosen to align with empirical evidence on how researchers interpret p-values. We have suggested some simple diagnostics to assess the plausibility of some of these assumptions (Section 4). If publication bias departs considerably from this model, for example because studies with large effects rather than merely $p < 0.05$ are favored, the proposed analyses may be compromised. However, for perhaps the most plausible violation of our modeling assumptions (namely, two-tailed instead of one-tailed selection), we have shown that our assumptions are often, though not always, conservative at least when considering the point estimate. Additionally, some meta-analyses may not contain any nonaffirmative studies because, for example, the population effects are very large or publication bias is extremely severe; in these cases, our methods cannot be applied because there would be no nonaffirmative studies to upweight. Also, it is important to note that sensitivity to publication bias does not imply that publication bias is in fact severe in practice, nor the inverse. Last, our sensitivity analyses characterize evidence strength using the standard meta-analytic point estimate and its confidence interval, but these metrics alone do not fully characterize evidence strength in a potentially heterogeneous distribution of effects (Mathur & VanderWeele, 2018). Other metrics may be useful additional targets of sensitivity analysis (Mathur & VanderWeele (2018); Mathur & VanderWeele (2020b)) but would require bias correction for $\widehat{\tau}^2$ as well as $\widehat{\mu}$, which proved challenging under publication bias in meta-analyses of realistic sizes (Supplement, Section 3).

In summary, we have proposed sensitivity analyses for publication bias in meta-analyses that are straightforward to conduct and intuitive to interpret. These methods can be easily implemented using the `R` package `PublicationBias`, and we believe that their widespread reporting would help calibrate confidence in meta-analysis results.

Reproducibility

All code required to reproduce the applied examples and simulation study is publicly available and documented (`https://osf.io/prwyc/`). Data from the Boehm et al. (2018) and Li et al.

(2018) meta-analyses are publicly available (linked at `https://osf.io/prwyc/`). Data from the Anderson et al. (2010) meta-analysis cannot be made public at the author's request, but will be made available upon request to individuals who have secured permission from Craig Anderson.

Acknowledgments

Craig Anderson provided raw data for re-analysis and answered analytic questions. Zachary McCaw contributed to the proof of Theorem 1.1 (Supplement, Section 1.1). Zachary Fisher provided helpful discussion regarding his R package `robumeta`. The participants of the Harvard University Applied Statistics Workshop provided many insightful comments and suggestions. This research was supported by NIH grant R01 CA222147 and John E. Fetzer Memorial Trust grant R2020-16; the funders had no role in the conduct or reporting of this research.

References

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., ... Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: A meta-analytic review. *Psychological Bulletin*, *136*(2), 151.

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*(8), 2766–94.

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*(6), 1996–2009.

Boehm, J. K., Chen, Y., Koga, H., Mathur, M. B., Vie, L. L., & Kubzansky, L. D. (2018). Is optimism associated with healthier cardiovascular-related behavior? Meta-analyses of 3 health behaviors. *Circulation Research*, *122*(8), 1119–1134.

Bom, P. R., & Rachinger, H. (2019). A kinked meta-regression model for publication bias correction. *Research Synthesis Methods*, *10*(4), 497–514.

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

Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *Journal of the American Medical Association*, *291*(20), 2457–2465.

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice*, *17*(2).

Dear, K. B., & Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*, 237–245.

Ding, P., & VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, *27*(3), 368.

Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634.

Ferguson, C. J., & Kilburn, J. (2010). Much ado about nothing: The misestimation and overinterpretation of violent video game effects in eastern and western nations: Comment on anderson et al.(2010).

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 665–694.

Fisher, Z., & Tipton, E. (2015). Robumeta: An R-package for robust variance estimation in meta-analysis. *arXiv preprint arXiv:1503.02220*.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3).

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, 246–255.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65.

Hilgard, J., Engelhardt, C. R., & Rouder, J. N. (2017). Overstated evidence for short-term effects of violent games on affect and behavior: A reanalysis of Anderson et al.(2010). *Psychological Bulletin*, *143*(7).

Ioannidis, J. P., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *Canadian Medical Association Journal*, *176*(8), 1091–1096.

Jin, Z.-C., Zhou, X.-H., & He, J. (2015). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine*, *34*(2), 343–360.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association*, *112*(517), 1–10.

Kepes, S., Bushman, B. J., & Anderson, C. A. (2017). Violent video game effects remain a societal concern: Reply to Hilgard, Engelhardt, and Rouder (2017). *Perspectives on Psychological Science*, *143*(7).

Lee, K. P., Boyd, E. A., Holroyd-Leduc, J. M., Bacchetti, P., & Bero, L. A. (2006). Predictors of publication: characteristics of submitted manuscripts associated with acceptance at major biomedical journals. *Medical Journal of Australia*, *184*(12), 621.

Lewis, M., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P., Cristia, A., & Frank, M. C. (2016). A quantitative synthesis of early language acquisition using meta-analysis.

Li, X., Dai, D., Chen, B., Tang, H., Xie, X., & Wei, W. (2018). Efficacy of PI3K/AKT/mTOR pathway inhibitors for the treatment of advanced solid cancers: A literature-based meta-analysis of 46 randomised control trials. *PloS One*, *13*(2).

Masicampo, E., & Lalande, D. R. (2012). A peculiar prevalence of p values just below. 05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279.

Mathur, M. B., & VanderWeele, T. J. (2018). New metrics for meta-analyses of heterogeneous effects. *Statistics in Medicine*. doi: 10.1002/sim.8057

Mathur, M. B., & VanderWeele, T. J. (2020a). Estimating publication bias in meta-analyses: A meta-meta-analysis across disciplines and journal tiers.
(Preprint retrieved from `https://osf.io/p3xyd/`)

Mathur, M. B., & VanderWeele, T. J. (2020b). Robust metrics and sensitivity analyses for meta-analyses of heterogeneous effects. *Epidemiology*, *31*(3), 356–358.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*(5), 730–749.

McShane, B. B., & Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, *112*(519), 885–895.

Olson, C. M., Rennie, D., Cook, D., Dickersin, K., Flanagin, A., Hogan, J. W., ... Pace, B. (2002). Publication bias in editorial decision making. *Journal of the American Medical Association*, *287*(21), 2825–2828.

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

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, *61*(10), 991–996.

Rice, K., Higgins, J., & Lumley, T. (2018). A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(1), 205–227.

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

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage Publications.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78.

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*(11), 1119–1129.

Tipton, E. (2015). Small-sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375.

VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, doi: 10.7326/M16-2607.

VanderWeele, T. J., Mathur, M. B., & Ding, P. (2019). Correcting misinterpretations of the e-value. *Annals of Internal Medicine*, *170*(2), 131–132.

Vevea, J. L., Clements, N. C., & Hedges, L. V. (1993). Assessing the effects of selection bias on validity data for the general aptitude test battery. *Journal of Applied Psychology*, *78*(6), 981.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428.

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.

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Wang, C.-C., & Lee, W.-C. (2019). A simple method to estimate prediction intervals and predictive distributions: Summarizing meta-analyses beyond means and confidence intervals. *Research Synthesis Methods*.

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

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:

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

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

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

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

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

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

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:

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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 $E\left[D_i^* w_i^* \mid F_i^*, A_i^*\right]$ as $E\left[D_i^* \mid A_i^*\right] E\left[w_i^* \mid F_i^*, A_i^*\right]$. □

Corollary 1.1 (Limiting result)**.**

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

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:*

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

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:

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

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

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

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

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

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 $\mathcal{N}^-$:

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

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:*

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

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 $\eta \geq 1$:

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

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:*

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

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 $\widehat{\mu}_\eta^t$, we have:

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

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

We therefore have four cases to consider:

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

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 $\widehat{\mu}_\eta^t$ are monotonically decreasing in η and $\widehat{\mu}_\eta \leq \widehat{\mu}_\eta^t$ 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 $\frac{\partial \widehat{\mu}_\eta^t}{\partial \eta} \geq 0$

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 $\frac{\partial \widehat{\mu}_\eta^t}{\partial \eta} \geq 0$

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 $\frac{\partial \widehat{\mu}_\eta^t}{\partial \eta} < 0$

Since $\widehat{\mu}_\eta \leq \widehat{\mu}_\eta^t$ for $\eta > 1$ by Lemma 1.3 and $\widehat{\mu}_\eta = \widehat{\mu}_\eta^t$ 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:

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

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

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

□

For example, if applying the proposed sensitivity analyses yields $S\left(t, q\right) = 10$, 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 $S\left(t, q\right)$, 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, $S\left(t, q\right)$, 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:

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

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.*

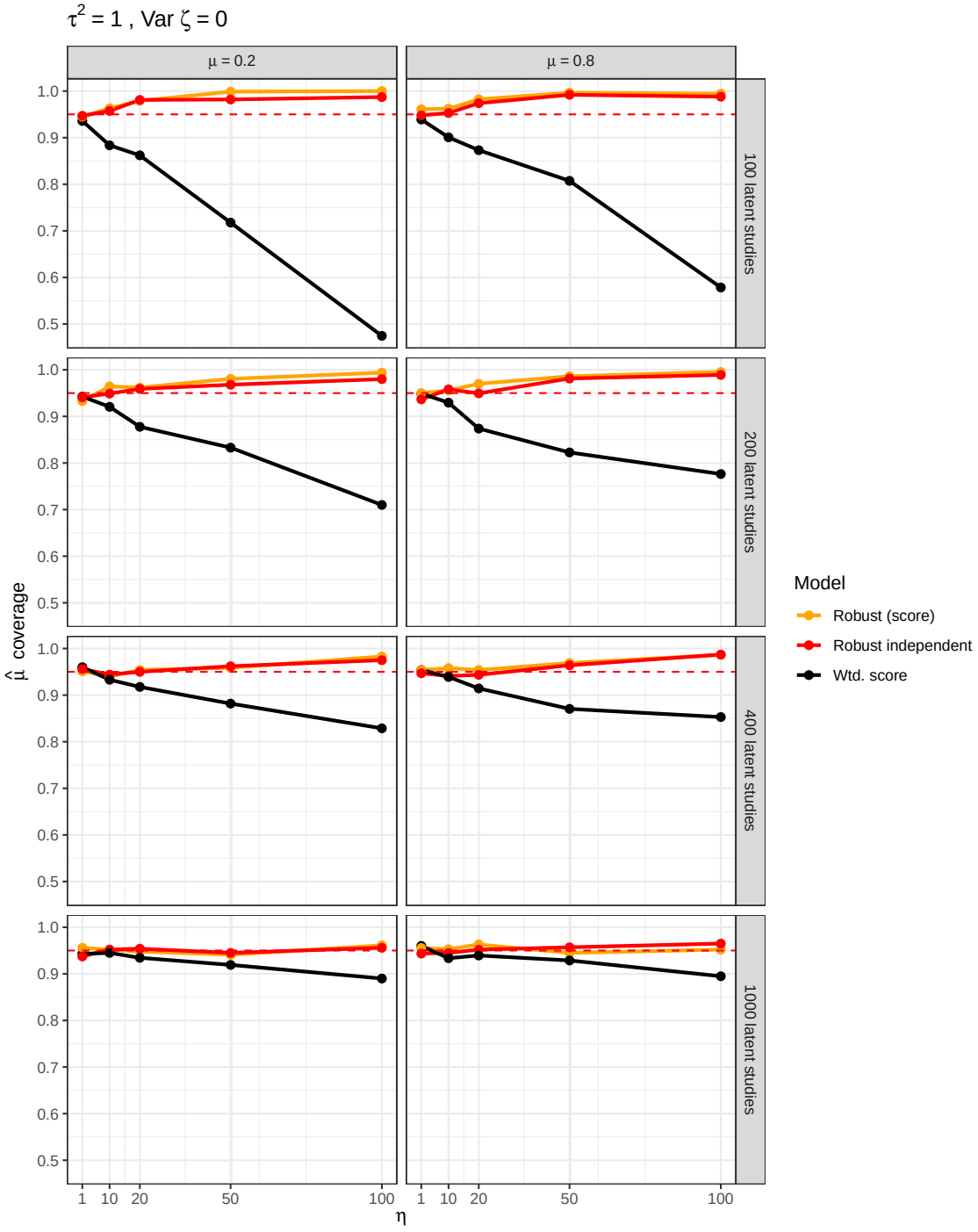


Figure S2: *Median width of confidence interval for $\widehat{\mu}_\eta$ in scenarios without clustering. "Robust (score)": Robust independent model in which $\widehat{\tau}^2$ is chosen by first fitting the weighted score model. "Robust independent": Robust independent model as in the main text, in 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.*

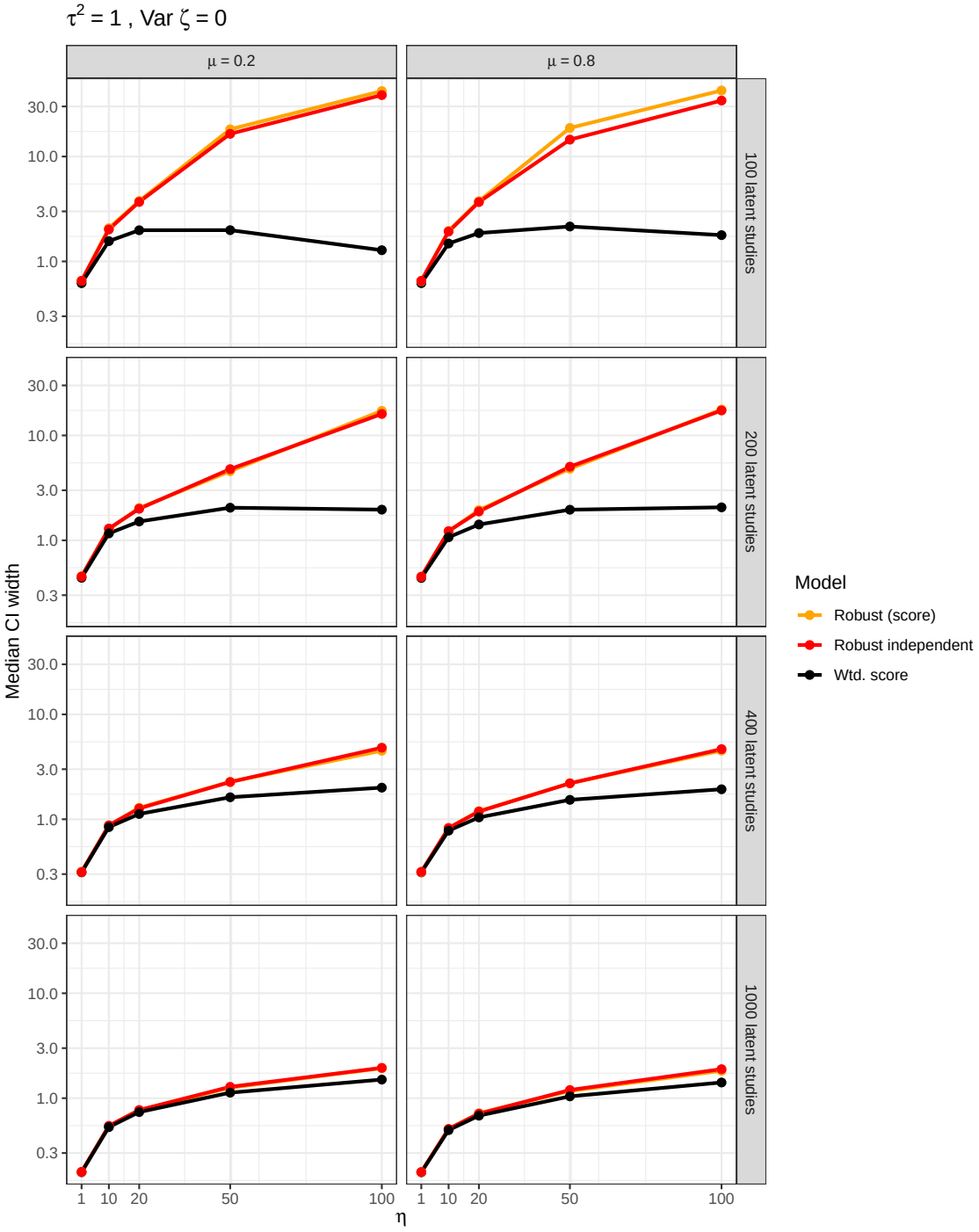


Figure S3: *Mean coverage in scenarios with clustering. "Robust (score)": Robust clustered model in which $\widehat{\tau}^2$ is chosen by first fitting the weighted score model. "Robust clustered": Robust clustered model as in the main text, in which $\widehat{\tau}^2$ is chosen by first fitting the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

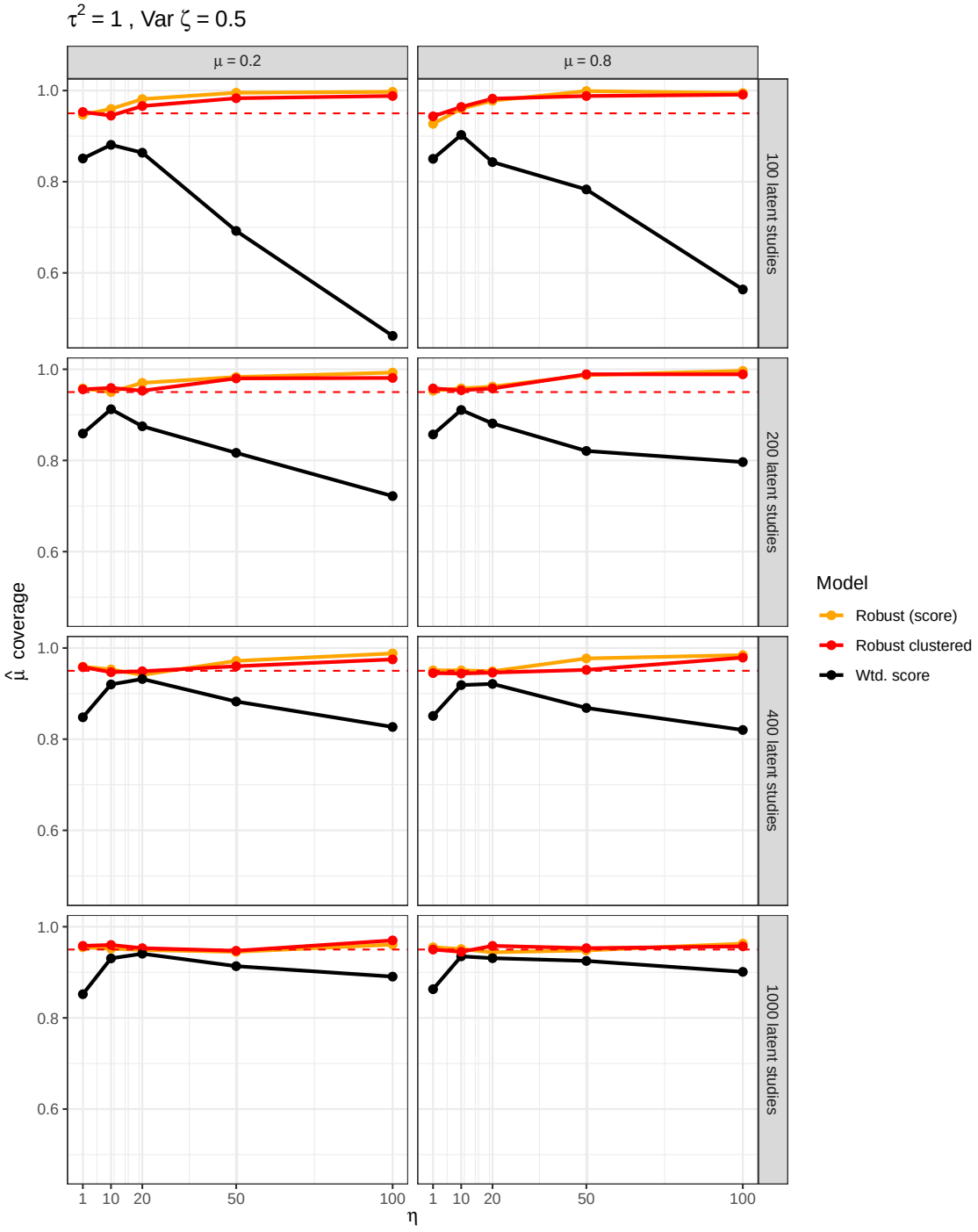
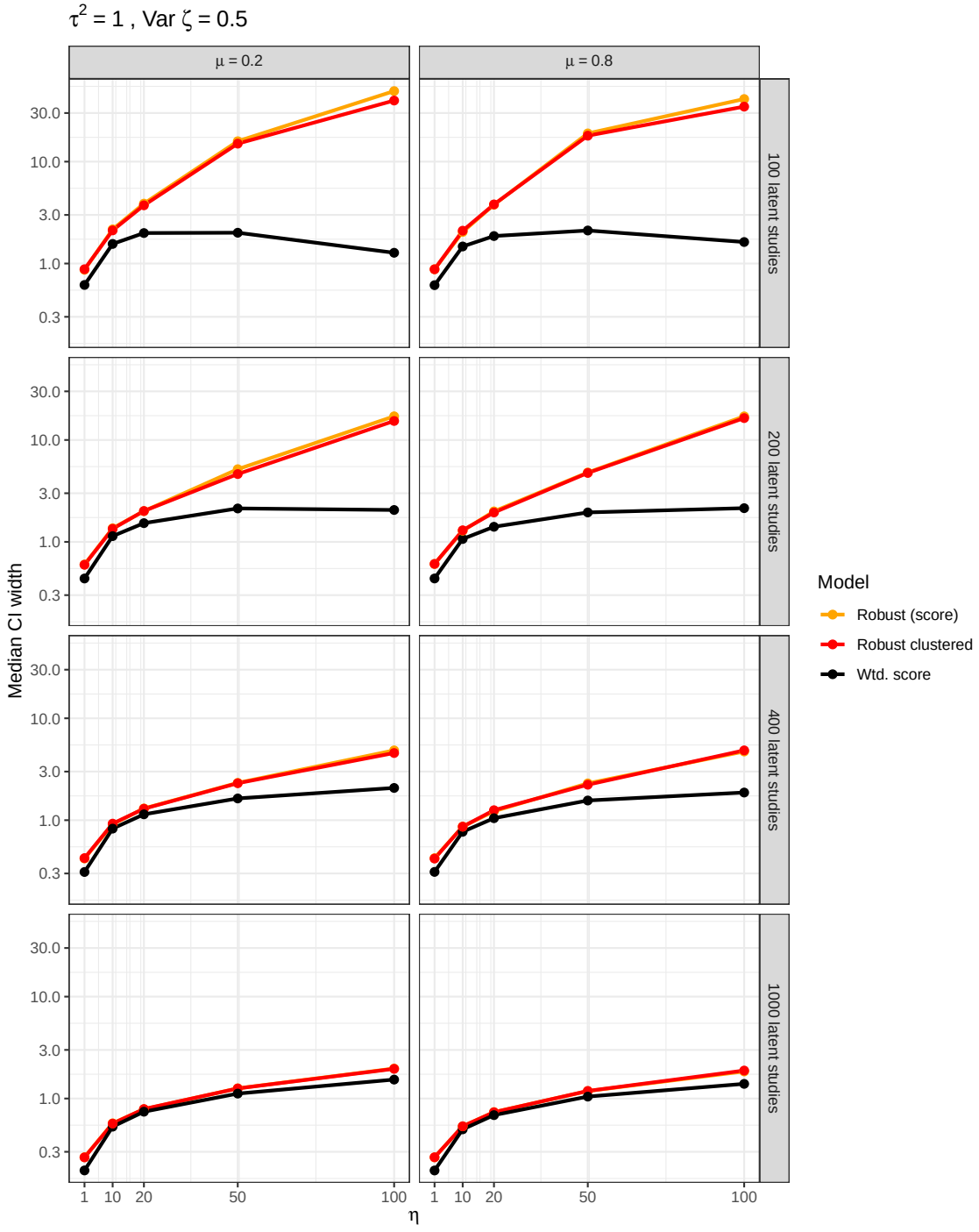


Figure S4: *Median width of confidence interval for $\widehat{\mu}_\eta$ in scenarios with clustering. "Robust (score)": Robust clustered model in which $\widehat{\tau}^2$ is chosen by first fitting the weighted score model. "Robust clustered": Robust clustered model as in the main text, in which $\widehat{\tau}^2$ 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.