Robust Bayesian Meta-Analysis: Model-Averaging Across Complementary Publication Bias Adjustment Methods

František Bartoš[1,2], Maximilian Maier[1,3], Eric-Jan Wagenmakers[1], Hristos Doucouliagos[4,5], T.D. Stanley[4,5]

1 Department of Psychology, University of Amsterdam

2 Institute of Computer Science, Czech Academy of Sciences, Prague, Czech Republic

3 Department of Experimental Psychology, University College London

4 Deakin Laboratory for the Meta-Analysis of Research (DeLMAR), Deakin University

5 Department of Economics, Deakin University

Author Note

This is the authors' version of the manuscript.

See https://doi.org/10.1002/jrsm.1594 for the version of record published in *Research Synthesis Methods*.

This project was supported in part by a Vici grant (#016.Vici.170.083) to EJW. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Abstract

Publication bias is a ubiquitous threat to the validity of meta-analysis and the accumulation of scientific evidence. In order to estimate and counteract the impact of publication bias, multiple methods have been developed; however, recent simulation studies have shown the methods' performance to depend on the true data generating process, and no method consistently outperforms the others across a wide range of conditions. Unfortunately, when different methods lead to contradicting conclusions, researchers can choose those methods that lead to a desired outcome. To avoid the condition-dependent, all-or-none choice between competing methods and conflicting results, we extend robust Bayesian meta-analysis and model-average across two prominent approaches of adjusting for publication bias: (1) selection models of p-values and (2) models adjusting for small-study effects. The resulting model ensemble weights the estimates and the evidence for the absence/presence of the effect from the competing approaches with the support they receive from the data. Applications, simulations, and comparisons to preregistered, multi-lab replications demonstrate the benefits of Bayesian model-averaging of complementary publication bias adjustment methods.

Keywords: Meta-Analysis, Publication Bias, Bayesian Model-Averaging, Selection Models, PET-PEESE

Robust Bayesian Meta-Analysis: Model-Averaging Across Complementary Publication Bias Adjustment Methods

Meta-analysis is essential to cumulative science (e.g., Borenstein et al., 2009). However, a common concern to meta-analysis is the overestimation of effect size due to publication bias, the preferential publishing of statistically significant studies (Masicampo & Lalande, 2012; Rosenthal & Gaito, 1964; Rothstein et al., 2005a; Scheel et al., 2021; Wicherts, 2017). In addition, this effect size exaggeration can be further increased by questionable research practices, that is, researchers' tendency to manipulate their data in a way that increases the effect size and the evidence for an effect (e.g., Simmons et al., 2011; Stefan & Schönbrodt, 2022). Indeed, descriptive surveys find that both problems are remarkably common. For example, John et al. (2012) estimate that about 78% of researchers failed to disclose all dependent measures and around 36% stopped data collection after achieving a significant result[1] (but see Fiedler and Schwarz, 2016 who argued that the survey by John et al. overestimated QRPs prevalence).

The results of the publication bias and questionable research practices are often viewed as a missing data problem; where some studies are missing from the published research record because they did not reach a statistical significance criterion while other estimates are observed after being "massaged" by researchers. Unfortunately, a perfect solution to the problem of missing data is impossible since we cannot know the unreported results nor the precise mechanism of omission. Multiple methods have been offered to adjust for likely publication bias from observable patterns contained in the reported research record (e.g., Andrews & Kasy, 2019; Bom & Rachinger, 2019; Citkowicz & Vevea,

[1] As measured by the geometric mean over researchers' self-admission rate, prevalence estimate (the estimate for the percent of other researchers who had engaged in a behavior), and an admission estimate. For the admission estimates the number of people reporting to have engaged in a given QRP was divided by researchers estimate for the proportion of other researchers that would admit that they engaged in this QRP.

2017; Copas, 1999; Duval & Tweedie, 2000; Egger et al., 1997; Iyengar & Greenhouse, 1988; Maier, Bartoš, & Wagenmakers, 2022; Simonsohn et al., 2014; Stanley & Doucouliagos, 2014, 2017; Stanley et al., 2017; Van Assen et al., 2015; Vevea & Hedges, 1995). All of these methods have been shown to thrive under different assumptions and simulation designs (e.g., Carter et al., 2019; Hong & Reed, 2020; Maier, Bartoš, & Wagenmakers, 2022; Renkewitz & Keiner, 2019).

Because different methods can lead to different conclusions, some meta-analysts suggest that we should not search for the "best" bias-adjusted effect size estimate. Instead, they suggest that multitude bias-adjusted effect size estimates should be offered as a sensitivity analysis for the original unadjusted value (e.g., Mathur & VanderWeele, 2020; Rothstein et al., 2005b; Vevea & Woods, 2005). The new method proposed here is another useful tool to accommodate publication bias, and researchers are free to supplement it with other methods.

Researchers interested in obtaining better bias-adjusted effect size estimates or selecting the most suitable set of methods for sensitivity analysis increasingly emphasize the importance of selecting an appropriate estimator conditional on the situation at hand. For instance, Hong and Reed (2020) argue:

> What is missing is something akin to a flow-chart that would map observable characteristics to experimental results which the meta-analyst could then use to select the best estimator for their situation. (p. 22)

and Carter et al. (2019) write:

> Therefore, we recommend that meta-analysts in psychology focus on sensitivity analyses—that is, report on a variety of methods, consider the conditions under which these methods fail (as indicated by simulation studies such as ours), and then report how conclusions might change depending on which conditions are most plausible. (p. 115)

In practice, researchers seldom have knowledge about the data-generating process nor do they have sufficient information to choose with confidence among the wide variety of proposed methods that aim to adjust for publication bias. Furthermore, this wide range of proposed methods often leads to contradictory conclusions (Carter et al., 2019). The combination of uncertainty about the data-generating process and the presence of conflicting conclusions can create a "breeding ground" for confirmation bias (e.g. Oswald & Grosjean, 2004): researchers may unintentionally select those methods that support the desired outcome. This freedom to choose can greatly inflate the rate of false positives, which can be a serious problem for conventional meta-analysis methods.

An alternative approach is to integrate the different approaches, explicitly, and let the data determine the contribution of each model based on its relative predictive accuracy for the observed data. To implement this approach we extend the robust Bayesian meta-analysis (RoBMA) framework outlined in Maier, Bartoš, and Wagenmakers (2022). The original RoBMA framework included selection models (operating on p-values) that have been shown to work well even under high heterogeneity (Carter et al., 2019; McShane et al., 2016; see also Guan and Vandekerckhove, 2016). The extended RoBMA framework also includes PET-PEESE, a method that adjusts for small-study effects by modeling the relationship between the effect sizes and standard errors (Stanley & Doucouliagos, 2014). PET-PEESE generally has low bias and performs well in applications (Carter et al., 2019; Kvarven et al., 2020). By including both p-value selection models as well as PET-PEESE, the extended version of RoBMA can apply both models simultaneously and optimally, relative to the observed research record.

Below we first provide a brief introduction to the RoBMA framework. We use an example on precognition (Bem, 2011) to illustrate both the general model-averaging methodology, RoBMA-PSMA (PSMA: publication selection model-averaging), and the way RoBMA-PSMA combines multiple weight functions including PET-PEESE. Second, we evaluate RoBMA-PSMA on comparisons with findings from preregistered multi-lab

replications (Kvarven et al., 2020), and across more than a thousand simulation environments employed by four different simulation studies (Hong & Reed, 2020).

Robust Bayesian Meta-Analysis: General Background

Because the true data generating process is unknown (effect present vs. effect absent; fixed-effect vs. random-effects; no publication bias vs. publication bias; and how publication bias expresses itself), many different models can be specified. RoBMA-PSMA accepts this multitude of models and uses Bayesian model-averaging to combine the estimates from individual models based on how well each model predicts the data (Hinne et al., 2020; Hoeting et al., 1999; Leamer, 1978). Consequently, the posterior plausibility for each individual model determines its contribution to the model-averaged posterior distributions (e.g., Gronau et al., 2021; Gronau et al., 2017; Maier, Bartoš, & Wagenmakers, 2022).

In this section, we provide a brief overview of Bayesian model-averaging, the work horse of RoBMA (for an in-depth treatment see Fragoso et al., 2018; Gronau et al., 2021; Hinne et al., 2020). First, the researcher needs to specify (1) the models $\mathcal{H}_.$ under consideration, that is, the probability of data under the different parameter values that $\mathcal{H}_.$ allows, $p(\text{data} \mid \theta, \mathcal{H}_.)$ (i.e., the likelihood), and (2) prior distributions for the model parameters θ, that is, the relative plausibility of the parameter values before observing the data, $p(\theta \mid \mathcal{H}_.)$. In the case of meta-analyses, the data are usually represented by the observed effect sizes (y_k) and their standard errors (se_k) from $k = 1, \ldots, K$ individual studies. For example, a fixed-effect meta-analytic model $\mathcal{H}_0$ assuming absence of the mean effect (i.e., $\mu = 0$) and no across-study heterogeneity (i.e., $\tau = 0$), can be defined as:

$$\begin{aligned} \mathcal{H}_0 : \mu = 0,\ \tau = 0 \\ p(\text{data} \mid \theta_0, \mathcal{H}_0) : y_k \sim \text{Normal}(0, se_k), \end{aligned} \tag{1}$$

where θ_0 denotes vector of parameters (μ and τ) belonging to the model $\mathcal{H}_0$.

In contrast, a fixed-effect meta-analytic model $\mathcal{H}_1$ assuming the presence of the mean effect (i.e., $\mu \neq 0$) needs to also specify a prior distribution for μ, $f(.)$:

$$\begin{aligned} \mathcal{H}_1 : \mu \sim f(.),\ \tau = 0 \\ p(\text{data} \mid \theta_1, \mathcal{H}_1) : y_k \sim \text{Normal}(\mu, se_k). \end{aligned} \tag{2}$$

Once the models have been specified, Bayes' rule dictates how the observed data update the prior distributions to posterior distributions, for each model separately:

$$\begin{aligned} p(\theta_0 \mid \mathcal{H}_0, \text{data}) &= \frac{p(\theta_0 \mid \mathcal{H}_0)\, p(\text{data} \mid \theta_0, \mathcal{H}_0)}{p(\text{data} \mid \mathcal{H}_0)}, \\ p(\theta_1 \mid \mathcal{H}_1, \text{data}) &= \frac{p(\theta_1 \mid \mathcal{H}_1)\, p(\text{data} \mid \theta_1, \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_1)}, \end{aligned} \tag{3}$$

where the denominators denote the marginal likelihood, that is, the average probability of the data under a particular model. Specifically, marginal likelihoods are obtained by integrating the likelihood over the prior distribution for the model parameters:

$$\begin{aligned} p(\text{data} \mid \mathcal{H}_0) &= \int p(\text{data} \mid \theta_0, \mathcal{H}_0)\, p(\theta_0 \mid \mathcal{H}_0)\, \mathrm{d}\theta_0, \\ p(\text{data} \mid \mathcal{H}_1) &= \int p(\text{data} \mid \theta_1, \mathcal{H}_1)\, p(\theta_1 \mid \mathcal{H}_1)\, \mathrm{d}\theta_1. \end{aligned} \tag{4}$$

Together with the likelihood, the prior parameter distribution determines the model's predictions. The marginal likelihood therefore quantifies a model's predictive performance in light of the observed data. Consequently, the marginal likelihood plays a pivotal role in model comparison and hypothesis testing (Jefferys & Berger, 1992). The ratio of two marginal likelihoods is known as the Bayes factor (BF; Etz & Wagenmakers, 2017; Kass & Raftery, 1995; Rouder & Morey, 2019b; Wrinch & Jeffreys, 1921), and it indicates the extent to which one model outpredicts another; in other words, it grades the relative support that the models receive from the data. For example, the Bayes factor that assesses the relative predictive performance of the fixed-effect meta-analytic model $\mathcal{H}_0 : \mu = 0$ to that of the fixed-effect model $\mathcal{H}_1 : \mu \neq 0$ is

$$\text{BF}_{10} = \frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}. \tag{5}$$

The resulting BF_{10} represents the outcome of a Bayesian hypothesis test for the presence vs. absence of an effect for the fixed-effect meta-analytic models. Unlike the p-value in Neyman-Pearson hypothesis testing, the BF value can be interpreted as a continuous measure of evidence. A BF_{10} value larger than 1 indicates support for the alternative hypothesis (in the nominator) and a value lower than 1 indicates support for the null hypothesis (in the denominator). As a general rule of thumb, Bayes factors between 1 and 3 (between 1 and $1/3$) are regarded as anecdotal evidence, Bayes factors between 3 and 10 (between $1/3$ and $1/10$) are regarded as moderate evidence, and Bayes factors larger than 10 (smaller than $1/10$) are regarded as strong evidence in favor of (against) a hypothesis (e.g., Jeffreys, 1939, Appendix I; Lee and Wagenmakers, 2013, p. 105). While this rule of thumb can aid interpretation, Bayes factors are inherently continuous measures of the strength of evidence and any attempt at discretization inevitably involves a loss of information.

Next, we incorporate the prior model probabilities that later allow us to weight the posterior model estimates by posterior probability of the considered models. It is common practice to divide the prior model probability equally across the different model types (i.e., $p(\mathcal{H}_0) = p(\mathcal{H}_1) = 1/2$; e.g., Bartoš et al., 2021; Clyde et al., 2011; Gronau et al., 2021; Hoeting et al., 1999). To obtain the posterior model probabilities, we apply Bayes' rule one more time, now on the level of models instead of parameters:

$$\begin{aligned} p(\mathcal{H}_0 \mid \text{data}) &= \frac{p(\mathcal{H}_0)\, p(\text{data} \mid \mathcal{H}_0)}{p(\text{data})}, \\ p(\mathcal{H}_1 \mid \text{data}) &= \frac{p(\mathcal{H}_1)\, p(\text{data} \mid \mathcal{H}_1)}{p(\text{data})}. \end{aligned} \tag{6}$$

The common denominator,

$$p(\text{data}) = p(\text{data} \mid \mathcal{H}_0)\, p(\mathcal{H}_0) + p(\text{data} \mid \mathcal{H}_1)\, p(\mathcal{H}_1), \tag{7}$$

ensures that the posterior model probabilities sum to one.

The relative predictive performance of the rival models determines the update from prior to posterior model probabilities; in other words, models that predict the data well

receive a boost in posterior probability, and models that predict the data poorly suffer a decline (Rouder & Morey, 2019a; Wagenmakers et al., 2016). Thus, the Bayes factor quantifies the degree to which the data change the prior model odds to posterior model odds:

$$\underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\text{Bayes factor}} = \underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\text{Posterior odds}} \Big/ \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\text{Prior odds}} . \tag{8}$$

We can combine the posterior parameter distributions from the two fixed-effect meta-analytic models by weighting the distributions according to the posterior model probabilities (p. 387, Wrinch and Jeffreys, 1921; p. 222, Jeffreys, 1935). The resulting model-averaged posterior distribution can be defined as a mixture distribution,

$$p(\theta \mid \text{data}) = p(\theta_0 \mid \mathcal{H}_0, \text{data})\, p(\mathcal{H}_0 \mid \text{data}) + p(\theta_1 \mid \mathcal{H}_1, \text{data})\, p(\mathcal{H}_1 \mid \text{data}). \tag{9}$$

In RoBMA, the overall model ensemble is constructed from eight model types that represent the combination of the presence/absence of the effect, heterogeneity, and publication bias (modeled with two types of selection models in the original version of RoBMA; Maier, Bartoš, & Wagenmakers, 2022). With more than two models in play, Equations 5 and 9 can be expanded to accommodate the additional models. Specifically, the *inclusion Bayes factor* can be defined as a comparison between sets of models. For example, BF_{10} quantifies the evidence for presence vs. absence of the effect by the change from prior to posterior odds for the set of models that include the effect versus the set of models that exclude the effect:

$$\underbrace{\text{BF}_{10}}_{\substack{\text{Inclusion Bayes factor} \\ \text{for effect}}} = \underbrace{\frac{\sum_{i \in I} p(\mathcal{H}_i \mid \text{data})}{\sum_{j \in J} p(\mathcal{H}_j \mid \text{data})}}_{\substack{\text{Posterior inclusion odds} \\ \text{for models assuming effect}}} \Big/ \underbrace{\frac{\sum_{i \in I} p(\mathcal{H}_i)}{\sum_{j \in J} p(\mathcal{H}_j)}}_{\substack{\text{Prior inclusion odds} \\ \text{for models assuming effect}}} , \tag{10}$$

where $i \in I$ refers to models that include the effect and $j \in J$ refers to models that exclude the effect (Gronau et al., 2021; Hinne et al., 2020).[2] In the same way, we can also assess the relative predictive performance of any model compared to the rest of the ensemble.

[2] Both simple and inclusion Bayes factors are commonly denoted as BF since they are still Bayes factors and the same rules and interpretations apply to them.

Finally, the model-averaged posterior distribution of θ is defined as a mixture distribution of the posterior distributions of θ from each model $\mathcal{H}_n$ weighted by the posterior model probabilities,

$$p(\theta \mid \text{data}) = \sum_{n=1}^{N} p(\theta_n \mid \mathcal{H}_n, \text{data})\, p(\mathcal{H}_n \mid \text{data}). \tag{11}$$

To complete the model-averaged ensemble with multiple models corresponding to each component (e.g., two weight functions as a way of adjusting for publication bias in the original RoBMA), we maintain our prior indifference towards each of the hypotheses (e.g., presence/absence of the effect) by setting the prior model probabilities of all models that compose one of these two components to sum to 1/2. Often, the data contain enough information to assign posterior model probabilities to a class of similar models, largely washing out the effect of prior model probabilities on the model-averaged posterior distribution. If the data do not contain enough information, the model-averaged posterior distribution will be more affected by the choice of prior model probabilities. If researchers have diverging views on plausibility of different models, they can modify these prior model probabilities (e.g., by decreasing the prior model probabilities of fixed-effect models, but see Bartoš et al., 2021).

In contrast to classical meta-analytic statistics, the advantages of the Bayesian approach outlined above are that RoBMA can: (1) provide evidence for the absence of an effect (and therefore distinguish between "absence of evidence" and "evidence of absence", Keysers et al., 2020; Robinson, 2019); (2) update meta-analytic knowledge sequentially, thus addressing recent concern about accumulation bias (ter Schure & Grünwald, 2019); (3) incorporate expert knowledge; (4) retain and incorporate all uncertainty about parameters and models, without the need to make all-or-none choices; (5) emphasize the model outcomes that are most supported by the data, allowing it to flexibly adapt to scenarios with high heterogeneity and small sample sizes.

Publication Bias Adjustment Method 1: Selection Models

One class of publication bias correction methods are selection models (e.g., Hedges, 1992; Iyengar & Greenhouse, 1988; Maier, VanderWeele, et al., 2022; McShane et al., 2016; Vevea & Hedges, 1995).[3] In general, selection models estimate the relative probability that studies with p-values within pre-specified intervals were published as well as the corrected meta-analytic effect size. In other words, they are directly accounting for the missing data, based on the modelled relation between statistical significance and probability of publication. Selection models differ mostly in the specified weight function (such as 3PSM and 4PSM from Vevea and Hedges, 1995 and AK1 and AK2 from Andrews and Kasy, 2019), or are fit only to the statistically significant results (e.g., p-curve, Simonsohn et al., 2014; p-uniform, Van Assen et al., 2015).

Selection models based on p-values are attractive for several reasons. First, the models provide a plausible account of the data generating process – statistically non-significant studies are less likely to be published than statistically significant studies (Masicampo & Lalande, 2012; Rosenthal & Gaito, 1964; Wicherts, 2017). Second, in recent simulation studies the unrestricted versions of selection models performed relatively well (Carter et al., 2019; Hong & Reed, 2020).

Selection models can be specified flexibly according to the assumed publication process. For example, we can distinguish between two-sided selection (i.e., significant studies are published regardless of the direction of the effect) and one-sided selection (only significant studies in the "correct" direction are preferentially reported). In the previous implementation of RoBMA, the selection models assumed two-sided selection, either at a p-value cutoff of 0.05 or also at a marginally significant cutoff of 0.10 (Maier, Bartoš, & Wagenmakers, 2022). In this paper, we extend RoBMA by adding 4 weight functions that encompass more ways in which the selection process might operate. The added weight

[3] See Copas (1999), Copas and Li (1997), and Copas and Shi (2001) for selection models based on effect sizes and standard errors.

functions assume one-sided selection for positive effect sizes with cutoffs on significant, marginally significant, and/or p-values corresponding to the expected effect size direction. Overall the six included weight functions are:

1. Two-sided (already included in RoBMA)
    (a) p-value cutoffs = 0.05
    (b) p-value cutoffs = 0.05 & 0.10
2. One-sided (new in RoBMA-PSMA)
    (a) p-value cutoffs = 0.05
    (b) p-value cutoffs = 0.025 & 0.05
    (c) p-value cutoffs = 0.05 & 0.50
    (d) p-value cutoffs = 0.025 & 0.05 & 0.50

Example - Feeling the Future

We illustrate this extended version of RoBMA on studies from the infamous 2011 article "Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect" (Bem, 2011). Across a series of nine experiments, Bem (2011) attempted to show that participants are capable of predicting the future through the anomalous process of precognition. In response to a methodological critique, Bem et al. (2011) later conducted a meta-analysis on the nine reported experiments in order to demonstrate that the experiments jointly contained strong support for existence of the effect. Publication of such an implausible result in the flagship journal of social psychology ignited an intense debate about replicability, publication bias, and questionable research practices in psychology (Simmons et al., 2011).

We analyze the data as described by Bem et al. (2011) in Table 1 with the updated version of `RoBMA` R package (Bartoš & Maier, 2021). For illustration, we specify the

publication bias adjustment part with the six weight functions outlined above. We use the default prior distributions for the effect size and heterogeneity (standard normal and inverse-gamma, respectively, as in the original version of RoBMA, see Appendix B for details). Internally, the package transforms the priors, the supplied Cohen's d, and their standard errors to the Fisher's z scale.[4] The estimates are transformed back to Cohen's d scale for ease of interpretation. R code and data for reproducibility are available on OSF https://osf.io/fgqpc/.

Our results do not provide notable evidence either for or against the presence of the anomalous effect of precognition: the model-averaged Bayes factor equals $\text{BF}_{10} = 1.91$ and the posterior model-averaged mean estimate of $\mu = 0.097$, 95% CI $[0.000, 0.232]$.[5] Figure 1 shows posterior model-averaged estimated weights with a re-scaled x-axis for easier readability. Because the meta-analysis is based on only nine estimates, the uncertainty in the estimated weights is relatively high.

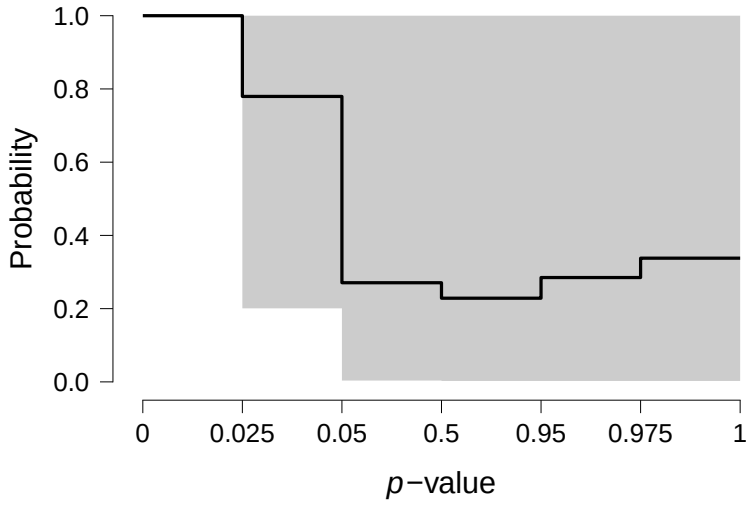
These results are an improvement from the original RoBMA implementation (with only two two-sided weight functions) that showed strong support for the effect: $\text{BF}_{10} = 97.89$, $\mu = 0.149$, 95% CI $[0.053, 0.240]$. The substantial difference in conclusions

[4] We use the Fisher's z scale for model fitting because it makes standard errors and effect sizes independent under the model without publication bias. This is an important prerequisite to test for the presence/absence of publication bias with the PET-PEESE models introduced later. The effect sizes and standard errors are transformed using the popular formulas for effect size transformations. See the Appendix in Haaf and Rouder (2020) for proof that Fisher's z is a variance stabilizing transformation for Cohen's d. Prior distributions are linearly re-scaled from Cohen's d to Fisher's z, in the same manner as in `metaBMA` R package (Heck et al., 2019).

[5] The reported lower bound of the credible interval of 0.000 is not a coincidence and will be encountered more often than in conventional frequentist methods. The 0.000 lower bound is a consequence of averaging posterior estimates across all models, including models that specify $\mu = 0$ (see Equation 9). When a notable proportion of the posterior model probabilities is accumulated by models assuming the absence of the effect, the model-averaged posterior distribution for effect size will include a sizeable point mass at $\mu = 0$. Consequently, the lower credible internal bound is then "shrunk" to 0.

Figure 1

The Model-Averaged Weight Function with 95% CI for Bem (2011).



Note. Results are model-averaged across the whole model ensemble, including models assuming no publication bias ($\omega = 1$).

between the original RoBMA and the RoBMA with four additional weight functions is due to the inclusion of one-sided selection models that seems to provide a better description for the Bem studies. More importantly, with the additional four weight functions RoBMA provides only moderate evidence for the effect even when adopting the $N(0, 0.304^2)$ prior distribution for effect size recommended by Bem et al. (2011): $\text{BF}_{10} = 5.59$, $\mu = 0.122$, 95% CI $[0.000, 0.234]$.

Limitations of Selection Models

While selection models have been shown to perform well in comparison to other methods in simulation studies (e.g., Carter et al., 2019; McShane et al., 2016), they often insufficiently adjust for publication bias when applied to actual meta-analytic data (e.g., Kvarven et al., 2020; Mathur & VanderWeele, 2019). This discrepancy arises because the simulation studies assume that the selection model is an accurate reflection of the true data-generating process; that is, the synthetic data obey a selection process that stipulates

publication probability is (a) based solely on p-values rather than on effect sizes; (b) based solely on a discretized p-value interval, within which the probability of publication is constant. The simulation studies largely ignore the possibility of model misspecification and therefore provide an upper bound on model performance (Carter et al., 2019). A key strength of the Bayesian model-averaging approach is that it can incorporate any number of models, increasing robustness and decreasing the potentially distorting effects of model misspecification. Therefore, we extend RoBMA with another method that adjusts for publication bias in an entirely different way – PET-PEESE (Stanley & Doucouliagos, 2014).

Publication Bias Adjustment Method 2: PET-PEESE

A prominent class of alternative approaches to the selection models outlined above are methods that adjust for publication bias by adjusting for small-study effects by estimating the relationship between effect sizes and their standard errors (e.g., Egger et al., 1997). The most well-known approaches include trim and fill (Duval & Tweedie, 2000) and PET-PEESE (Stanley & Doucouliagos, 2014). Here, we focus only on PET-PEESE since its regression-based framework, which fits the model to all observed studies, allows us to compare the model fit directly to the selection model-based approaches.

PET-PEESE method is an attractive addition to the RoBMA methodology since it often performs better than selection models in meta-analytic applications (Kvarven et al., 2020; for applications in the field of ego depletion and antidepressant effectiveness see Carter and McCullough, 2014 and Moreno et al., 2009, respectively). PET-PEESE is a conditional (two-step) estimator composed of two models, PET model (i.e., Precision Effect Test) that is correctly specified when the effect is absent and PEESE model (i.e., Precision Effect Estimate with Standard Error) that provides a better approximation when the effect is present (Stanley & Doucouliagos, 2014). The individual PET and PEESE models are the linear and the quadratic meta-regression approximations, respectively, to the incidentally truncated selection model (Stanley & Doucouliagos, 2014). The choice between the PET

and the PEESE model proceeds as follows: the test for the effect size coefficient based on PET (with $\alpha = 0.10$ for model selection only) is used to decide whether the PET ($p > \alpha$) or the PEESE ($p < \alpha$) effect size estimator is employed (Stanley, 2017).

In order to add PET and PEESE models as a way of adjusting for publication bias with RoBMA, we modify them in the following way. Instead of following PET-PEESE conditional selection of either PET or PEESE as proposed by Stanley and Doucouliagos (2014), we include both PET and PEESE models, separately, in the RoBMA ensemble (alongside the weight functions model) and model-average over the entire ensemble. Furthermore, instead of using an unrestricted weighted least squares estimator (Stanley & Doucouliagos, 2014), we specify both fixed-effects and random-effects versions of these models, for consistency with the remaining RoBMA models. Consequently, the PET and PEESE models implemented in RoBMA correspond to meta-regressions of effect size on either the standard errors or the variances with conventional fixed-effects and random-effect flavors (see Equation 13 in Appendix A).

In sum, we created a new RoBMA ensemble adjusting for publication bias using PET and PEESE models. Instead of using the model estimates conditionally, we model-average across the fixed- and random-effects PET and PEESE models assuming either absence or presence of the effect and the corresponding fixed- and random-effects models without publication bias adjustment.

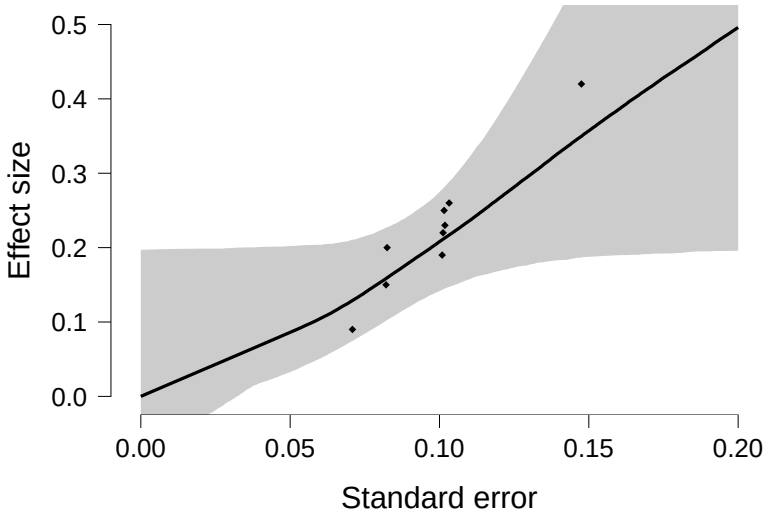
Example - Feeling the Future

We revisit the Bem (2011) example. For illustration, we now specify only the PET and PEESE models as the publication bias adjustment part of the RoBMA ensemble (we include the six weight functions specified above in the next subsection). Again, we use the RoBMA package with the same default priors for effect size and heterogeneity; we assign Cauchy$(0, 1)$ and Cauchy$(0, 5)$ priors restricted to the positive range to the regression coefficients on standard errors and variances, respectively (see Appendix B for details).

RoBMA version model-averaging across the PET and PEESE models provides moderate evidence for the absence of an effect, $BF_{10} = 0.226$ (the reciprocal quantifying the evidence for the null hypothesis, $BF_{01} = 4.42$), with the posterior model-averaged mean estimate $\mu = 0.013$, 95% CI $[-0.078, 0.197]$. Figure 2 shows the estimated relationship between standard errors and effect sizes, where the effect size at standard error 0 corresponds to the posterior model-averaged bias-corrected estimate.

Figure 2

The Relationship between the Standard Errors and Model-Averaged Effect Size Estimate with 95% CI for Bem (2011).



Note. Results are model-averaged across the entire model ensemble. Models assuming no publication bias have both PET and PEESE coefficients set to 0. Black diamonds correspond to the individual study estimates and standard errors.

In this application, these results seem to provide an even better adjustment than the RoBMA version of selection models discussed previously. Furthermore, the RoBMA ensemble with PET-PEESE models resolves a seeming inconsistency in the original conditional PET-PEESE estimator. The frequentist PET model resulted in a significant negative effect size, $\mu = -0.182$, $t(7) = -3.65$, $p = 0.008$, indicating that the effect size estimate from PEESE should be used, $\mu = 0.024$, $t(7) = 0.86$, $p = 0.418$, which however is

not notably different from zero. In addition, the RoBMA ensemble with PET-PEESE does not provide evidence for precognition even under the more informed $N(0, 0.304^2)$ prior distribution for effect size recommended in Bem et al. (2011): $\text{BF}_{10} = 0.670$, $\mu = 0.028$, 95% CI $[-0.160, 0.215]$. However, under this more informed prior, the data no longer provide moderate evidence against precognition.

Limitations of PET-PEESE

While PET-PEESE shows less bias and overestimation compared to other bias correction methods (Kvarven et al., 2020) its key limitation is that the estimates can have very high variability. In simulation studies, PET-PEESE can have high RMSE (root mean square error; Carter et al., 2019; Hong & Reed, 2020; Maier, Bartoš, & Wagenmakers, 2022). Therefore, when PET-PEESE based models are applied to an area of research for which they are ill-suited, the resulting estimates may be inaccurate and unreliable. Stanley (2017) shows how the performance of PET-PEESE can be especially problematic at very high levels of heterogeneity ($\tau \geq .5$), with low number of studies (i.e., $k \leq 10$), and under uniformly low power.

Combining Selection Models and PET-PEESE

In order to obtain the best of both PET-PEESE and selection models, we combine them into an overarching model: RoBMA-PSMA. Specifically, RoBMA-PSMA includes the 6 weight functions outlined in the section "Publication Bias Adjustment Methods 1: Selection Models" (assuming either presence or absence of the effect and heterogeneity, this yields 24 models) as well as the two PET and PEESE regression models outlined in the section "Publication Bias Adjustment Methods 2: PET-PEESE" section (assuming either presence or absence of the effect and heterogeneity, this yields 8 models). We set the prior probability for the publication bias-adjusted models to 0.5 (cf. Maier, Bartoš, and Wagenmakers, 2022) and divide this 0.5 probability equally across selection models and PET-PEESE models (cf. Jeffreys, 1961, p. 47). Finally, adding models assuming absence

of the publication bias (assuming either presence or absence of the effect and heterogeneity, this yields 4 models) results in a total of $24 + 8 + 4 = 36$ models that together comprise RoBMA-PSMA. The entire model ensemble is summarized in Table 1.

As mentioned above, RoBMA-PSMA draws inference about the data by considering all models simultaneously. Specific inferences can be obtained by interrogating the model ensemble and focusing on different model classes. Concretely, the evidence for presence vs. absence of the effect is quantified by the inclusion Bayes factor BF_{10} (Equation 10) obtained by comparing the predictive performance of models assuming the effect is present ($i = 19, \ldots, 36$ in Table 1) to that of models assuming the effect is absent ($j = 1, \ldots, 18$ in Table 1). In the Bem example, substituting the prior and posterior model probabilities from Table 1 yields $\text{BF}_{10} = 0.479$. This Bayes factor indicates that the posterior inclusion odds for the models assuming the effect is present are slightly lower than the prior inclusion odds. In other words, models assuming that the effect is absent predicted the data about $1/0.479 \approx 2.09$ times better than models assuming the effect is present. This result aligns with the common scientific understanding of nature, which the presence of precognition would effectively overturn.

The remaining Bayes factors are calculated similarly. The Bayes factor for the presence vs. absence of heterogeneity, BF_{rf}, compares the predictive accuracy of models assuming heterogeneity ($i = 10, \ldots, 18$ and $28, \ldots 36$ in Table 1) with models assuming homogeneity ($j = 1, \ldots, 9$ and $19, \ldots 27$ in Table 1). With $\text{BF}_{rf} = 0.144$, the data disfavor the models assuming heterogeneity; that is, the data are $\text{BF}_{fr} = 1/0.144 \approx 6.94$ times more likely to occur under homogeneity than under heterogeneity. Analogously, the Bayes factor for the presence vs. absence of publication bias, BF_{pb}, compares predictive performance of models assuming publication bias is present ($i = 2, \ldots, 9, 11, \ldots 18, 20 \ldots 27,$ and $29, \ldots 36$ in Table 1) to that of models assuming publication bias is absent ($j \in 1, 10, 19,$ and 28 in Table 1). Here, the results show strong support in favor of the models assuming publication bias is present, $\text{BF}_{pb} = 16.31$.

Table 1

RoBMA-PSMA Model Ensemble Together with Prior Parameter Distributions (Columns 1-3), Prior Model Probabilities, (Column 4), and Posterior Model Probabilities (Column 5) Based on an Application to the Data from Bem (2011).

i	Effect Size	Heterogeneity	Publication Bias	Prior prob.	Posterior prob.
1	$\mu = 0$	$\tau = 0$	None	0.125	0.000
2	$\mu = 0$	$\tau = 0$	$\omega_{\text{Two-sided}(.05)} \sim \text{CumDirichlet}(1, 1)$	0.010	0.000
3	$\mu = 0$	$\tau = 0$	$\omega_{\text{Two-sided}(.1,.05)} \sim \text{CumDirichlet}(1, 1, 1)$	0.010	0.000
4	$\mu = 0$	$\tau = 0$	$\omega_{\text{One-sided}(.05)} \sim \text{CumDirichlet}(1, 1)$	0.010	0.012
5	$\mu = 0$	$\tau = 0$	$\omega_{\text{One-sided}(.05,.025)} \sim \text{CumDirichlet}(1, 1, 1)$	0.010	0.034
6	$\mu = 0$	$\tau = 0$	$\omega_{\text{One-sided}(.5,.05)} \sim \text{CumDirichlet}(1, 1, 1)$	0.010	0.001
7	$\mu = 0$	$\tau = 0$	$\omega_{\text{One-sided}(.5,.05,.025)} \sim \text{CumDirichlet}(1, 1, 1, 1)$	0.010	0.004
8	$\mu = 0$	$\tau = 0$	$\text{PET} \sim \text{Cauchy}(0, 1)_{[0,\infty]}$	0.031	0.281
9	$\mu = 0$	$\tau = 0$	$\text{PEESE} \sim \text{Cauchy}(0, 5)_{[0,\infty]}$	0.031	0.254
10	$\mu = 0$	$\tau \sim \text{InvGamma}(1, 0.15)$	None	0.125	0.000
11	$\mu = 0$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\omega_{\text{Two-sided}(.05)} \sim \text{CumDirichlet}(1, 1)$	0.010	0.000
12	$\mu = 0$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\omega_{\text{Two-sided}(.1,.05)} \sim \text{CumDirichlet}(1, 1, 1)$	0.010	0.000
13	$\mu = 0$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\omega_{\text{One-sided}(.05)} \sim \text{CumDirichlet}(1, 1)$	0.010	0.014
14	$\mu = 0$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\omega_{\text{One-sided}(.05,.025)} \sim \text{CumDirichlet}(1, 1, 1)$	0.010	0.020
15	$\mu = 0$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\omega_{\text{One-sided}(.5,.05)} \sim \text{CumDirichlet}(1, 1, 1)$	0.010	0.006
16	$\mu = 0$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\omega_{\text{One-sided}(.5,.05,.025)} \sim \text{CumDirichlet}(1, 1, 1, 1)$	0.010	0.010
17	$\mu = 0$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\text{PET} \sim \text{Cauchy}(0, 1)_{[0,\infty]}$	0.031	0.021
18	$\mu = 0$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\text{PEESE} \sim \text{Cauchy}(0, 5)_{[0,\infty]}$	0.031	0.017
19	$\mu \sim \text{Normal}(0,\ 1)$	$\tau = 0$	None	0.125	0.051
20	$\mu \sim \text{Normal}(0,\ 1)$	$\tau = 0$	$\omega_{\text{Two-sided}(.05)} \sim \text{CumDirichlet}(1, 1)$	0.010	0.007
21	$\mu \sim \text{Normal}(0,\ 1)$	$\tau = 0$	$\omega_{\text{Two-sided}(.1,.05)} \sim \text{CumDirichlet}(1, 1, 1)$	0.010	0.031
22	$\mu \sim \text{Normal}(0,\ 1)$	$\tau = 0$	$\omega_{\text{One-sided}(.05)} \sim \text{CumDirichlet}(1, 1)$	0.010	0.030
23	$\mu \sim \text{Normal}(0,\ 1)$	$\tau = 0$	$\omega_{\text{One-sided}(.05,.025)} \sim \text{CumDirichlet}(1, 1, 1)$	0.010	0.035
24	$\mu \sim \text{Normal}(0,\ 1)$	$\tau = 0$	$\omega_{\text{One-sided}(.5,.05)} \sim \text{CumDirichlet}(1, 1, 1)$	0.010	0.018
25	$\mu \sim \text{Normal}(0,\ 1)$	$\tau = 0$	$\omega_{\text{One-sided}(.5,.05,.025)} \sim \text{CumDirichlet}(1, 1, 1, 1)$	0.010	0.022
26	$\mu \sim \text{Normal}(0,\ 1)$	$\tau = 0$	$\text{PET} \sim \text{Cauchy}(0, 1)_{[0,\infty]}$	0.031	0.047
27	$\mu \sim \text{Normal}(0,\ 1)$	$\tau = 0$	$\text{PEESE} \sim \text{Cauchy}(0, 5)_{[0,\infty]}$	0.031	0.046
28	$\mu \sim \text{Normal}(0,\ 1)$	$\tau \sim \text{InvGamma}(1, 0.15)$	None	0.125	0.007
29	$\mu \sim \text{Normal}(0,\ 1)$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\omega_{\text{Two-sided}(.05)} \sim \text{CumDirichlet}(1, 1)$	0.010	0.001
30	$\mu \sim \text{Normal}(0,\ 1)$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\omega_{\text{Two-sided}(.1,.05)} \sim \text{CumDirichlet}(1, 1, 1)$	0.010	0.003
31	$\mu \sim \text{Normal}(0,\ 1)$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\omega_{\text{One-sided}(.05)} \sim \text{CumDirichlet}(1, 1)$	0.010	0.005
32	$\mu \sim \text{Normal}(0,\ 1)$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\omega_{\text{One-sided}(.05,.025)} \sim \text{CumDirichlet}(1, 1, 1)$	0.010	0.005
33	$\mu \sim \text{Normal}(0,\ 1)$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\omega_{\text{One-sided}(.5,.05)} \sim \text{CumDirichlet}(1, 1, 1)$	0.010	0.003
34	$\mu \sim \text{Normal}(0,\ 1)$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\omega_{\text{One-sided}(.5,.05,.025)} \sim \text{CumDirichlet}(1, 1, 1, 1)$	0.010	0.004
35	$\mu \sim \text{Normal}(0,\ 1)$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\text{PET} \sim \text{Cauchy}(0, 1)_{[0,\infty]}$	0.031	0.004
36	$\mu \sim \text{Normal}(0,\ 1)$	$\tau \sim \text{InvGamma}(1, 0.15)$	$\text{PEESE} \sim \text{Cauchy}(0, 5)_{[0,\infty]}$	0.031	0.004

Note. μ corresponds to the effect size parameter, τ to the heterogeneity parameter, ω to the weight parameters with an appropriate selection process (either one or two-sided with given cutoffs), PET to the regression coefficient on the standard errors, and PEESE to the regression coefficient on variances. All prior distributions are specified on the Cohen's d scale.

Model-averaging can also be used to compare the different types of publication bias adjustment methods. Specifically, the predictive performance of the selection models ($i = 2, \ldots, 7, 11, \ldots, 16, 20, \ldots, 25,$ and $29, \ldots, 34$) may be contrasted to that of the PET-PEESE models ($j = 8, 9, 17, 18, 26, 27, 35,$ and 36), yielding BF = 0.397 (cf. Equation 10); this result indicates that the posterior probability increases more for the PET-PEESE models ($0.25 \rightarrow 0.675$) than it does for the selection models ($0.25 \rightarrow 0.268$), especially selection model assuming one-sided selection that were better supported by the data ($0.166 \rightarrow 0.225$) than the two-sided selection models ($0.083 \rightarrow 0.043$). However, this Bayes factor only modestly favors the PET-PEESE models, and consequently the results from the selection models also contribute substantially towards the final posterior model-averaged estimate.

Predictive performance of individual models may be contrasted to that of the rest of the ensemble (cf. Equation 10). For Bem (2011), the data most strongly supported the PET and PEESE models assuming no effect and no heterogeneity, BF = 12.13 and BF = 10.57, respectively – the corresponding model probabilities increased from 0.031 to 0.280 and 0.255.

The posterior model-averaged effect size estimate μ is obtained by combining the 36 estimates across all models in the ensemble, weighted according to their posterior model probabilities. Some of the models assume the effect is absent, and concentrate all prior probability mass on $\mu = 0$; therefore, the model-averaged posterior distribution is a combination of a "spike" at 0 and a mixture of continuous probability densities that correspond to the alternative models. When the alternative models are strongly supported by the data, the impact of the spike is minimal and the model-averaged posterior distribution reduces to a mixture of continuous densities. In the Bem (2011) example, RoBMA-PSMA gives a posterior model-averaged mean estimate $\mu = 0.038$, 95% CI $[-0.034, 0.214]$ (cf. Equation 9). The posterior model-averaged estimates for the remaining parameters, e.g., the heterogeneity estimate τ or the publication weights ω, are obtained

similarly.

The overall results would, again, remain similar even when using the Bem et al. (2011) more informed prior distribution for effect size, $N(0, 0.304^2)$: $\text{BF}_{10} = 1.41$, $\mu = 0.067$, 95% CI $[-0.111, 0.226]$. These results are in line with failed replication studies (e.g., Galak et al., 2012; Ritchie et al., 2012; Schlitz et al., 2021; Wagenmakers et al., 2012), evidence of QRPs (Alcock, 2011; Francis, 2012; Schimmack, 2012), and common sense (Hoogeveen et al., 2020; see also Bem et al., 2011; Rouder and Morey, 2011; Schimmack, 2015, 2018; Wagenmakers et al., 2011).

Evaluating RoBMA Through Registered Replication Reports

Kvarven et al. (2020) compared the effect size estimates from 15 meta-analyses of psychological experiments to the corresponding effect size estimates from Registered Replication Reports (RRR) of the same experiment.[6] RRRs are accepted for publication independently of the results and should be unaffected by publication bias. The original meta-analyses reveal considerable heterogeneity; thus, any single RRR is unlike to directly correspond to the true mean meta-analytic effect size. As a result, the comparison of meta-analysis results to RRRs will inflate RMSE and can be considered a highly conservative way of evaluating bias detection methods. However, when averaged over 15 RRRs, we would expect little systematic net heterogeneity and a notable reduction in aggregate bias. In this way, average bias adjusted estimates should randomly cluster around the average of the RRR estimates. In other words, we would expect little overall bias, relative to RRRs. Hence the comparison to RRRs can be used to gauge the performance of publication bias adjustment methods, while keeping in mind that the studies are heterogeneous and limited in number. Kvarven et al. (2020) found that

[6] This RRR category includes "replications published according to the 'registered replication report' format in the journals 'Perspectives on Psychological Science' and 'Advances in Methods and Practices in Psychological Science'; and (2) The 'Many Labs' projects in psychology" (Kvarven et al., 2020, p. 424).

conventional meta-analysis methods resulted in substantial overestimation of effect size. In addition, Kvarven et al. (2020) examined three popular bias detection methods: trim and fill (TF; Duval and Tweedie, 2000), PET-PEESE (Stanley & Doucouliagos, 2014), and 3PSM (Hedges, 1992; Vevea & Hedges, 1995). The best performing method was PET-PEESE; however, its estimates still have notable RMSE.

Here we use the data analyzed by Kvarven et al. (2020) as one way of comparing the performance of RoBMA-PSMA in relation to a series of alternative publication bias correction methods. These methods include those examined by Kvarven et al. (2020) –PET-PEESE, 3PSM, and TF– as well as a set of seven other methods (cf. Hong and Reed, 2020): 4PSM (Vevea & Hedges, 1995), AK1 and AK2 (Andrews & Kasy, 2019), p-curve (Simonsohn et al., 2014), p-uniform (Van Assen et al., 2015), WAAP-WLS (Stanley et al., 2017), and endogenous kink (EK; Bom & Rachinger, 2019). For completeness, we also show results for the original implementation of RoBMA-old (Maier, Bartoš, & Wagenmakers, 2022). The RoBMA-old, 3PSM, 4PSM, AK1, AK2, p-curve, and p-uniform can be viewed as selection models operating on p-values that mostly differ in thresholds of the weight function and estimation algorithm. The PET-PEESE, TF, and EK can be viewed as methods correcting for publication bias based on relationship between effect sizes and standard errors. Finally, RoBMA-PSMA is a method that combines both types of publication bias corrections.

Following Kvarven et al. (2020), we report all meta-analytic estimates on the Cohen's d scale, with one exception for a meta-analysis that used Cohen's q scale. As in the Bem example, RoBMA internally transforms effect sizes from the Cohen d scale to the Fisher z scale.[7] Each method is evaluated on the following five metrics (cf. Kvarven et al., 2020): (1) false positive rate (FPR), that is, the proportion of cases where the RRR fails to reject the null hypothesis (i.e., $p > .05$) whereas the meta-analytic method concludes that

[7] We also tried to estimate the remaining methods on the Fisher z scale; however, doing so reduced the performance of some of the other methods.

Table 2

Performance of 13 Publication Bias Correction Methods for the Kvarven et al. (2020) Test Set Comprised of 15 Meta-analyses and 15 Corresponding "Gold Standard" Registered Replication Reports (RRR).

Method	FPR / Undecided	FNR / Undecided	OF	Bias	RMSE
RoBMA-PSMA	0.143 / 0.857	0.000 / 0.750	1.160	0.026	0.164
AK2	*0.000 / —*	*0.250 / —*	*1.043*	*-0.070*	*0.268*
PET-PEESE	0.143 / —	0.500 / —	1.307	0.050	0.256
EK	0.143 / —	0.500 / —	1.399	0.065	0.283
RoBMA-old	0.714 / 0.286	0.000 / 0.000	2.049	0.171	0.218
4PSM	0.714 / —	0.500 / —	1.778	0.127	0.268
3PSM	0.714 / —	0.125 / —	2.193	0.195	0.245
TF	*0.833 / —*	*0.000 / —*	*2.315*	*0.206*	*0.259*
AK1	0.857 / —	0.000 / —	2.352	0.221	0.264
p-uniform	0.500 / —	0.429 / —	2.375	0.225	0.288
p-curve			2.367	0.223	0.289
WAAP-WLS	0.857 / —	0.125 / —	2.463	0.239	0.295
Random Effects (DL)	1.000 / —	0.000 / —	2.586	0.259	0.310

Note. FPR / Undecided = false positive rate / undecided evidence under no effect, FNR/ Undecided = false negative rate/ undecided evidence under an effect, OF = overestimation factor, and RMSE = root mean square error. The results in *gray italic* are conditional on convergence: trim and fill did not converge in one case and AK2 did not converge in 10 cases. The rows are ordered based on combined log scores performance of the abs(log(OF), abs(Bias), and RMSE (not shown).

the data offer support for the presence of the effect (i.e., $p < 0.05$ or $BF_{10} > 10$); (2) false negative rate (FNR), that is, the proportion of cases where the RRR rejects the null hypothesis (i.e., $p < .05$) whereas the meta-analytic method fails to reject the null/finds

evidence for the absence of the effect (i.e., $p > 0.05$ or $BF_{10} < 1/10$)[8]; (3) overestimation factor (OF), that is, the meta-analytic mean effect size divided by the RRR mean effect size; (4) bias, that is, the mean difference between the meta-analytic and RRR effect size estimates; and (5) root mean square error (RMSE), that is, the square root of the mean of squared differences between the meta-analytic and RRR effect size estimates. Note that when evaluating the methods' qualitative decisions (i.e., FPR and FNR), the RoBMA methods do not necessarily lead to a strong claim about the presence or absence of the effect; in the Bayesian framework, there is no need to make an all-or-none decision based on weak evidence, and here we have defined an in-between category of evidence that does not allow a confident conclusion (i.e., Undecided, $1/10 < BF_{10} < 10$; for a discussion on the importance of this in-between category see Robinson, 2019). Furthermore, selecting a different significance level or Bayes factor thresholds would lead to different false positive and false negative rates.

The main results are summarized in Table 2. Evaluated across all metrics simultaneously, RoBMA-PSMA generally outperforms the other methods. RoBMA-PSMA has the lowest bias, the second-lowest RMSE, and the second lowest overestimation factor. The only methods that perform better in one of the categories (i.e., AK2 with the lowest overestimation factor; PET-PEESE and EK with the second and third lowest bias, respectively), showed considerably larger RMSE, and AK2 converged in only 5 out of 15 cases. Furthermore, RoBMA-PSMA resulted in conclusions that are qualitatively similar to those from the RRR studies. Specifically, for cases where the RRR was statistically significant, RoBMA-PSMA never showed evidence for the absence of the effect (i.e., FNR $= 0/8 = 25\%$) but often did not find compelling evidence for the presence of the effect either (i.e., Undecided $= 6/8 = 75\%$). Furthermore, for cases where the RRR was not statistically

[8] This corresponds to the definition of the FPR and FNR indices from Kvarven et al. (2020). However, it is important to note that a statistically non-significant result is not generally a valid reason to "accept" the null hypothesis (e.g., Aczel et al., 2018; Goodman, 2008).

significant, RoBMA-PSMA showed evidence for the presence of the effect only once (i.e., FPR = $1/7 \approx 14.3\%$) and did not find compelling evidence for the absence of the effect in the remaining meta-analyses (i.e., Undecided = $6/7 \approx 85.7\%$). After adjusting for publication selection bias with RoBMA, the original meta-analyses often did not contain sufficient evidence for firm conclusions about the presence vs. absence of the effect.[9] This highlights the oft-hidden reality that the data at hand do not necessarily warrant strong conclusions about the phenomena under study; consequently, a final judgement needs to be postponed until more data accumulates.

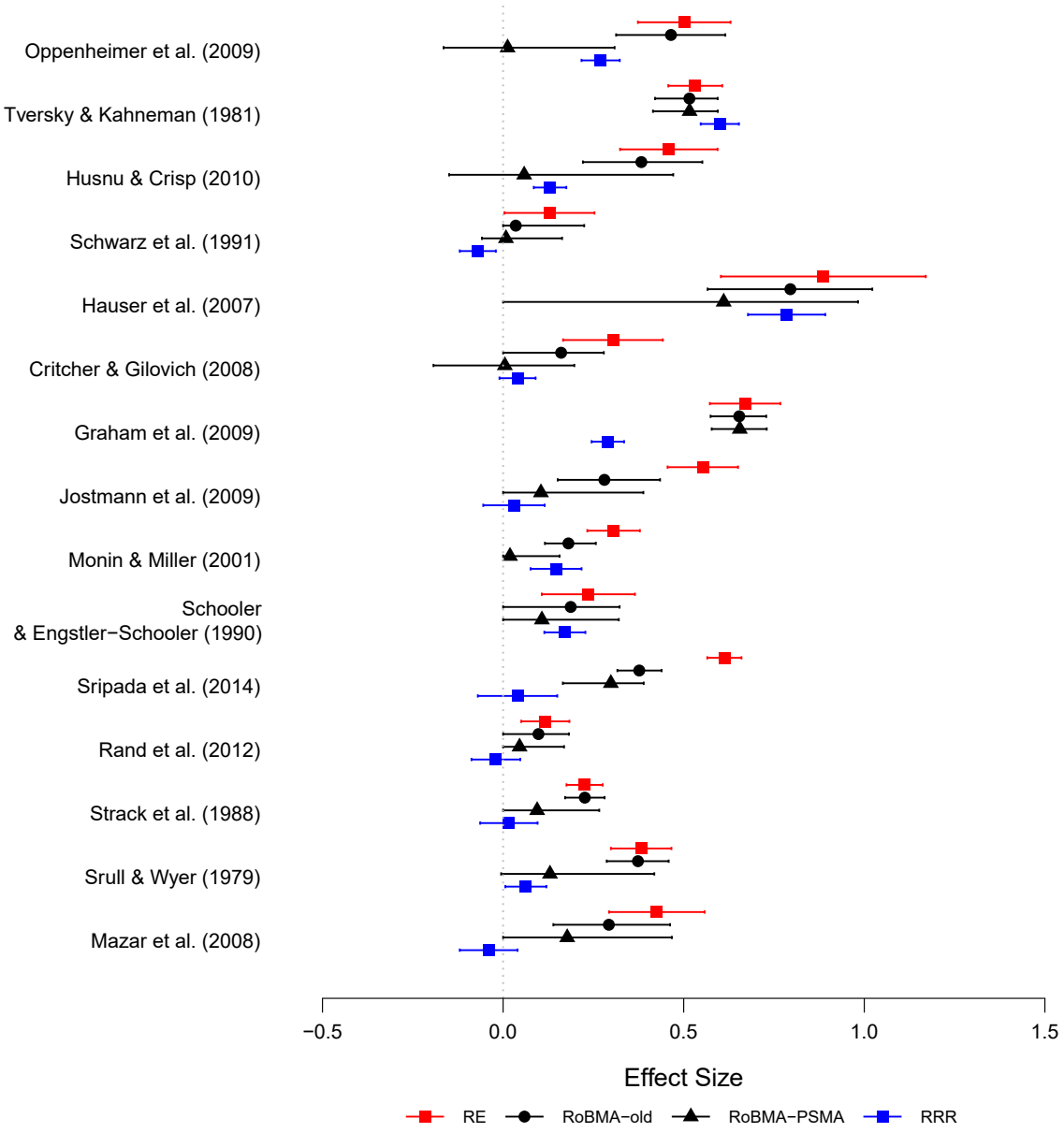
Figure 3 shows the effect size estimates from the RRRs for each of the 15 cases, together with the estimates from a random effects meta-analysis and the posterior model-averaged estimates from RoBMA and RoBMA-PSMA (figures comparing all methods for each RRR are available in the "Kvarven et al/estimates figures" folder in the online supplementary materials at https://osf.io/fgqpc/files/). Because RoBMA-PSMA corrects for publication bias, its estimates are shrunken toward zero. In addition, the estimates from RoBMA-PSMA also come with wider credible intervals (reflecting the additional uncertainty about the publication bias process) and are generally closer to the RRR results. The most anomalous case concerns the Graham et al. (2009) study, where all four methods yield similar intervals, but the RRR shows an effect size that is twice as small. This result may be due to cultural differences and the choice of the social or economic dimension that all contributed to heterogeneity in the original meta-analysis (Kivikangas et al., 2021).

Appendix C demonstrates robustness of our findings by estimating RoBMA under different parameter prior distributions. Appendix D presents a non-parametric bootstrap analysis of the RRR comparison, showing high uncertainty in the FPR and FNR, but

[9] Note that this is not a general pattern and RoBMA often results in compelling evidence, either in favor of the absence or in favor of the presence of an effect (e.g., Maier, Bartoš, Oh, et al., 2022; Maier, Bartoš, Stanley, et al., 2022).

Figure 3

Effect Size Estimates with 95% CIs from a random-effects meta-analysis, two RoBMA models, and the RRR for the 15 Experiments Included in Kvarven et al. (2020).



Note. Estimates are reported on the Cohen's d scale.

qualitatively robust conclusions about the overestimation factor, bias, and RMSE. Appendix E demonstrates that our findings are not a result of a systematic underestimation of effect sizes by estimating RoBMA on 28 sets of Registered Replication

Reports from Many Labs 2 (Klein et al., 2018).

Evaluating RoBMA Through Simulation Studies

We evaluate the performance of the newly developed RoBMA methods using simulation studies (cf. Hong and Reed, 2020). As in Hong and Reed (2020), we tested the methods in four simulation environments, namely those developed by Stanley et al. (2017, SD&I), Alinaghi and Reed (2018, A&R), Bom and Rachinger (2019, B&R), and Carter et al. (2019, CSG&H). These environments differ in terms of assumptions concerning effect sizes, heterogeneity, sample sizes, and publication bias; moreover, CSG&H include questionable research practices (QRP; John et al., 2012). Briefly, the SD&I environment relates to the settings usually found in medicine where a difference between two groups is assessed with either continuous or dichotomous outcomes. The A&R environment is similar to the settings encountered in economics and business and consists of relationships between two continuous variables with multiple estimates originating from a single study. The B&R environment considers situations where regression coefficients are routinely affected by an omitted-variables bias. The CSG&H environment is most typical for psychology with effect sizes quantifying differences in a continuous measure between groups. For each condition from each of the four simulation environments, Hong and Reed (2020) generated 3,000 synthetic data sets that were then analyzed by all of the competing methods.

Here we used the code, data, and results publicly shared by Hong and Reed (2020). Because our Bayesian methods require computationally intensive Markov chain Monte Carlo estimation, we used only 300 synthetic data sets per condition (10% of the original replications).[10] Nevertheless, our simulations still required ~25 CPU years to complete. A detailed description of the simulation environments (consisting of a total of 1620

[10] For the methods used in Hong and Reed (2020), we recalculated the result based on a sample matching the 300 replications per conditions used for the RoBMA methods, using the per-replication estimates shared by the authors.

conditions) and the remaining methods can be found in Hong and Reed (2020) and the corresponding original simulation publications. We compared the performance of RoBMA-PSMA to all methods used in the previous section. See Supplementary Materials at https://osf.io/bd9xp/ for comparison of methods after removing 5% of the most extreme estimates from each method, as done by Hong and Reed (2020), with the main difference being an improved performance of AK1 and AK2.

Tables 3 and 4 summarize the aggregated results for mean square error (MSE) and bias, respectively, separately for each simulation environment. Although no single estimator dominates across all simulation environments and criteria, RoBMA-PSMA is at or near the top in most cases. The exception is that RoBMA-PSMA produces below-average performance in the CSG&H environment. Tables F1 and F2 in Appendix F show that RoBMA-PSMA overcorrects the effect size estimates and performs relatively poorly only in conditions with p-hacking strong enough to introduce significant skew in the distributions of effect sizes.[11]

Following Hong and Reed (2020), Table 5 averages performance across all four simulation environments. While the results confirm that RoBMA-PSMA performs the best with regard to type I error rates and coverage, it is important to note that both the coverage and error rate were far above the nominal levels. The results also appear

[11] The QRPs simulated by Carter et al. (2019) results in a strongly positively skewed distribution of effect sizes. While RoBMA-PSMA contains selection models with weight functions that well adjusts for publication bias simulated by Carter et al. (2019), the additional skew generated by these QRPs results in misspecification of the best fitting models and consequent overcorrection of the meta-analytic effect size. The simpler original RoBMA does not contain the appropriate one-sided weight function. The skewed distribution of effect sizes does not introduce a strong systematic bias and it, as other methods, ironically result in better performance in the QRP environment of Carter et al. (2019) due to more layers of specific model misspecification. The remaining simulation environments produce bias in a more diverse manner and do not lead to such strongly skewed distributions of effect sizes. As a result, RoBMA-PSMA does not generally suffer from systematic bias.

Table 3

Ordered Performance of the Methods According to MSE for Each Simulation Environment. Rank 1 has the lowest MSE. See text for details.

Rank	SD&I	MSE	A&R	MSE	B&R	MSE	CSG&H	MSE
1	RoBMA-PSMA	0.009	RoBMA-PSMA	0.222	RoBMA-PSMA	0.098	RoBMA-old	0.012
2	*AK2**	*0.013*	TF	0.273	*p-uniform*	*0.185*	WAAP-WLS	0.018
3	RoBMA-old	0.017	*AK2**	*0.277*	WAAP-WLS	0.193	TF	0.022
4	TF	0.025	RoBMA-old	0.327	RoBMA-old	0.221	*3PSM*	*0.023*
5	WAAP-WLS	0.025	4PSM	0.365	TF	0.321	PET-PEESE	0.027
6	PET-PEESE	0.028	AK1*	0.389	EK	0.375	*p*-uniform	0.028
7	EK	0.031	Random Effects (DL)	0.511	PET-PEESE	0.378	*4PSM*	*0.031*
8	Random Effects (DL)	0.034	3PSM	0.511	4PSM	0.492	EK	0.033
9	*p*-uniform	0.050	WAAP-WLS	0.546	3PSM	0.493	RoBMA-PSMA	0.036
10	3PSM	0.238	PET-PEESE	0.605	Random Effects (DL)	0.526	Random Effects (DL)	0.046
11	*p*-curve	1.228	EK	0.760	*p*-curve	0.850	*p*-curve	0.075
12	4PSM	3.375	*p*-curve	3.514	AK1*	2.806	AK1*	0.280
13	AK1*	6.231	*p*-uniform	3.621	*AK2**	*5.816*	*AK2**	*2.849*

Note. Methods in *gray italic* converged in fewer than 90% repetitions in a given simulation environment.

* The performance difference in terms of MSE for AK1 and AK2 between our implementation and that of Hong and Reed (2020) is due to the fact that we did not omit the 5% most extreme estimates.

Table 4

Ordered Performance of the Methods According to Bias for Each Simulation Environment. Rank 1 has the lowest bias. See text for details.

Rank	SD&I	Bias	A&R	Bias	B&R	Bias	CSG&H	Bias
1	*AK2*	*0.029*	RoBMA-PSMA	0.159	EK	0.095	PET-PEESE	0.059
2	RoBMA-PSMA	0.034	*AK2*	*0.207*	*AK2*	*0.105*	WAAP-WLS	0.062
3	3PSM	0.040	EK	0.221	4PSM	0.108	RoBMA-old	0.064
4	PET-PEESE	0.049	PET-PEESE	0.259	RoBMA-PSMA	0.121	AK1	0.067
5	EK	0.053	WAAP-WLS	0.266	PET-PEESE	0.129	EK	0.072
6	RoBMA-old	0.062	TF	0.288	3PSM	0.156	*3PSM*	*0.081*
7	AK1	0.082	4PSM	0.302	WAAP-WLS	0.189	TF	0.091
8	WAAP-WLS	0.083	RoBMA-old	0.354	RoBMA-old	0.228	*4PSM*	*0.096*
9	TF	0.088	AK1	0.397	TF	0.240	*p*-uniform	0.106
10	4PSM	0.088	3PSM	0.475	AK1	0.277	RoBMA-PSMA	0.110
11	Random Effects (DL)	0.108	Random Effects (DL)	0.556	Random Effects (DL)	0.363	*AK2*	*0.117*
12	*p*-uniform	0.147	*p*-curve	1.530	*p*-uniform	0.374	*p*-curve	0.118
13	*p*-curve	0.422	*p*-uniform	1.555	*p*-curve	0.522	Random Effects (DL)	0.150

Note. Methods in *gray italic* converged in fewer than 90% repetitions in a given simulation environment.

Table 5

Aggregated Results Over all Simulation Conditions from Hong and Reed (2020).

Rank	Rank(Bias)	Bias	Rank(MSE)	MSE	Rank(|Coverage - .95|)	|Coverage - .95|	Rank(ERR)	ERR
1	EK	0.079	RoBMA-PSMA	0.054	*AK2*	*0.167*	RoBMA-PSMA	0.093
2	PET-PEESE	0.083	RoBMA-old	0.085	RoBMA-PSMA	0.172	*AK2*	*0.129*
3	*AK2*	*0.099*	WAAP-WLS	0.085	3PSM	0.213	EK	0.257
4	RoBMA-PSMA	0.099	TF	0.121	4PSM	0.265	3PSM	0.259
5	4PSM	0.103	PET-PEESE	0.149	PET-PEESE	0.306	PET-PEESE	0.286
6	3PSM	0.105	EK	0.155	EK	0.307	4PSM	0.290
7	WAAP-WLS	0.110	*p*-uniform	0.161	RoBMA-old	0.317	RoBMA-old	0.485
8	RoBMA-old	0.121	Random Effects (DL)	0.203	WAAP-WLS	0.319	WAAP-WLS	0.525
9	TF	0.141	3PSM	0.223	AK1	0.341	AK1	0.573
10	AK1	0.143	*p*-curve	0.623	TF	0.407	*p*-uniform	0.585
11	Random Effects (DL)	0.217	4PSM	0.851	Random Effects (DL)	0.510	TF	0.597
12	*p*-uniform	0.230	AK1	2.258	*p*-uniform	0.576	Random Effects (DL)	0.649
13	*p*-curve	0.336	*AK2*	*3.316*	*p*-curve		*p*-curve	

Note. Ranking and values of aggregated bias, mean square error (MSE), absolute difference from .95 CI coverage (|Coverage - .95|), and type I error rate (ERR) averaged across all simulation environments in (Hong and Reed, 2020; the type I error rate for RoBMA methods is based on BF $>$ 10).

* The performance difference in terms of MSE for AK1 and AK2 between our implementation and that of Hong and Reed (2020) is due to the fact that we did not omit the 5% most extreme estimates.

Methods in *gray italic* converged in fewer than 90% repetitions in a given simulation environment.

favorable to AK2, as it has the lowest bias in SD&I environment and the second lowest biases in the A&R and B&R environments. However, AK2 failed to converge in over 10% of these simulated meta-analyses. Even when AK2 converges, its MSE in the B&R and CSG&H environments is relatively large.

Table 6

Ordered Performance of the Methods Across Simulation Environments According to Log Scoring Rule

Rank	Bias	LogScore(Bias)	MSE	LogScore(MSE)
1	*AK2*	*0.801*	RoBMA-PSMA	0.831
2	RoBMA-PSMA	0.801	RoBMA-old	0.682
3	EK	0.778	TF	0.519
4	PET-PEESE	0.746	WAAP-WLS	0.504
5	WAAP-WLS	0.616	*AK2*	*0.392*
6	3PSM	0.615	PET-PEESE	0.369
7	4PSM	0.602	EK	0.327
8	RoBMA-old	0.579	*p*-uniform	0.324
9	AK1	0.515	3PSM	0.316
10	TF	0.500	4PSM	0.315
11	Random Effects (DL)	0.329	Random Effects (DL)	0.310
12	*p*-uniform	0.304	AK1	0.183
13	*p*-curve	0.242	*p*-curve	0.114

Note. Methods in *gray italic* converged in less than 90% repetitions.

It should be noted that the averaging operation is valid only for coverage and type I error rates, as these are fully comparable across the different simulation environments. In contrast, bias and MSE cannot be directly averaged or aggregated, as these are based on very different effect-size metrics (Hong & Reed, 2020); for instance, the best method in

A&R environment has five times the bias as the best method in SD&I environment. In order to make the metrics commensurate, we employ a relative order-preserving logarithmic transformation to obtain an average ranking across these four different simulation environments (1 corresponds to the best relative performance, 0 to the worst relative performance; Ioannidis et al., 2019).[12] Table 6 displays the average relative ranks of bias and MSE for these alternative methods across all simulation environments. RoBMA-PSMA is ranked highest according to MSE and type I error rates, and is the second best according to both bias and confidence interval coverage. Again, the closest competition appears to come from AK2, but AK2 often does not converge and may yield high MSEs – see Table 3. Table 5 also shows that RoBMA-PSMA does well when bias and MSE are simply averaged across these simulations, but those comparisons need to be interpreted with caution.

Concluding Comments

We have extended the robust Bayesian meta-analytic framework with one-sided weight functions and PET-PEESE regression models. This extension allows researchers to draw inferences using a multitude of otherwise competing approaches (i.e., selection models based on p-values and models estimating the relationship between effect sizes and standard errors). Consequently, researchers interested in obtaining the best possible adjusted meta-analytic effect size estimate do not need to speculate about the type of publication bias in order to select the best method for their setting. Instead, RoBMA weights its inference in proportion to how well each method accounts for the data.

The extended version of RoBMA resolves the tension between the selection models and PET-PEESE. Furthermore, we demonstrated that RoBMA-PSMA outperforms previous methods when applied to actual meta-analyses for which a gold standard is available (Kvarven et al., 2020). Finally, the new RoBMA methods performed well in

[12] This transformation has been used to compare and rank top scientists across different criteria of scientific impact (Ioannidis et al., 2019).

simulation studies. However, it is important to note that RoBMA-PSMA did not perform well in simulation settings of Carter et al. (2019) with prominent p-hacking where it overcorrected the effect sizes.

The RoBMA framework can be further extended in multiple ways: to account for multilevel structures, to estimate within study clusters, to deal with multivariate outcomes, and to include of explanatory variables. Many of those extensions will, however, increase computational complexity, making them practically unfeasible for selection models. Therefore, further research is need in developing efficient algorithms or approximations that will allow the further extensions, currently unachievable under the RoBMA-PSMA framework.

Out of the remaining methods, p-curve, p-uniform, and random effects meta-analysis were dominated by the other estimators, and AK2 failed to converge in many cases. Overall, Bayesian model-averaging greatly improved both PET-PEESE and selection models: RoBMA-PSMA reduces PET-PEESE's MSE and bias as well as the selection models' MSE. Importantly, RoBMA-PSMA takes uncertainty about the type of publication bias into account and combines the best of the two worlds. Even though RoBMA outperforms other methods in many cases in both the simulation study and the comparison of meta-analyses and registered replication reports, it should be considered merely a new tool in the toolbox of publication selection bias detection.

In cases where the data generating process is known and depending on the metric that researchers want to optimize (e.g., bias vs. RMSE) an appropriate method can be selected via the results from our simulation study or the meta-showdown explorer https://tellmi.psy.lmu.de/felix/metaExplorer/. If there is considerable uncertainty about the data generating process, we believe that RoBMA is a sensible default. Nevertheless, researchers may wish to check the conclusions of RoBMA against methods that are not part of the RoBMA ensemble, such as WAAP-WLS. As there is no principled way of

averaging these methods with RoBMA,[13] researchers should view these comparisons as sensitivity analyses. If alternative methods come to the same conclusions as RoBMA, this suggests that the results are robust; If alternative methods come to a qualitatively different conclusion, this suggests that the results are fragile; in this case we recommend a more in-depth consideration of the data-model relationship, and a transparent report that the conclusions vary based on the selected meta-analytic technique.

We believe that the extended version of RoBMA with the outlined default prior distributions presents a reasonable setup for anyone interested in performing a meta-analysis. However, the RoBMA framework is flexible and allows researchers to specify different prior distributions for any of the model parameters or include/exclude additional models (see "Appendix B: Specifying Different Priors" in Bartoš et al., in press, or many of the `R` package vignettes). Consequently, researchers with substantial prior knowledge can test more specific hypotheses than those specified with the default model ensemble (e.g., Bartoš et al., 2021; Gronau et al., 2020; Gronau et al., 2017) or incorporate prior knowledge about the research environment. For instance, when prior research has established that the effect of interest shows considerable between-study heterogeneity, researchers may decide to trim the default RoBMA ensemble by assigning prior probability zero to the fixed effects models, and consequently drawing conclusions from only the random effects models.

We have implemented RoBMA-PSMA in a new version of the `RoBMA R` package (Bartoš & Maier, 2021). Also, for researchers with little programming expertise we will implement the methodology in the open-source statistical software package JASP (**ly2021bayesian**; JASP Team, 2021). We hope that these publicly-shared statistical packages will encourage researchers across different disciplines to adopt these new methods for accommodating potential publication bias and draw conclusions that are rich, robust, and reliable.

[13] Many of these methods either remove data (WAAP-WLS) or impute data (trim-and-fill), which makes a comparison via marginal likelihood impossible.

Data Availability Statement

The data and `R` scripts for performing the analyses and simulation study are openly available on OSF at https://osf.io/fgqpc/.

Author Contributions

All authors jointly generated the idea for the study. F. Bartoš programmed the analysis, conducted the simulation study, and analyzed the data. F. Bartoš and M. Maier wrote the first draft of the manuscript and all authors critically edited it. All authors approved the final submitted version of the manuscript.

Conflicts of Interest

F. Bartoš declares that he owns a negligible amount of shares in semiconductor manufacturing companies that might benefit from a wider application of computationally intensive methods such as RoBMA-PSMA. The authors declare that there were no other conflicts of interest with respect to the authorship or the publication of this article.

Highlights

- There are multiple publication bias adjustment methods that often lead to conflicting conclusions.
- Their performance depends on factors that are unknown to the analysts, making it difficult to choose the right estimator for a given situation.
- We combine two main approaches to publication bias adjustment (selection models and PET-PEESE) with Bayesian model-averaging. The resulting model-averaged ensemble bases the inference according to the prior predictive performance of included models.

- We illustrate the performance of the proposed method across multiple examples and simulation studies. We provide an R package implementing the method.
- We demonstrate that Bayesian model-averaging is a viable approach to publication bias adjustment that provides a rich, robust, and reliable inference.

References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science*, *1*(3), 357–366.

Alcock, J. (2011). Back from the future: Parapsychology and the Bem affair. *Skeptical Inquirer*, *35*(2), 31–39. https://skepticalinquirer.org/exclusive/back-from-the-future/

Alinaghi, N., & Reed, W. R. (2018). Meta-analysis and publication bias: How well does the FAT-PET-PEESE procedure work? *Research Synthesis Methods*, *9*(2), 285–311. https://doi.org/10.1002/jrsm.1298

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*(8), 2766–94. https://doi.org/10.1257/aer.20180310

Bartoš, F., Gronau, Q. F., Timmers, B., Otte, W. M., Ly, A., & Wagenmakers, E.-J. (2021). Bayesian model-averaged meta-analysis in medicine. *Statistics in Medicine*. https://onlinelibrary.wiley.com/doi/10.1002/sim.9170

Bartoš, F., & Maier, M. (2021). *RoBMA: An R Package for Robust Bayesian Meta-Analyses* [R package version 2.0.0]. https://CRAN.R-project.org/package=RoBMA

Bartoš, F., Maier, M., Quintana, D., & Wagenmakers, E.-J. (in press). Adjusting for publication bias in JASP and R — Selection models, PET-PEESE, and robust Bayesian meta-analysis. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.31234/osf.io/75bqn

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, *101*(4), 716–719. https://doi.org/10.1037/a0024777

Bom, P. R., & Rachinger, H. (2019). A kinked meta-regression model for publication bias correction. *Research Synthesis Methods*, *10*(4), 497–514. https://doi.org/10.1002/jrsm.1352

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). Publication bias. In M. Borenstein (Ed.), *Introduction to Meta-Analysis* (pp. 277–292). Wiley.

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*, 823. https://doi.org/10.3389/fpsyg.2014.00823

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Citkowicz, M., & Vevea, J. L. (2017). A parsimonious weight function for modeling publication bias. *Psychological Methods*, *22*(1), 28–41. https://doi.org/10.1037/met0000119

Clyde, M. A., Ghosh, J., & Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, *20*(1), 80–101. https://doi.org/10.1198/jcgs.2010.09049

Copas, J. (1999). What works?: Selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(1), 95–109. https://doi.org/10.1111/1467-985X.00123

Copas, J., & Li, H. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(1), 55–95. https://doi.org/10.1111/1467-9868.00055

Copas, J., & Shi, J. Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research*, *10*(4), 251–265. https://doi.org/10.1177/096228020101000402

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. https://doi.org/10.1111/j.0006-341X.2000.00455.x

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. https://doi.org/10.1136/bmj.315.7109.629

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*, 313–329. https://doi.org/10.1214/16-STS599

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Fragoso, T. M., Bertoli, W., & Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, *86*(1), 1–28. http://dx.doi.org/10.1111/insr.12243

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*(2), 151–156. https://doi.org/10.3758/s13423-012-0227-9

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, *103*(6), 933–948. https://doi.org/10.1037/a0029709

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2021). A primer on Bayesian model-averaged meta-analysis. *Advances in Methods and Practices in Psychological Science*, *4*(3). https://doi.org/10.1177%5C%2F25152459211031256

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian t-tests. *The American Statistician*, *74*(2), 137–143. https://doi.org/10.1080/00031305.2018.1562983

Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*(1), 123–138. https://doi.org/10.1080/23743603.2017.1326760

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review*, *23*(1), 74–86. https://doi.org/10.3758/s13423-015-0868-6

Haaf, J. M., & Rouder, J. N. (2020). Does every study? Implementing ordinal constraint in meta-analysis. https://doi.org/10.31234/osf.io/hf9se

Heck, W., D., Gronau, F., Q., Wagenmakers, & E.-J. (2019). *MetaBMA: Bayesian model averaging for random and fixed effects meta-analysis.* https://CRAN.R-project.org/package=metaBMA

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, *7*(2), 246–255.

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215. https://doi.org/10.1177/2515245919898657

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–401.

Hong, S., & Reed, W. R. (2020). Using Monte Carlo experiments to select meta-analytic estimators. *Research Synthesis Methods*, *12*, 192–215. https://doi.org/10.1002/jrsm.1467

Hoogeveen, S., Sarafoglou, A., & Wagenmakers, E.-J. (2020). Laypeople can predict which social-science studies will be replicated successfully. *Advances in Methods and Practices in Psychological Science*, *3*(3), 267–285. https://doi.org/10.1177/2515245920919667

Ioannidis, J. P., Baas, J., Klavans, R., & Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLoS Biology*, *17*(8). https://doi.org/10.1371/journal.pbio.3000384

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, *3*(1), 109–117.

JASP Team. (2021). JASP (Version 0.15)[Computer software]. https://jasp-stats.org/

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*, 64–72.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, *31*, 203–222. https://doi.org/10.1017/S030500410001330X

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*, 788–799. https://doi.org/10.1038/s41593-020-0660-4

Kivikangas, J. M., Fernández-Castilla, B., Järvelä, S., Ravaja, N., & Lönnqvist, J.-E. (2021). Moral foundations and political orientation: Systematic review and meta-analysis. *Psychological Bulletin*, *147*(1), 55–94. https://doi.org/10.1037/bul0000308

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/515245918810225

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., et al. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*, 451–479. https://doi.org/10.1037/bul0000308

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data* (Vol. 53). Wiley New York.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Maier, M., Bartoš, F., Oh, M., Wagenmakers, E.-J., Shanks, D., & Harris, A. (2022). *Publication bias in research on construal level theory*. https://doi.org/10.31234/osf.io/r8nyu

Maier, M., Bartoš, F., Stanley, T., Shanks, D. R., Harris J.L., A., & Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, *119*(31). https://doi.org/10.1073/pnas.2200300119

Maier, M., Bartoš, F., & Wagenmakers, E.-J. (2022). Robust Bayesian meta-analysis: Addressing publication bias with model-averaging. *Psychological Methods*. https://doi.org/10.1037/met0000405

Maier, M., VanderWeele, T. J., & Mathur, M. B. (2022). Using selection models to assess sensitivity to publication bias: A tutorial and call for more routine use. *Campbell Systematic Reviews*, *18*(3), e1256. https://doi.org/10.1002/cl2.1256

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. https://doi.org/10.1080/17470218.2012.711335

Mathur, M. B., & VanderWeele, T. J. (2019). Finding common ground in meta-analysis "wars" on violent video games. *Perspectives on Psychological Science*, *14*(4), 705–708.

Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(5), 1091–1119.

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. https://doi.org/10.1177/1745691616662243

Moreno, S. G., Sutton, A. J., Turner, E. H., Abrams, K. R., Cooper, N. J., Palmer, T. M., & Ades, A. (2009). Novel methods to deal with publication biases: Secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *BMJ*, *339*, b2981. https://doi.org/10.1136/bmj.b2981

Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 79–96). Psychology Press.

Renkewitz, F., & Keiner, M. (2019). How to detect publication bias in psychological research. *Zeitschrift für Psychologie*, *227*(4), 261–279. https://doi.org/10.1027/2151-2604/a000386

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'Retroactive facilitation of recall'effect. *PloS One*, *7*(3), e33423. https://doi.org/10.1371/journal.pone.0033423

Robinson, G. K. (2019). What properties might statistical inferences reasonably be expected to have? – Crisis and resolution in statistical inference. *The American Statistician*, *73*, 243–252. https://doi.org/10.1080/00031305.2017.1415971

Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in interpretation of levels of significance. *Psychological Reports*, *15*(2), 570. https://doi.org/10.2466/pr0.1964.15.2.570

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005a). *Publication bias in meta-analysis*. John Wiley & Sons.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005b). Publication bias in meta-analysis. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, 1–7. https://doi.org/10.1002/0470870168.ch1

Rouder, J. N., & Morey, R. D. (2019a). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, *73*, 186–190. https://doi.org/10.1080/00031305.2017.1341334

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*(4), 682–689. https://doi.org/10.3758/s13423-011-0088-7

Rouder, J. N., & Morey, R. D. (2019b). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, *73*(2), 186–190. https://doi.org/10.1080/00031305.2017.1341334

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2). https://doi.org/10.1177/25152459211007467

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551. https://doi.org/10.1037/a0029487

Schimmack, U. (2015, May 9). *Why psychologists should not change the way they analyze their data: The devil is in the default prior* [Replicability-index]. Retrieved July 6, 2021, from https://replicationindex.com/2015/05/09/why-psychologists-should-not-change-the-way-they-analyze-their-data-the-devil-is-in-the-default-piror/

Schimmack, U. (2018, January 21). *My email correspondence with Daryl J. Bem about the data for his 2011 article "Feeling the future"* [Replicability-index]. Retrieved October 21, 2019, from https://replicationindex.com/2018/01/20/my-email-correspondence-with-daryl-j-bem-about-the-data-for-his-2011-article-feeling-the-future/

Schlitz, M., Bem, D. J., Marcusson-Clavertz, D., CardeñA, E., Lyke, J., Grover, R., Blackmore, S., Tressoldi, P., Roney-Dougal, S., Bierman, D., et al. (2021). Two replication studies of a time-reversed (psi) priming task and the role of expectancy in reaction times. *Journal of Scientific Exploration*, *35*(1), 65–90.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2). https://doi.org/10.1037/a0033242

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, *8*(5), 581–591. https://doi.org/10.1177/1948550617693062

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

Stanley, T. D., & Doucouliagos, H. (2017). Neither fixed nor random: Weighted least squares meta-regression. *Research Synthesis Methods*, *8*(1), 19–42. https://doi.org/10.1002/jrsm.1211

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. (2017). Finding the power to reduce publication bias. *Statistics in Medicine*, *36*(10), 1580–1598. https://doi.org/10.1002/sim.7228

Stefan, A., & Schönbrodt, F. D. (2022). Big little lies: A compendium and simulation of p-hacking strategies. https://doi.org/10.31234/osf.io/xy2dk

ter Schure, J., & Grünwald, P. (2019). Accumulation Bias in meta-analysis: The need to consider time in error control. *F1000Research*, *8*, Article 962. https://dx.doi.org/10.12688/f1000research.19375.1

Van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293–309. https://doi.org/10.1037/met0000025

van Erp, S., Verhagen, J., Grasman, R. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990–2013. *Journal of Open Psychology Data*, *5*(1), Article 4. http://doi.org/10.5334/jopd.33

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. https://doi.org/10.1007/BF02294384

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428. https://psycnet.apa.org/doi/10.1037/1082-989X.10.4.428

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176. https://doi.org/10.1177/0963721416643289

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–32. http://dx.doi.org/10.2139/ssrn.2001721

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wicherts, J. M. (2017). The weak spots in contemporary science (and how to fix them). *Animals*, *7*(12), 90–119. https://doi.org/10.3390/ani7120090

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390. https://doi.org/10.1080/14786442108633773

Appendix A

Model Specifications

Most of the model specification is identical to the one described in Appendix A of Maier, Bartoš, and Wagenmakers (2022). The individual models employed in the ensemble (Table 1) differ in terms of the prior distribution specified over the effect size parameter (μ), the heterogeneity parameter (τ), and the way they adjust for publication bias (ω). If we group the models according to the way they adjust for publication bias, we can differentiate between the following model types based on the likelihood function.

Models Assuming No Publication Bias

Models assuming no publication bias use a normal likelihood to model the observed effect sizes y based on the observed standard errors se from K studies,

$$y_k \sim \text{Normal}(\mu, \tau^2 + se_k^2). \tag{12}$$

If the specific model assumes absence of an effect or heterogeneity, it further simplifies by setting $\mu = 0$ and $\tau = 0$. Otherwise, the corresponding prior distributions for μ and τ needs to be specified to obtain the complete model.

Models Adjusting for Publication Bias Based on The Relationship Between Standard Errors and Effect Sizes

Models correcting for publication bias by adjusting for the relationship between standard errors/variances and effect sizes use a normal likelihood as the models assuming no publication bias; however, they add a regression parameter that adjusts for the relationship between effect sizes and standard errors (PET) or the effect sizes and variances (PEESE),

$$\begin{aligned} y_k &\sim \text{Normal}(\mu + \text{PET} \times se_k, \tau^2 + se_k^2), \\ y_k &\sim \text{Normal}(\mu + \text{PEESE} \times se_k^2, \tau^2 + se_k^2). \end{aligned} \tag{13}$$

As before, in the case that the specific model assumes absence of the effect, or heterogeneity, it further simplifies by setting $\mu = 0$ and $\tau = 0$. Otherwise, the corresponding prior distributions for μ, τ, and PET or PEESE needs to be specified to obtain the complete model.

Selection Models

Selection models use a weighted likelihood function to incorporate the publication probabilities, ω, into the likelihood function for the observed effect sizes,

$$y_k \sim \text{Weighted-normal}(\mu, \tau^2 + se_k^2, \omega). \tag{14}$$

Weighted-normal stands for a likelihood function of a weighted normal distribution, with mean μ, variance σ^2, weights ω, and a cumulative probability function of a standard normal distribution Φ, that is further differentiated accordingly whether the one-sided or two-sided selection is assumed,

$$\begin{aligned}
\text{Weighted-normal}_{\text{one-sided}}(y \mid \mu, \sigma^2, \omega) &= \frac{\text{Normal}(y \mid \mu, \sigma^2) \times w(\omega, p, c)}{\int \text{Normal}(x \mid \mu, \sigma^2) \times w(\omega, 1 - \Phi(x/\sigma), c)dx}, \\
\text{Weighted-normal}_{\text{two-sided}}(y \mid \mu, \sigma^2, \omega) &= \frac{\text{Normal}(y \mid \mu, \sigma^2) \times w(\omega, p, c)}{\int \text{Normal}(x \mid \mu, \sigma^2) \times w(\omega, (1 - \Phi(|x/\sigma|)) \times 2, c)dx},
\end{aligned} \tag{15}$$

where the weights ω are assigned based on the one or two-sided p-values, p, and N cutoffs c through the weight function w,

$$w(\omega, p, c) = \begin{cases} \omega_1, & \text{if } p > c_1 \\ \omega_n, & \text{if } c_n < p \leq c_{n+1} \\ ... & \\ 1, & \text{if } p \leq c_N \end{cases} \tag{16}$$

Again, in the case that the specific model assumes absence of the effect, or heterogeneity, it further simplifies by setting $\mu = 0$ and $\tau = 0$ respectively. Otherwise, the corresponding prior distributions for μ, τ, and ω needs to be specified to obtain the complete model.

Appendix B

Parameter Prior Distributions

Table 1 outlines the default prior distributions used throughout the manuscript. The specified default prior distributions can be viewed as a sensible starting options tested in simulations. However, one of the advantages of Bayesian statistics is that it allows researchers to flexibly specify and test different hypotheses. We urge researchers to specify their own prior distributions directly corresponding to the hypotheses of interest. See Bartoš et al. (in press) for a tutorial where we explain how to specify different alternative and/or null hypotheses with RoBMA. Here, we outline the rationale for the default prior distributions used in this manuscript.

Effect Size (μ)

For the effect size μ, we use a standard normal, Normal$(0, 1)$, as the default prior distribution. We already used this distribution when introducing the previous version of the method (Maier, Bartoš, & Wagenmakers, 2022). The standard normal prior distribution specifies a wide a range of plausible values for effect sizes, yet it has thinner tail than a frequently used Cauchy$(0, 1)$ prior distribution. The thinner tails of the standard normal distribution reduce the prior probability of very large effect sizes that we deem as less plausible in meta-analytic settings. Another choice for prior distribution for μ, used in the robustness analysis of Kvarven et al. (2020, Appendix C), might be Student-t$_{[0,\infty]}(0.35, 0.102, 3)$, so-called "Oosterwijk prior" (Gronau et al., 2020). The Oosterwijk prior is a shifted and scaled student-t distribution with location 0.35, scale 0.102, and three degrees of freedom, truncated to have mass only on positive effect sizes. We consider Oosterwijk prior to be a reasonable specification for effects that are known to be of small-to-medium size and it has been used in previous studies (e.g., Gronau et al., 2017; Landy et al., 2020).

Heterogeneity (τ)

For the heterogeneity τ, we use an inverse-gamma, InvGamma$(1, 0.15)$, as the default prior distribution. We also used this distribution in Maier, Bartoš, and Wagenmakers (2022) and it is based on heterogeneity estimates from meta-analyses in psychology recorded by van Erp et al. (2017). This prior distribution was used in previous studies (e.g., Gronau et al., 2017; Landy et al., 2020) and it is also the default choice of the `metaBMA` R package (Heck et al., 2019).

Publication Bias Regression Coefficients For PET and PEESE

For the PET and PEESE regression coefficients, we use Cauchy, Cauchy$_{[0,\infty]}(0, 1)$ and Cauchy$_{[\infty]}(0, 5)$, as the default prior distributions. Equation 13 shows that the PET and PEESE regression coefficients can be thought as a bias of studies with a given standard error or variance. Since standard errors (and subsequently variances) are dependent on the sample size for standardized effect size measures, we derived the range of plausible values for the prior distribution based on the regression coefficients based on small sample studies (N $= 25, 50, 100$) and values of medium to large bias (bias $= 0.30, 0.40, 0.50$). The resulting values of PET regression coefficients are summarized in Table B1 and PEESE regression coefficients in Table B2. We conclude that the Cauchy$_{[0,\infty]}(0, 1)$ and Cauchy$_{[0,\infty]}(0, 5)$ prior distribution cover the range reasonably well and still allow for larger values in case that our initial assessment was incorrect.

Table B1

PET Regression Coefficients Based on Theoretical Sample Sizes and Degrees of Bias

Bias	0.30	0.40	0.50
N = 25	0.75	1.00	1.25
N = 50	1.06	1.41	1.77
N = 100	1.50	2.00	2.50

Table B2

PEESE Regression Coefficients Based on Theoretical Sample Sizes and Degrees of Bias

Bias	0.30	0.40	0.50
N = 25	1.88	2.50	3.13
N = 50	3.75	5.00	6.25
N = 100	7.50	10.00	12.50

To assess the robustness of our results in the Kvarven et al. (2020) example, we collected the estimated PET and PEESE regression coefficients from conditions assuming presence of the publication bias in the simulation study. We fitted gamma distributions to the simulation-based PET and PEESE regression coefficients using maximum likelihood and obtained Gamma$(2.84, 2.19)$ and Gamma$(2.32, 0.86)$ shape and rate parameterized prior distributions for PET and PEESE regression coefficients. Both of the prior distributions show a more concentrated prior probability density around 1.30 and 2.70 with a much thinner tail, making them more informed.

Notably, when transforming prior distributions for PET and PEESE regression coefficients to a different effect size scale, the prior distribution for the PET regression coefficient does not change – the scaling of the effect size corresponds to scaling of the standard error when using approximate linear transformation, however, the PEESE regression coefficient changes with the inverse of the approximate linear transformation applied to the effect size and standard errors.

Publication Bias Weights (ω)

For the publication bias weights ω, we use unit cumulative Dirichlet distributions, as the default prior distributions. In the case of a weight function with only one step, the unit cumulative Dirichlet distribution simplifies to a uniform distribution on interval from zero to one. In the more complex cases, the unit cumulative Dirichlet distributions assigns

prior probabilities across the possible weights, constraining them to be increasing, bound between zero and one, and allowing for variation in the predicted values.

Similarly to the prior distribution for the PET and PEESE regression coefficient, we assess the robustness of our results in the Kvarven et al. (2020) example by fitting cumulative Dirichlet distributions to the estimated publication bias weights based on the simulation study using maximum likelihood. Table B3 summarizes the simulation-based prior distribution and shows that the first parameter is usually larger resulting in a smaller step from significant to the non-significant studies, in other words, a more optimistic prediction regarding publication bias.

Table B3

Simulation-Based Prior Distribution for Publication Bias Weight Functions

Weight function	Prior distribution
$\omega_{\text{Two-sided(.05)}}$	CumDirichlet$(2.49, 0.83)$
$\omega_{\text{Two-sided(.1,.05)}}$	CumDirichlet$(2.88, 0.98, 0.99)$
$\omega_{\text{One-sided(.05)}}$	CumDirichlet$(2.61, 0.89)$
$\omega_{\text{One-sided(.05,.025)}}$	CumDirichlet$(2.92, 0.95, 0.75)$
$\omega_{\text{One-sided(.5,.05)}}$	CumDirichlet$(3.17, 0.80, 0.83)$
$\omega_{\text{One-sided(.5,.05,.025)}}$	CumDirichlet$(3.24, 1.02, 0.68, 0.66)$

Appendix C

Robustness of the Kvarven et al. (2020) Results Across Different Prior Specifications

To assess the robustness of the results on the empirical data sets provided by Kvarven et al. (2020), we repeated the analysis conducted in the "Evaluating RoBMA on Registered Replication Reports" section with different parameter prior distributions. First, we exchanged the default standard normal prior for the effect size μ with the Oosterwijk prior distribution, Student-$t_{[0,\infty]}(0.35, 0.102, 3)$. Second, we exchanged the default prior distributions for the PET and PEESE regression coefficients and the publication bias weights ω for the simulation-based prior distributions. Finally, we exchanged the default prior distributions for both parameters simultaneously (a detailed description of the prior distributions can be found in Appendix B).

We found that all RoBMA models performed the best under the simulation-based prior distribution for the PET and PEESE regression coefficients and publication bias weights and the worst under the Oosterwijk prior distribution. We attribute the inferior performance of the Oosterwijk prior distribution to the fact that three replication studies resulted in negative estimates that are unattainable under the prior distribution restricted to positive numbers. However, even though there were some differences in how the RoBMA models performed under different prior distributions, the results were still in line with our previous conclusions.

Figure C1 compares the model-averaged posterior effect size estimate from RoBMA-PSMA with the default prior distribution specification (blue) to model-averaged posterior effect size estimates from RoBMA-PSMA with the above three alternative prior distributions for each of the 15 RRRs from Kvarven et al. (2020). In general, the figure shows consistent results across prior distributions; however, a closer look reveals that the informed Oosterwijk prior distribution pulls the model-averaged posterior for effect size towards its prior location (i.e., $\mu = 0.35$). Also note that under the Oosterwijk prior, the lower bounds of the credible interval do not cross zero – it is impossible to obtain a

negative estimate when this has been ruled out a priori by restricting the prior range to the positive real line.

Table C1

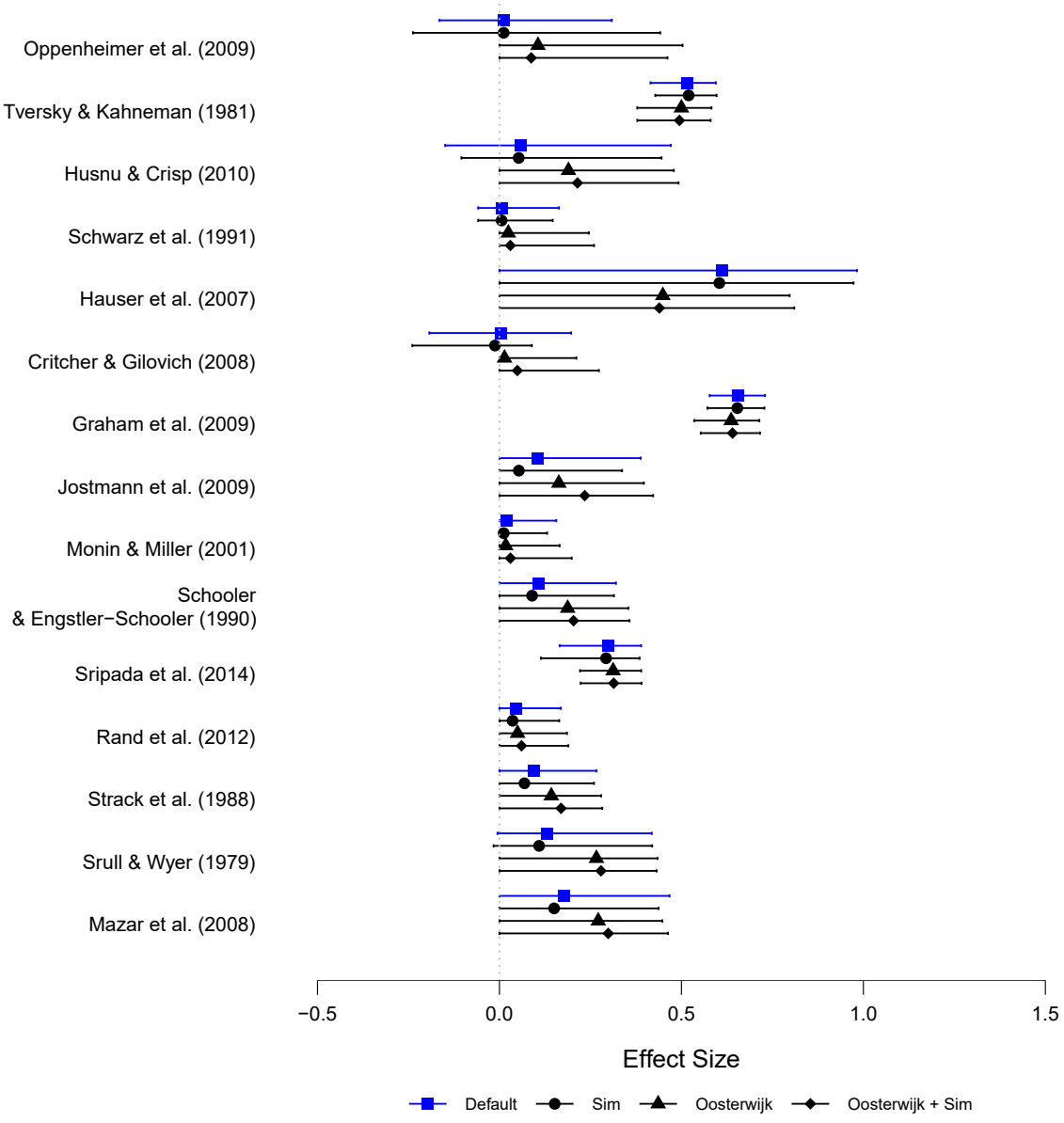
Performance of RoBMA with Different Priors in the Kvarven et al. (2020) example.

Method	FPR	FNR	Undecided	OF	Bias	RMSE
Oosterwijk prior (μ)						
RoBMA-PSMA	0.286	0.000	0.667	1.446	0.073	0.204
RoBMA-old	0.714	0.000	0.133	2.050	0.172	0.224
Simulation-based prior						
RoBMA-PSMA	0.143	0.000	0.800	1.080	0.013	0.160
RoBMA-old	0.714	0.000	0133.	2.021	0.167	0.212
Oosterwijk prior (μ) & simulation-based prior						
RoBMA-PSMA	0.143	0.000	0.667	1.358	0.059	0.192
RoBMA-old	0.714	0.000	0.133	2.032	0.169	0.219

Note. FPR = false positive rate, FNR = false negative rate, Undecided = undecided evidence, OF = overestimation factor, and RMSE = root mean square error.

Figure C1

Robustness Check of Effect Size Estimates with 95% CIs Comparing RoBMA-PSMA with the Default Prior Distribution to Three Alternative Prior distribution Specifications for the 15 Experiments Included in Kvarven et al. (2020).



Note. Estimates are reported on the Cohen's d scale. "Sim" corresponds to simulation based priors for the publication bias adjustment part and "Oosterwijk" corresponds to an informed prior distribution expecting small-to-medium effect sizes (Student-$t_{[0,\infty]}(0.35, 0.102, 3)$).

Appendix D

Robustness of the Kvarven et al. (2020) to the Selection of Registered Replication Reports

To further assess the robustness of the results on the empirical data sets provided by Kvarven et al. (2020), we repeated the analysis conducted in the "Evaluating RoBMA on Registered Replication Reports" section with a non-parametric bootstrap of the data set. We performed a 1000 repetitions, each of which sampled 15 RRR data sets with replacement which allowed us to assess the dependency of the reported results on the particular data.

Table D1 summarizes results from the non-parametric bootstrap as 95% quantile intervals. We found that the false positive and false negative rates estimates were highly variable since each of them was based only on a subset of the RRR (seven not statistically significant and eight statistically significant). Despite the added uncertainty depicted via the quantile intervals of overestimation, bias, and RMSE estimates, the results were aligned with our previous conclusions. Moreover, the higher quantile limit of bias and RMSE of the RoBMA-PSMA was around or bellow lower quantile limits of many other methods.

Table D1

Robustness of Performance of 13 Publication Bias Correction Methods to Selection of the Kvarven et al. (2020) Test Set Comprised of 15 Meta-analyses and 15 Corresponding "Gold Standard" Registered Replication Reports (RRR).

Method	FPR / Undecided	FNR / Undecided	Overestimation	Bias	RMSE
RoBMA-PSMA	[0.000, 0.444] / [0.556, 1.000]	[0.000, 0.000] / [0.429, 1.000]	[0.744, 2.261]	[-0.054, 0.114]	[0.106, 0.214]
AK2	*[0.000, 0.000] / —*	*[0.000, 1.000] / —*	*[-0.758, 3.551]*	*[-0.326, 0.098]*	*[0.040, 0.453]*
PET-PEESE	[0.000, 0.429] / —	[0.143, 0.857] / —	[0.447, 2.343]	[-0.067, 0.176]	[0.108, 0.361]
EK	[0.000, 0.429] / —	[0.143, 0.857] / —	[0.527, 2.541]	[-0.063, 0.202]	[0.124, 0.403]
RoBMA-old	[0.333, 1.000] / [0.000, 0.667]	[0.000, 0.000] / [0.000, 0.000]	[1.417, 4.345]	[0.107, 0.237]	[0.162, 0.264]
3PSM	[0.333, 1.000] / —	[0.000, 0.400] / —	[1.462, 4.980]	[0.117, 0.271]	[0.177, 0.309]
4PSM	[0.363, 1.000] / —	[0.143, 0.833] / —	[1.025, 4.543]	[0.005, 0.243]	[0.209, 0.324]
TF	*[0.500, 1.000] / —*	*[0.000, 0.000] / —*	*[1.541, 5.319]*	*[0.127, 0.289]*	*[0.185, 0.325]*
AK1	[0.544, 1.000] / —	[0.000, 0.000] / —	[1.571, 5.378]	[0.145, 0.291]	[0.202, 0.320]
p-curve		/ —	[1.606, 5.350]	[0.134, 0.319]	[0.208, 0.362]
p-uniform	[0.000, 1.000] / —	[0.000, 0.750] / —	[1.587, 5.349]	[0.136, 0.316]	[0.210, 0.365]
WAAP-WLS	[0.544, 1.000] / —	[0.000, 0.400] / —	[1.675, 5.296]	[0.157, 0.330]	[0.186, 0.392]
RE	[1.000, 1.000] / —	[0.000, 0.000] / —	[1.686, 6.032]	[0.174, 0.342]	[0.225, 0.385]

Note. FPR / Undecided = false positive rate / undecided evidence under no effect, FNR/ Undecided = false negative rate/ undecided evidence under an effect, OF = overestimation factor, and RMSE = root mean square error. The table presents 95% bootstrapped percentile intervals from 1000 samples with results conditional on convergence. The results in *gray italic* are conditional on convergence: trim and fill did not converge in one case and AK2 did not converge in 10 cases.

Appendix E

Performance Under the Absence of Publication Bias

In the "Evaluating RoBMA on Registered Replication Reports" section, all Registered Replication Reports (RRR) found lower effect size estimates than the original meta-analyses. To assess whether RoBMA's performance can be explained by a systematic underestimation of effect sizes, we estimated RoBMA and the remaining publication bias correction methods on data from Many Labs 2 (Klein et al., 2018). Many Labs 2 is a collection of different RRR attempting to replicate 28 classic and contemporary findings from psychology across multiple participating labs (N = 125). Since each finding was replicated by about half of the labs following the same RRR protocol, we can be certain about the absence of publication bias in the collection of the lab estimates. Consequently, we can establish the "Gold Standard" effect size estimate for each of the 28 psychological findings by applying a fixed-effect meta-analysis to the corresponding effect size estimates from the different labs (only one heterogeneity estimate τ was larger than 0.2). If RoBMA's previous performance was a result of systematic underestimation, we would expect to find significantly underestimated effect size estimates (the positive bias of the random-effect meta-analytic models in the "Evaluating RoBMA on Registered Replication Reports" section was 0.259).

The results are summarized in Table E1. We found that RoBMA-PSMA showed a small negative bias and slightly larger RMSE than most of the remaining methods. However, the increase in bias and RMSE in the Many Labs 2 data is decisively outweighed by the advantage in performance obtained in the "Evaluating RoBMA on Registered Replication Reports" section. This result, in conjunction with the results reported in the "Evaluating RoBMA Through Simulation Studies" section, show that the RoBMA's performance cannot be explained by a systematic underestimation of effect sizes.

Table E1

Performance of 13 Publication Bias Correction Methods for 28 Meta-Analysis from Many Labs 2 Compared to a "Gold Standard" established with Fixed-Effect Meta-Analytic Models.

Method	FPR	FNR	Undecided	OF	Bias	RMSE
WAAP-WLS	0.000	0.056		1.005	0.002	0.011
TF	0.100	0.000		0.991	-0.004	0.044
Random Effects (DL)	0.000	0.000		1.035	0.013	0.035
3PSM	0.000	0.000		1.033	0.013	0.042
RoBMA-old	0.000	0.000	0.071	0.961	-0.015	0.043
AK1	0.000	0.056		1.083	0.017	0.044
4PSM	0.000	0.167		1.048	0.019	0.063
AK2	*0.167*	*0.200*	*0.000*	*0.631*	*0.020*	*0.051*
RoBMA-PSMA	0.000	0.000	0.214	0.897	-0.040	0.070
PET-PEESE	0.100	0.444		0.820	-0.070	0.170
p-curve				*1.352*	*0.058*	*0.195*
EK	0.100	0.444		0.733	-0.103	0.259
$p-$uniform	0.571	0.182		1.353	0.155	0.745

Note. FPR = false positive rate, FNR = false negative rate, Undecided = undecided evidence, OF = overestimation factor, and RMSE = root mean square error. The results in *gray italic* are conditional on convergence: p-uniform and p-curve did not converge in four cases (p-uniform also did not provide test for the effect in 10 cases) and AK2 did not converge in 17 cases. The rows are ordered based on combined log scores performance of the abs(log(OF), abs(Bias), and RMSE (not shown).

Appendix F

Additional Results from the Hong and Reed (2020) Simulation Study

Table F1

Mean Square Error (MSE) in the Carter et al. (2019) simulation environment stratified by publication bias.

Rank	No-QRP	MSE	Medium-QRP	MSE	High-QRP	MSE
1	*3PSM*	*0.012*	RoBMA-old	0.011	RoBMA-old	0.011
2	RoBMA-old	0.013	WAAP-WLS	0.018	WAAP-WLS	0.018
3	RoBMA-PSMA	0.014	*p*-uniform	0.021	*p*-uniform	0.019
4	WAAP-WLS	0.018	TF	0.023	TF	0.025
5	TF	0.018	*3PSM*	*0.023*	*p*-curve	0.029
6	*4PSM*	*0.021*	PET-PEESE	0.027	PET-PEESE	0.032
7	PET-PEESE	0.022	EK	0.033	*3PSM*	*0.035*
8	EK	0.027	*4PSM*	*0.033*	EK	0.038
9	Random Effects (DL)	0.039	*AK2**	*0.034*	*4PSM*	*0.040*
10	*p*-uniform	0.042	RoBMA-PSMA	0.039	Random Effects (DL)	0.052
11	*p*-curve	0.156	*p*-curve	0.041	RoBMA-PSMA	0.056
12	AK1*	0.620	Random Effects (DL)	0.047	AK1*	0.134
13	*AK2**	*2.515*	AK1*	0.086	*AK2**	*6.127*

Note. *The difference of performance in terms of MSE for AK1 and AK2 between our and Hong and Reed (2020) is a result of us not omitting 5% of the most extreme estimates.

Methods in *gray italic* converged in less than 90% repetitions in a given simulation environment.

Different columns correspond to the conditions described in Carter et al. (2019); "no-QRP" condition corresponds to lack of questionable research practices (QRPs), "medium-QRP" condition corresponds to mix of no QRPs (30%), moderate QRPs (50%), and strong QRPs (20%), and "high-QRP" condition corresponds to a mix of no QRPs (10%), moderate strategy QRPs (40%), and strong QRPs (50%).

Table F2

Bias in the Carter et al. (2019) simulation environment stratified by publication bias.

Rank	None	Bias	Medium	Bias	High	Bias
1	*3PSM*	*0.028*	PET-PEESE	0.059	WAAP-WLS	0.063
2	*4PSM*	*0.041*	WAAP-WLS	0.062	RoBMA-old	0.067
3	RoBMA-PSMA	0.045	RoBMA-old	0.063	PET-PEESE	0.070
4	PET-PEESE	0.047	AK1	0.065	AK1	0.071
5	EK	0.050	EK	0.075	EK	0.093
6	WAAP-WLS	0.061	*3PSM*	*0.091*	*p*-uniform	0.094
7	RoBMA-old	0.063	TF	0.094	*p*-curve	0.096
8	AK1	0.064	*p*-uniform	0.095	TF	0.100
9	*AK2*	*0.067*	*p*-curve	0.100	*3PSM*	*0.124*
10	TF	0.080	*4PSM*	*0.113*	*4PSM*	*0.134*
11	Random Effects (DL)	0.128	*AK2*	*0.123*	RoBMA-PSMA	0.161
12	*p*-uniform	0.129	RoBMA-PSMA	0.124	*AK2*	*0.166*
13	*p*-curve	0.158	Random Effects (DL)	0.154	Random Effects (DL)	0.167

Note. Methods in *gray italic* converged in less than 90% repetitions in a given simulation environment.

Different columns correspond to the conditions described in Carter et al. (2019); "no-QRP" condition corresponds to lack of questionable research practices (QRPs), "medium-QRP" condition corresponds to mix of no QRPs (30%), moderate QRPs (50%), and strong QRPs (20%), and "high-QRP" condition corresponds to a mix of no QRPs (10%), moderate strategy QRPs (40%), and strong QRPs (50%).