

# Robust Bayesian Meta-Analysis: Addressing Publication Bias With Model-Averaging

Maximilian Maier<sup>1</sup>, František Bartoš<sup>1, 2</sup>, and Eric-Jan Wagenmakers<sup>1</sup>

<sup>1</sup> Department of Psychology, University of Amsterdam

<sup>2</sup> Department of Psychology, Charles University



## Abstract

Meta-analysis is an important quantitative tool for cumulative science, but its application is frustrated by publication bias. In order to test and adjust for publication bias, we extend model-averaged Bayesian meta-analysis with selection models. The resulting robust Bayesian meta-analysis (RoBMA) methodology does not require all-or-none decisions about the presence of publication bias, can quantify evidence in favor of the absence of publication bias, and performs well under high heterogeneity. By model-averaging over a set of 12 models, RoBMA is relatively robust to model misspecification and simulations show that it outperforms existing methods. We demonstrate that RoBMA finds evidence for the absence of publication bias in Registered Replication Reports and reliably avoids false positives. We provide an implementation in R so that researchers can easily use the new methodology in practice.

## Translational Abstract

Meta-analysis is an essential tool to synthesize information from a series of primary studies. However, the application of meta-analysis is often frustrated by publication bias – the fact that statistically significant results are published more often than nonsignificant results. To alleviate the problem of publication bias we developed a robust Bayesian meta-analysis (RoBMA). RoBMA applies a series of meta-analytic models to the data simultaneously and estimates the effect size by taking all models into account. RoBMA can quantify evidence for the presence as well as the absence of publication bias, RoBMA can correct for publication bias in cases where the true effect size differs between studies, and RoBMA does not require all-or-none decisions. We illustrate RoBMA with a meta-analysis on violent video games and aggressive behavior, and apply it to the ManyLabs 2 data for which we know publication bias to be absent. Simulations suggest that RoBMA provides a valuable complement to current methods for meta-analysis.

**Keywords:** evidence, heterogeneity, selection models

**Supplemental materials:** <https://doi.org/10.1037/met0000405.supp>

This article was published Online First May 19, 2022.

Maximilian Maier  <https://orcid.org/0000-0002-9873-6096>

František Bartoš  <https://orcid.org/0000-0002-0018-5573>

Eric-Jan Wagenmakers  <https://orcid.org/0000-0003-1596-1034>

Maximilian Maier and František Bartoš contributed equally to the article.

This project was supported in part by a Vici Grant (016.Vici.170.083) to Eric-Jan Wagenmakers. 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.

Part of this work was presented at the 2020 Annual Meeting of the Psychometric Society (IMPS). A preprint has been deposited on PsyArXiv at <https://psyarxiv.com/u4cns/> and a blog post that summarizes the preprint is at <https://tinyurl.com/yymahyvz>. Parts of this work were also based on an internship project of František Bartoš at the Psychological Methods Unit of the University of Amsterdam.

 The data are available at <https://osf.io/y354c/>

 The experiment materials are available at <https://osf.io/y354c/>

Correspondence concerning this article should be addressed to Maximilian Maier, who is now at Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom. Email: [maximilianmaier0401@gmail.com](mailto:maximilianmaier0401@gmail.com)

Meta-analysis is the method of choice for aggregating evidence across a set of primary studies. However, studies that report statistically significant results (i.e.,  $p < .05$ ) are more likely to get published (Rosenthal & Gaito, 1964; Scheel et al., 2020; Wicherts, 2017), and this preferential publishing results in an inflated effect size estimate—a phenomenon known as publication bias (Rothstein et al., 2005). A variety of methods have been developed to test and adjust for publication bias. Citation counts suggest that funnel-based methods such as trim-and-fill (Duval & Tweedie, 2000) and Egger’s regression (Egger et al., 1997) are most popular.<sup>1</sup> Alternative methods such as  $p$ -curve (Simonsohn et al., 2014) and PET-PEESE (Carter & McCullough, 2014) have recently been proposed and are increasing in popularity (for reviews see Carter et al., 2019 and Renkewitz & Keiner, 2019). The existing methods are predominantly based on the frequentist framework for statistical inference. In other words, these methods aim to control the

<sup>1</sup> At the time of writing, January 19th 2021, Google Scholar indicated that these methods had 7, 710 and 32, 286 citations, respectively.

long-term rate of false positives such that it does not exceed some frequency  $\alpha$ . For an individual test, the methodology usually reduces to computing a  $p$ -value (i.e., the probability of encountering a test statistic at least as extreme as the one observed, under the assumption that the null-hypothesis of no difference holds) and assessing whether or not  $p < \alpha$ , with  $\alpha$  usually set to 5% (Bartoš & Maier, 2019; Neyman, 1977). However, most existing methods to test and adjust for publication bias share several limitations.

A first limitation is that frequentist methods require researchers to decide whether or not to adjust for publication bias in all-or-none fashion, based on whether the  $p$ -value is smaller than the  $\alpha$  level. This choice fails to take model uncertainty into account and, consequently, slight differences in  $p$ -values (e.g., .049 vs. .051) can lead to consequential changes in mean estimates. This problem is exacerbated because the number of primary studies in a meta-analysis is often so low that confident inferences about publication bias are beyond reach.

A second limitation is that frequentist methods cannot quantify evidence in favor of the absence of publication bias; a non-significant  $p$ -value may indicate evidence of absence or absence of evidence (Gallistel, 2009; Rouder et al., 2009). In practice, researchers often assume that they do not need to worry about publication bias whenever a test for publication bias is not significant. However, publication bias might still be present, with the test lacking the statistical power to detect it (Renkewitz & Keiner, 2019).

A third limitation is that frequentist methods assume knowledge of the sampling plan with which the studies accumulate over time (ter Schure & Grünwald, 2019). When executing a meta-analysis on  $K$  primary studies, the implicit assumption is that an experimenter planned to conduct  $K$  studies and then stopped. This sampling plan is not realistic; studies accumulate over time, either indefinitely or until a point has been sufficiently proven or disproven. Assume a frequentist meta-analysis has been executed based on the standard assumption that  $K$  studies (no more, no less) were ever going to be reported. Next, study  $K + 1$  comes in. For frequentist methods, this presents a conundrum, as conducting another test invokes a multiple comparisons problem, with all of the  $\alpha$  already spent on the first test.

A fourth limitation is that most existing methods fail under high across-study heterogeneity, as shown in simulation studies (Carter et al., 2019; McShane et al., 2016; Renkewitz & Keiner, 2019). A notable exception are selection models (e.g., Iyengar & Greenhouse, 1988; Vevea & Hedges, 1995) which perform well even under high heterogeneity, although they have other problems such as poor convergence under few primary studies (Field & Gillett, 2010; Terrin et al., 2003).

To overcome the above limitations, we developed an extension of Bayesian model-averaged meta-analysis (BMA). Below we start by introducing BMA (Gronau, van Erp et al. (2017)) which is the first component of our method. Next, we introduce selection models as a framework to test and adjust for publication bias—the second component of our method. These two components are then brought together to yield Robust Bayesian Meta-Analysis (RoBMA), a model-averaged meta-analytical framework in the presence of publication bias.

We illustrate the methods by applying them to two data sets. The first data set concerns the meta-analysis by Anderson et al. (2010) on violent video games and aggressive behavior. Anderson et al. (2010) concluded that “exposure to violent video games is a causal risk factor for increased aggressive behavior.” However, several researchers have questioned this conclusion, suggesting instead that the results are contaminated by publication bias (e.g., Ferguson et al., 2010; Hilgard et al., 2017). We focus our analysis on the 27 experimental studies that meet Anderson and colleagues’ (2010) best practice criteria.<sup>2</sup> The second data set concerns 28 meta-analyses from the Many Labs 2 data (Klein et al., 2018); this data set is particularly useful because we know the ground truth: Publication bias is absent. Consequently, the different tests for publication bias can be compared based on how many false positives they produce. Finally, we present a simulation study comparing RoBMA to other methods under different conditions for heterogeneity, effect size, and publication bias.

## Component 1: Model-Averaged Bayesian Meta-Analysis

Our method is based on BMA as introduced in Gronau, van Erp et al. (2017). This method applies four models simultaneously (see Figure 1), differentiated according to whether they assume an overall effect to be absent or present, and whether they assume across-study heterogeneity to be absent (i.e., fixed effects, when the effect size is the same across the primary studies) or present (i.e., random effects, when effect sizes differ across the primary studies). Thus, BMA features the following four models:

1. The fixed effects null hypothesis ( $\mathcal{H}_0^f$ )
2. The fixed effects alternative hypothesis ( $\mathcal{H}_1^f$ )
3. The random effects null hypothesis ( $\mathcal{H}_0^r$ )
4. The random effects alternative hypothesis ( $\mathcal{H}_1^r$ )

The default setting for BMA is to assume that all four models are equally likely a priori (cf. Figure 1). The relative plausibility of the four rival models is then updated according to Bayes’ theorem: Models that predict the observed data relatively well receive a boost in plausibility, whereas models that predict the data relatively poorly suffer a decline (Wagenmakers et al., 2016).

With only two models under consideration, their relative predictive performance is given by the Bayes factor (Etz & Wagenmakers, 2017; Kass & Raftery, 1995; Rouder & Morey, 2019; Wrinch & Jeffreys, 1921). To illustrate, Equation 1 shows that the Bayes factor for the random effects alternative hypothesis  $\mathcal{H}_1^r$  versus the fixed effects alternative hypothesis  $\mathcal{H}_1^f$  equals the change from prior to posterior odds brought about by the data:

<sup>2</sup>The data are available at <https://github.com/Joe-Hilgard/Anderson-meta>.

$$\underbrace{\frac{p(\text{data} | \mathcal{H}_1^r)}{p(\text{data} | \mathcal{H}_1^f)}}_{\text{Bayes factor}} = \underbrace{\frac{p(\mathcal{H}_1^r | \text{data})}{p(\mathcal{H}_1^f | \text{data})}}_{\text{Posterior odds}} \bigg/ \underbrace{\frac{p(\mathcal{H}_1^r)}{p(\mathcal{H}_1^f)}}_{\text{Prior odds}}. \quad (1)$$

When  $\mathcal{H}_1^r$  and  $\mathcal{H}_1^f$  are equally plausible a priori, the prior odds equal 1 and the Bayes factor reduces to the posterior odds. For instance, if  $\mathcal{H}_1^r$  predicted the observed data three times better than  $\mathcal{H}_1^f$  (i.e., the Bayes factor equals three) then the posterior odds also equal three, meaning that the posterior model probability for  $\mathcal{H}_1^r$  equals  $3/4 = .75$ , leaving  $.25$  for  $\mathcal{H}_1^f$ . A common rule of thumb is to regard Bayes factors between one and three as weak or anecdotal evidence, Bayes factors between three and ten as moderate evidence, and Bayes factors larger than ten as strong evidence (e.g., Jeffreys, 1939, Appendix I; Lee & Wagenmakers, 2013, p. 105). However, “This set of labels facilitates scientific communication but should only be considered an approximate descriptive articulation of different standards of evidence” (Lee & Wagenmakers, 2013, p. 105), by reducing a continuous measure for the strength of evidence to a discrete category, information is inevitably lost (Jeffreys, 1938).<sup>3</sup>

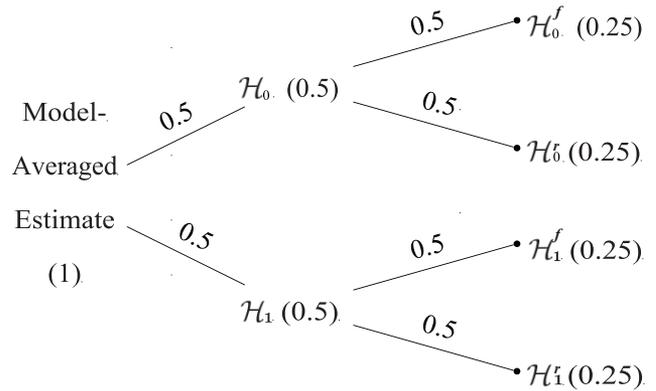
With more than two models under consideration, as is the case for BMA, one may consider the change from prior to posterior odds for sets of models defined by including versus excluding a specific parameter. This “inclusion Bayes factor” allows BMA to quantify evidence for the presence of a meta-analytic effect and for across-study heterogeneity by taking all four models into account simultaneously. For example, in the context of assessing the evidence for and against the presence of a meta-analytic effect the inclusion Bayes factor  $\text{BF}_{10}$  compares  $\mathcal{H}_1 = \{\mathcal{H}_1^f, \mathcal{H}_1^r\}$  to  $\mathcal{H}_0 = \{\mathcal{H}_0^f, \mathcal{H}_0^r\}$ . Specifically, the inclusion Bayes factor for assessing the evidence for the presence of a meta-analytic effect is given by:

$$\underbrace{\text{BF}_{10}}_{\text{Inclusion Bayes factor for effect}} = \frac{p(\mathcal{H}_1^f | \text{data}) + p(\mathcal{H}_1^r | \text{data})}{p(\mathcal{H}_0^f | \text{data}) + p(\mathcal{H}_0^r | \text{data})} \bigg/ \frac{p(\mathcal{H}_1^f) + p(\mathcal{H}_1^r)}{p(\mathcal{H}_0^f) + p(\mathcal{H}_0^r)}. \quad (2)$$

The inclusion Bayes factor  $\text{BF}_{rf}$  for the test of heterogeneity is obtained similarly, by contrasting  $\mathcal{H}_r = \{\mathcal{H}_0^r, \mathcal{H}_1^r\}$  to  $\mathcal{H}_f = \{\mathcal{H}_0^f, \mathcal{H}_1^f\}$ . Note that inclusion Bayes factors are able to distinguish between “absence of evidence” (i.e., when  $\text{BF}_{10}$  and  $\text{BF}_{rf}$  are close to 1) and “evidence of absence” (i.e., when  $\text{BF}_{10}$  and  $\text{BF}_{rf}$  are close to 0, or, equivalently, when  $\text{BF}_{01}$  and  $\text{BF}_{fr}$  are much larger than 1).<sup>4</sup> Also note that Bayes factors quantify evidence on a continuous scale, and this evidence can be seamlessly updated as more studies accumulate (Berger & Wolpert, 1988; Rouder, 2014).

In addition to testing for effect size and across-study heterogeneity, the BMA framework can also be used to estimate parameters. For instance, the overall posterior distribution for the meta-analytic effect is a weighted combination of the estimates from each of the four individual models, with the weights equal to the model’s posterior probability. This process is known as Bayesian model-averaging (Hoeting et al., 1999; for a conceptual introduction see Hinne et al., 2020) and it obviates the need to base inference on a single model in all-or-none fashion. This is particularly advantageous when there is considerable posterior uncertainty about whether a fixed effects or a random effects

**Figure 1**  
Prior Model Probabilities of BMA



Note. Total probabilities are displayed on the nodes and conditional probabilities on the edges.  $\mathcal{H}_0$  denotes the models assuming the null hypotheses to be true,  $\mathcal{H}_1$  denotes the models assuming the alternative hypotheses to be true.  $\mathcal{H}^f$  denotes fixed effects models,  $\mathcal{H}^r$  random effects models. BMA = Bayesian model-averaged meta-analysis.

model is most appropriate. In such scenarios, the Bayesian model-averaging result is substantially influenced by both model classes.

Recently, BMA has been used to estimate the effect of power posing (Gronau, van Erp, et al., 2017), to test compensatory control theory (Hoogveen et al., in press), study the effectiveness of descriptive social norms (Scheibehenne et al., 2017), to analyze a registered replication report on facial feedback (Hinne et al., 2020), to investigate how researchers subjective decisions when designing studies influence the results (Landy et al., 2020), and to conduct a Bayesian multiverse analysis on the data of Many Labs 4 (Haaf et al., 2020). A primer on BMA is presented in Gronau, Heck, et al. (2020). However, most past applications concern prospective meta-analyses, that is, meta-analyses on registered reports. The main reason for this is that BMA cannot yet adjust for publication bias.

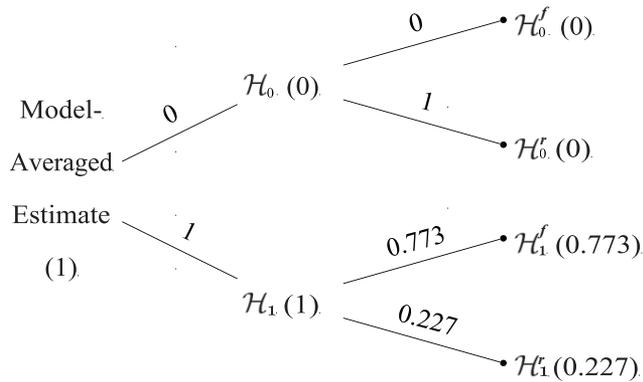
We illustrate BMA using the data from Anderson et al. (2010).<sup>5</sup> As can be seen in Figure 1 we assume equal prior probabilities for the four different models. Figure 2 shows the associated posterior model probabilities. The model that best predicts the data is  $\mathcal{H}_1^f$ , the fixed effects model that assumes the presence of an effect. This model has a posterior probability of .773. The second-best model is  $\mathcal{H}_1^r$ , the random effects model that assumes the presence of an effect with a posterior probability of .277. When comparing sets of

<sup>3</sup> On p. 378, Jeffreys (1938) discusses thresholds on  $\text{BF}_{01}$  (which he termed  $K$ ) as follows: “we are at liberty to surround  $K = 1$  by two other values and say that within this range the data are not sufficiently decisive, and even this device would be purely one of convenience and sacrifice some information given by the actual values of  $K$ .”

<sup>4</sup> The first and second subscript of the Bayes factor denote the model in the numerator and the denominator, respectively; so  $\text{BF}_{AB}$  means  $p(\text{data} | \mathcal{H}_A) / p(\text{data} | \mathcal{H}_B)$ , and  $\text{BF}_{BA}$  means  $p(\text{data} | \mathcal{H}_B) / p(\text{data} | \mathcal{H}_A)$ . When Bayes factors are lower than one, it is recommended to switch numerator and denominator, because “ $\text{BF}_{AB} = 0.2$ , the data are 0.2 times more likely under  $\mathcal{H}_A$  than under  $\mathcal{H}_B$ ” is more difficult to interpret than the equivalent statement “ $\text{BF}_{BA} = 5$ , the data are five times more likely under  $\mathcal{H}_B$  than under  $\mathcal{H}_A$ ”.

<sup>5</sup> We fit BMA using the `metaBMA` R package (Heck et al., 2017; see also Gronau, Singmann, & Wagenmakers, 2020).

**Figure 2**  
 Posterior Model Probabilities for a BMA of Anderson et al. (2010)



*Note.* Total probabilities are displayed on the nodes and conditional probabilities on the edges.  $\mathcal{H}_0$  denotes the models assuming the null hypotheses to be true,  $\mathcal{H}_1$  denotes the models assuming the alternative hypotheses to be true.  $\mathcal{H}^f$  denotes fixed effects models,  $\mathcal{H}^r$  random effects models. BMA = Bayesian model-averaged meta-analysis.

models using the inclusion Bayes factor, we find weak support favoring the fixed effects models over the random effects model (i.e.,  $\text{BF}_{fr} = 3.41$ )<sup>6</sup> and overwhelming support for the alternative of an effect of violent video games on aggressive behavior ( $\text{BF}_{10} = 8.75 \times 10^9$ ).

The model-averaged posterior mean effect size estimate  $r = .211$ , 95% CI [.171, .247] is then obtained after weighting the effect size implied by each model according to its posterior probability. In this particular case the estimate is based most strongly on  $\mathcal{H}_1^f$  and  $\mathcal{H}_1^r$ , and almost not at all influenced by  $\mathcal{H}_0^f$  and  $\mathcal{H}_0^r$ .

However, Ferguson et al. (2010) and Hilgard et al. (2017) have argued that the original effect size reported in Anderson et al. (2010) was overestimated due to publication bias. BMA cannot test or adjust for the possibility of publication bias; to do so, we now turn to selection models.

## Component 2: Selection Models

Our approach to publication bias is based on selection models that operate on  $p$ -values (e.g., Hedges, 1992; Iyengar & Greenhouse, 1988; Vevea & Hedges, 1995). These selection models use weighted distributions to account for the proportion of studies that are missing because they yielded nonsignificant results. The researcher specifies the  $p$ -value cut-offs that drive publication bias (usually  $p = .05$ ). The selection model then estimates how likely studies in nonsignificant intervals are to be published compared with the interval with the highest publication probability (usually  $p < .05$ ). The pooled effect size estimate accounts for the estimated publication bias by giving more weight to studies in intervals with lower publication probability (usually nonsignificant studies). Selection models can also test for publication bias. To do so, the unadjusted model—either a fixed or a random effects meta-analytic model—is compared with the corresponding selection model using a likelihood ratio test. This

test uses a  $\chi^2$  statistic to reject the null hypothesis of no publication bias whenever the  $p$ -value is lower than the conventional 5%  $\alpha$  level, or lower than an  $\alpha$  level of 10%, a more lenient threshold which is sometimes advocated because it increases power (Renkewitz & Keiner, 2019).

Another type of selection models, so-called “Copas” selection models, specify the probability of a study being published as a function of both effect sizes and standard errors (Copas, 1999; Copas & Li, 1997; Copas & Shi, 2001; for recent Bayesian versions see Bai et al., 2020; Mavridis et al., 2013). In this article we focus on the selection models based on  $p$ -values; these selection models have been extensively tested in simulation studies and are most popular in practice (e.g., Carter et al., 2019).

Selection models are one of the few methods that have been shown to work well even under high heterogeneity (Carter et al., 2019; McShane et al., 2016) and can be specified flexibly according to the assumed model of publication bias (e.g., assuming different probability of publication for marginally significant studies with  $p \in [.05, .10]$ ). Finally, the assumed model of the publication process (i.e., selection based on statistical significance) is a well-founded model for how publication bias operates in practice (Masicampo & Lalande, 2012; Rosenthal & Gaito, 1964; Wicherts, 2017).

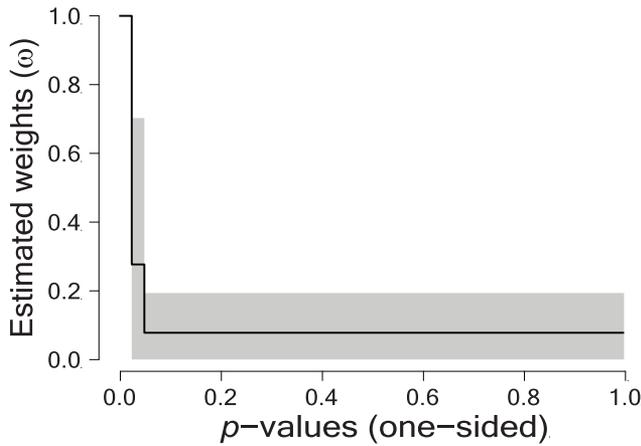
To illustrate the methodology of selection models we revisit the example of Anderson et al. (2010). We use the `weightR` R package (Coburn et al., 2019) and initially specified a two-sided selection with different probabilities for significant and marginally significant studies (i.e., with cutoffs at  $p = .05, .10$ ).<sup>7</sup> Because all effects were positive we were forced to alter the model and include only  $p$ -value cut-offs in the expected direction (i.e., one-sided selection). We first tested for presence of heterogeneity (with  $\alpha = .05$ ). The test is not significant,  $Q(22) = 15.09$ ,  $p = .89$ , so we proceeded with a fixed effect model (Bartoš et al., 2020). The fixed effect model indicates significant publication bias (i.e.,  $\chi^2(2) = 13.58$ ,  $p = .001$ , using  $\alpha = .10$ ). Estimates for the relative publication probability in the different intervals indicate that marginally significant results are only 28% as likely to be published than significant findings and nonsignificant results are only about 8% as likely to be published (see Figure 3). Adjusting for this publication bias corrects the effect size estimate downward although the effect remains significant,  $r = .140$ , 95% CI [.077, .201],  $z = 4.33$ ,  $p < .001$ .

However, in addition to the general limitations of frequentist methods (i.e., ignoring model uncertainty, inability to quantify evidence for the absence of an effect, and dependence on an unknown sampling plan), selection models have more specific limitations as well. In particular, they often cannot be estimated under small sample size, especially when there are few studies in any of the  $p$ -value intervals (e.g., Terrin et al., 2003; but see Citkowicz & Vevea, 2017; for a recent attempt

<sup>6</sup> As a calculation example based on Equation 2: The prior odds are  $(0.25 + 0.25)/(0.25 + 0.25) = 1$  and the posterior odds  $(0.773 + 0)/(0.227 + 0) = 3.41$  corresponding to a Bayes factor of  $3.41/1 = 3.41$ .

<sup>7</sup> The `weightR` R package allows the specification of  $p$ -values cut-offs on one-sided  $p$ -values directly. To achieve two-sided selection with significant and marginally significant studies, the following one-sided  $p$ -value cutoffs need to be specified (0.025, 0.05, 0.95, 0.975).

**Figure 3**  
 Estimated Weight Function With 95% CI for Anderson et al. (2010)



*Note.* Results are based on the selection model outlined in Vevea and Hedges (1995). The weight for significant studies is set to one. “Marginally” significant and nonsignificant studies are then weighted according to their relative probability of publication.

to solve this problem by modeling the weight function as a beta distribution). Furthermore, selection models are mainly available in the `weightr` package (Coburn et al., 2019) and the `selectMeta` package (Rufibach, 2015) in the programming language R, limiting their accessibility. This is arguably

the reason why they have not been adopted more widely despite their desirable properties (Card, 2015, p. 275).

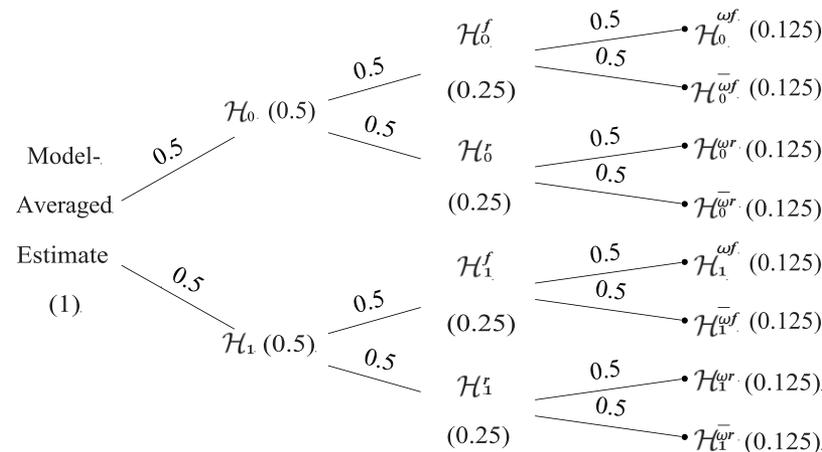
### Combining Model-Averaged Bayesian Meta-Analysis and Selection Models: Robust Bayesian Meta-Analysis (RoBMA)

We propose robust Bayesian meta-analysis (RoBMA), a Bayesian multimodel method that aims to overcome the limitations of existing procedures. RoBMA is an extension of BMA obtained by adding selection models to account for publication bias. This allows model-averaging across a larger set of models, ones that assume publication bias and ones that do not.

As shown in Figure 4, RoBMA partitions the model space according to three characteristics. As in BMA, we differentiate between models that do ( $\mathcal{H}_1$ ) and do not ( $\mathcal{H}_0$ ) assume the presence of an effect, and between models that assume homogeneity (fixed effects;  $\mathcal{H}^f$ ) and heterogeneity (random effects;  $\mathcal{H}^r$ ). RoBMA generalizes BMA by also differentiating models assuming the absence of publication bias ( $\mathcal{H}^{\bar{\omega}}$ ) and models assuming its presence ( $\mathcal{H}^{\omega}$ ). Comparing the predictive accuracy of the latter two model types allows us to test for the presence or absence of publication bias.

The models assuming the absence of publication bias are identical to the models described by Gronau, Sarafoglou et al. (2017). For the models that assume the presence of publication bias, we specified two selection processes (not shown in Figure 4). The first selection process is a two-step model that distinguishes only between significant and nonsignificant studies assuming that significant studies are more likely to be published. The second selection process is a three-step model that includes marginally significant

**Figure 4**  
 Different Models Included in RoBMA and Their Prior Probabilities



*Note.* RoBMA features two publication bias models (i.e., two and three-steps) that are not distinguished in the plot. Total probabilities are displayed on the nodes and conditional probabilities on the edges.  $\mathcal{H}_0$  denotes the models assuming the null hypotheses to be true,  $\mathcal{H}_1$  denotes the models assuming the alternative hypotheses to be true.  $\mathcal{H}^f$  denotes fixed effects models,  $\mathcal{H}^r$  random effects models,  $\mathcal{H}^{\omega}$  publication bias models, and  $\mathcal{H}^{\bar{\omega}}$  models without publication bias. RoBMA = resulting robust Bayesian meta-analysis.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

results as a separate category. We specify both processes to be independent of the direction of the effect. Therefore, marginal significance also corresponds to one-sided significance at the  $\alpha = .05$  level. By including several models for publication bias the overall approach is relatively robust to model misspecification. Overall, RoBMA includes the following 12 (i.e.,  $2 \times 2 \times 3$ ) different models of the data generating process:

1. The fixed effects null hypothesis without publication bias ( $\mathcal{H}_0^{of}$ )
2. The fixed effects alternative hypothesis without publication bias ( $\mathcal{H}_1^{of}$ )
3. The random effects null hypothesis without publication bias ( $\mathcal{H}_0^{or}$ )
4. The random effects alternative hypothesis without publication bias ( $\mathcal{H}_1^{or}$ )
5. The fixed effects weighted null hypothesis with publication bias ( $\mathcal{H}_0^{of}$ ) which subdivides into
  - (a) a model with two-step selection ( $\mathcal{H}_0^{of2f}$ )
  - (b) a model with three-step selection ( $\mathcal{H}_0^{of3f}$ )
6. The fixed effects weighted alternative hypothesis with publication bias ( $\mathcal{H}_1^{of}$ ) which subdivides into
  - (a) a model with two-step selection ( $\mathcal{H}_1^{of2f}$ )
  - (b) a model with three-step selection ( $\mathcal{H}_1^{of3f}$ )
7. The random effects weighted null hypothesis with publication bias ( $\mathcal{H}_0^{or}$ ) which subdivides into
  - (a) a model with two-step selection ( $\mathcal{H}_0^{or2r}$ )
  - (b) a model with three-step selection ( $\mathcal{H}_0^{or3r}$ )
8. The random effects weighted alternative hypothesis ( $\mathcal{H}_1^{or}$ ) which subdivides into
  - (a) a model with two-step selection ( $\mathcal{H}_1^{or2r}$ )
  - (b) a model with three-step selection ( $\mathcal{H}_1^{or3r}$ )

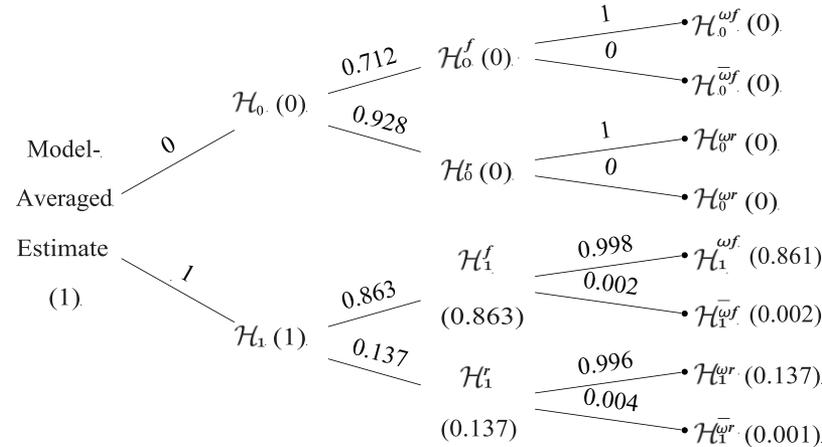
As can be seen in Figure 4, we distribute prior probability evenly across the different pairs of model classes. In other words, a priori we assume a probability of .5 of publication bias, a .5 probability for heterogeneity, and a .5 probability for a true effect (see Figure 4). In the case of the publication bias models, the prior probability is further divided, assigning .5 probability to the two-step models and .5 probability to the three-step models. Regarding the probability of the individual models, this results in a probability of .125 (.5/4) for each of the models assuming no publication bias and .0625 (.5/8) for each of the models assuming publication bias.

Regarding the parameter prior distributions, we used a normal distribution with mean zero and a standard deviation of 1 for the effect size. This prior distribution assigns high probability to a wide range of effect sizes, without being as heavy-tailed as the popular Cauchy distribution. For heterogeneity, we used inverse gamma with  $\alpha = 1$  and  $\beta = .15$ . This estimate is based on empirical heterogeneities for the field of psychology (van Erp et al., 2017). The prior distribution on the weights parameter is the cumulative sum of the Dirichlet distribution with  $\alpha$  set to (1, 1) for the two-interval model and (1, 1, 1) for the three-interval model. Thus, the weight for the significant studies is always equal to one and the weights increase monotonically with decreasing  $p$ -values. We fit the models with the RoBMA (Bartoš & Maier, 2020) R package. The package estimates the individual models via MCMC sampling using JAGS (Plummer, 2003), specifically the `runjags` R package (Denwood, 2016). While JAGS incorporates most common distributions we also added custom weighted distributions that were required to implement the selection models. To obtain the marginal likelihoods, the package uses bridge sampling (Gronau, Sarafoglou, et al., 2017) implemented in the `bridgesampling` R package (Gronau, Singmann, & Wagenmakers, 2020). More detailed specifications of RoBMA can be found in Appendix A.

Although this is the default setup and the setup that will be used in the examples and in the simulation study of this paper, we wish to stress that RoBMA can be customized in several ways. For example, some methodologists have argued that only random effects meta-analysis should be used (e.g., Borenstein, 2019). This advice can be followed by simply removing the fixed effects models from RoBMA. Apart from adding and removing models one can also alter the assignment of prior model probabilities. For instance, the belief that the null hypothesis is relatively unlikely may be incorporated by lowering the prior model probabilities for the associated models. Finally, in practice researchers should also feel free to use different parameter prior distributions when conducting meta-analyses, depending on their prior knowledge and the specific hypotheses under consideration. For example, when the possibility of very large effect sizes cannot be ruled out a Cauchy prior might be preferable over the normal prior distribution. An R vignette that explains how to modify RoBMA is available at <https://cran.microsoft.com/web/packages/RoBMA/vignettes/CustomEnsembles.html>.

To illustrate, we apply RoBMA to the example of Anderson et al. (2010). As can be seen in Figure 4, all model pairs are equally likely a priori. Figure 5 shows the associated posterior probabilities. After updating by the data the two most likely models are the weighted fixed effects alternative ( $H_1^{of}$ ) and the weighted random effects ( $H_1^{or}$ ) alternative. Those two models take up almost all posterior probability (86% and 13.6%). When comparing sets of models using the inclusion Bayes factor the data overwhelmingly support the presence of effect size ( $BF_{10} = 1.05 \times 10^7$ ) and publication bias ( $BF_{\omega\bar{\omega}} = 381.69$ ), while support for the fixed effects models is moderate ( $BF_{f\bar{f}} = 6.29$ ). The final inference is, therefore, based most strongly on the effect-present fixed and random effects models that include publication bias; consequently, the model-averaged posterior mean corrects the meta-analytical effect size estimate downward in comparison to BMA,  $r = .151$ , 95% CI [.094, .207].

**Figure 5**  
 Posterior Model Probabilities for the Different Types of Models in RoBMA After Updating on the Data of Anderson et al. (2010)



*Note.* RoBMA features two publication bias models (i.e., two and three steps) that are not distinguished in the plot. Total probabilities are displayed on the nodes and conditional probabilities on the edges.  $\mathcal{H}_0$  denotes the models assuming the null hypotheses to be true,  $\mathcal{H}_1$  denotes the models assuming the alternative hypotheses to be true.  $\mathcal{H}^f$  denotes fixed effects models,  $\mathcal{H}^r$  random effects models,  $\mathcal{H}^{wf}$  publication bias models and  $\mathcal{H}^{wr}$  models without publication bias. RoBMA = resulting robust Bayesian meta-analysis.

**RoBMA Benefits**

The first benefit of RoBMA is that Bayes factors can quantify evidence for the absence of publication bias. This is especially important because frequentist tests for publication bias usually have low power to detect publication bias if it is present (Renkewitz & Keiner, 2019). Therefore, a nonsignificant test for publication bias does not usually imply the absence of publication bias. RoBMA allows researchers to differentiate between absence of evidence and evidence of absence.

The second benefit of RoBMA is that the Bayesian framework mitigates the estimation problems that arise when some of the  $p$ -value intervals contain only few studies. If one of the  $p$ -value intervals misses sufficient observations, the weights are still constrained by prior distributions. Therefore, the model can also be estimated on few primary studies. However, in this context it is important to choose priors carefully and evaluate the robustness to prior specification by examining how the conclusions change when using different parameter priors on the weights.

The third benefit of RoBMA is that the Bayesian model-averaging removes the need for a variety of all-or-none  $p$ -value-based decisions required by frequentist methods. In other words, RoBMA allows researchers to retain all models, weighting their inferential impact by their posterior probability. The option to include multiple models of the publication process also makes selection models more robust to model misspecification. Model-averaging should be especially beneficial when there is substantial uncertainty about the best model. In addition, model-averaging avoids the danger of making incorrect analysis choices since the data will directly guide the inference to be based most strongly on the best model.

Unsurprisingly, we are not the first to realize these benefits. Non model-averaging Bayesian selection models have been proposed over 20 years ago (Givens et al., 1995; Larose & Dey, 1998; Silliman, 1997; Smith et al., 2001) and more recently Guan and Vandekerckhove (2016) showed model-averaging over fixed effects models. However, heterogeneous effect sizes are the norm, rather than the exception in psychology (e.g., McShane et al., 2016). Therefore, we extend model-averaging to include random effects models. In addition, we add Bayes factor tests for the presence or absence of publication bias. Finally, we implement the Bayesian model-averaged meta-analysis in an R package RoBMA (Bartoš & Maier, 2020) and we also recently developed a version for the open-source statistical software program JASP (Bartoš et al., 2020; JASP Team, 2020). This allows researchers without programming experience to conduct state-of-the-art publication bias-adjusted meta-analysis.

In the remainder of this article, we test RoBMA and compare it with existing meta-analytic techniques. First, we used the 28 studies from Many Labs 2—here we know that publication bias is absent—to assess the possible tendency to produce false positives when testing for publication bias. Second we performed a simulation study to compare RoBMA with other methods under different conditions for publication bias, effect size, heterogeneity, and number of primary studies.

**Testing False Positives on Many Labs 2**

It is hard to assess the performance of different methods on published meta-analyses because the true parameters are usually unknown. However, it is possible to assess the false positive rate

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

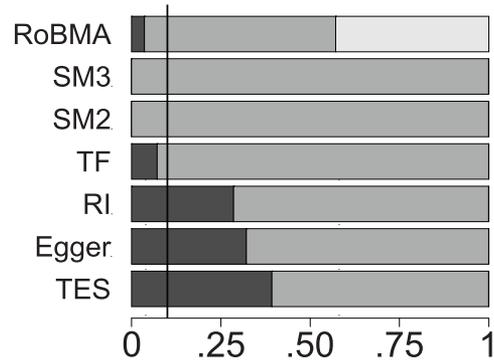
of tests for publication bias using Registered Reports (Chambers, 2013, 2019). Registered Reports are subject to a two-stage peer review process. In the first, predata stage, peer review concerns only the introduction and method sections; when the experimental design and data analysis plan are sufficiently convincing, the article receives “in principle acceptance,” which means that the journal commits to publishing the results independent of the outcome (Chambers et al., 2015). Therefore, in the case of Registered Reports we know that all primary studies are published regardless of the result; in other words, if a method detects publication bias, this is a false positive finding. In addition, Registered Reports allow an empirical test of RoBMA’s ability to quantify evidence in favor of the absence of publication bias. Therefore, we compared RoBMA to different methods for publication bias using the registered replication report Many Labs 2 (Klein et al., 2018). This large-scale project attempted to replicate 28 classic and contemporary effects from psychology. Each effect was subject to a replication attempt from about half of the 125 participating labs (comprising 15,305 participants from 36 countries in total). The median Cohen’s  $d$  across the 28 replicated effects was .15. One effect had  $\tau$  larger .2, while the others all had values near  $\tau = .10$ .<sup>8</sup>

In addition to RoBMA, we included Vevea and Hedges (1995) selection models with two-steps or three-steps and two-sided selection. Similar to the Anderson example, we first applied the test for heterogeneity before proceeding to the test for publication bias.<sup>9</sup> In addition to RoBMA and the selection models, we also test for publication bias using trim and fill (Duval & Tweedie, 2000), replicability index (Schimmack, 2020), Egger regression (Egger et al., 1997; which corresponds to testing for publication bias using the PET part of PET-PEESE; Stanley & Doucouliagos, 2014), and the test for excess significance (Ioannidis & Trikalinos, 2007). For an explanation of these methods see Renkewitz and Keiner (2019) and Schimmack (2020). For Egger regression we used code from Carter et al. (2019); for replicability index and test for excess significance we used code provided by Dr. Schimmack; for trim and fill we used the `trimfill` function from the `metafor` (Viechtbauer, 2010) R package with the `RO` estimator; for selection models we used the `weightr` R package (Coburn et al., 2019), and for the RoBMA models we used the `ROBMA` R package (Bartoš & Maier, 2020). For frequentist methods we used  $\alpha = .10$  as significance level (Renkewitz & Keiner, 2019) whereas for RoBMA we used the inclusion  $BF_{w\omega} > 3$  as a discrete “evidence threshold” (e.g., Jeffreys, 1939). In practice, we recommend researchers interpret Bayes factors as a continuous measure of evidence rather than by reference to discrete thresholds. We use thresholds here only to interpret RoBMA’s performance in terms of false positive rate and power, thereby enabling a comparison to frequentist methods.

As can be seen in Figure 6, RoBMA, the two selection models (i.e., “SM2” and “SM3”), and trim and fill had either no or few false positives (0% for SM2 and SM3; 3.6% or 1/28 for RoBMA; 7.1% or 2/28 for trim and fill; i.e., results with  $p < .10$  or  $BF_{w\omega} > 3$ ). However, in contrast to the two frequentist selection models, RoBMA also provided evidence in favor of the absence of publication bias in 43% of cases (i.e., for 12/28 studies). The other methods (i.e., replicability index, Egger regression, and the test for excess significance) fare worse and incorrectly find evidence for publication bias in 29% of cases or more (i.e., 8/28 for replicability Index, 9/28 for Egger regression, 11/28 for the test for excess significance).

**Figure 6**

*False Positives of Different Methods Using the 28 Effects Studied in Many Labs 2 (Klein et al., 2018)*



*Note.* SM3 = three-step selection model; SM2 = two-step selection model; RI = replicability index; TES = test for excess significance; RoBMA = resulting robust Bayesian meta-analysis; TF = trim and fill. Black areas indicate evidence for publication bias ( $BF > 3, p < .10$ ), gray areas indicate ambiguous evidence regarding publication bias ( $1/3 < BF < 3, p > .10$ ), and white areas indicate evidence for the absence of publication bias ( $BF < 1/3$ , only applicable to RoBMA).

## Simulation Study

### Method

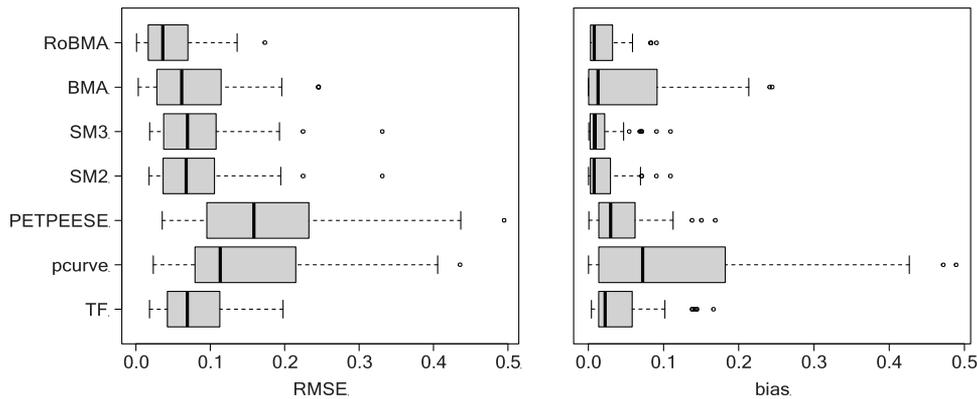
To test the performance of RoBMA under different conditions we designed a simulation study in which we manipulated the following four parameters:

1. The population mean:  $\mu = (0, 0.2, 0.5)$
2. The heterogeneity:  $\tau = (0, 0.2)$
3. The number of primary studies in each meta-analysis:  $K = (10, 30, 100)$
4. Three different conditions for publication bias with  $\omega$  denoting the probability of publication for (a) nonsignificant findings, (b) marginally significant findings, and (c) significant findings:
  - (a) No publication bias:  $\omega_1 = 1, \omega_2 = 1, \omega_3 = 1$
  - (b) Moderate publication bias:  $\omega_1 = 0.2, \omega_2 = 0.5, \omega_3 = 1$

<sup>8</sup> We used the reported  $t$ -statistics and sample sizes for studies analyzed using (Welch)  $t$ -tests (16 effects). For the remaining 12 effects, we computed  $t$ -statistic from the reported correlation coefficients.

<sup>9</sup> Because the frequentist implementation of selection models requires at least three  $p$ -values in each interval, we join intervals with fewer than three  $p$ -values. If there was only one interval left, we used either a random or fixed meta-analytical model based on the test of residual heterogeneity (with  $\alpha = .05$ ). Otherwise, we fitted a selection model, first tested for residual heterogeneity (with  $\alpha = .05$ ), selected either fixed or random effect model and then used a likelihood ratio test (with  $\alpha = .10$ ) to determine whether or not to use an adjusted model. If an adjusted model was selected but all of the estimated weights were higher than 1, we set the publication bias  $p$ -value to 1 and proceeded with the unadjusted model.

**Figure 7**  
Boxplots of RMSE and Bias Across Simulation Conditions



*Note.* SM3 = three-step selection model; SM2 = two-step selection model; TF = trim and fill; RoBMA = resulting robust Bayesian meta-analysis; PETPEESE = precision-effect test and precision-effect estimate with standard error; RMSE = root mean squared error. We omit 10 conditions with RMSE > .5 for *p*-curve, 1 condition with RMSE > .5 for PET-PEESE, and 4 conditions with bias > .5 for *p*-curve.

(c) Severe publication bias:  $\omega_1 = 0$ ,  $\omega_2 = 0$ ,  $\omega_3 = 1$

Significance was assessed based on two-sided *p*-values. The sample size of the primary studies in the meta-analysis was sampled from a negative binomial distribution based on empirical sample sizes that we extracted from Stanley et al. (2018; figure in Appendix B). We also conducted one simulation to assess method performance under extreme *p*-hacking of a null effect ( $\mu = 0$ ,  $\tau = 0$ ). The results of this simulation can be found in Appendix C. Each of the presented conditions was repeated 500 times.

In all simulations, we compared the performance of RoBMA to that of several different methods. To test for publication bias we used the same methods as in the Many Labs 2 example presented above. For effect size estimation and testing, we kept RoBMA, two and three-step selection models, and trim and fill, because the other methods are only suited to test for publication bias. Instead we added *p*-curve (Simonsohn et al., 2014), PET-PEESE (Stanley & Doucouliagos, 2014), and vanilla BMA (Gronau, van Erp, et al., 2017) to test for and estimate the meta-analytical effect size. A review and explanation of the different frequentist methods can be found in (Carter et al., 2019; Renkewitz & Keiner, 2019; Schimmack, 2020). The implementation of the additional methods for effect size estimation and testing, PET-PEESE, *p*-curve and Trim and Fill (estimation)<sup>10</sup> was based on code from Carter et al. (2019). For the frequentist methods, we used  $\alpha = .05$  as significance level for the effect size test and  $\alpha = .10$  for the test of publication bias. For the Bayesian methods, we used  $\text{BF} > 10$  as “evidence threshold” for effect size testing and  $\text{BF} > 3$  for the test of publication bias. For estimation with RoBMA and BMA we used the model-averaged posterior mean as the point estimate. As before we do not recommend interpreting Bayes factors in terms of these thresholds; we used them here solely to facilitate the comparison to frequentist methods.

## Results

Our results indicate that RoBMA outperforms other meta-analytic methods in most conditions. Averaging across all conditions RoBMA usually had the lowest root mean squared error (RMSE) and bias (cf.

Figure 7). RoBMA performs best in 65% of conditions when ranked according to RMSE (Figure 8a) and in 36% of conditions when ranked according to bias (Figure 8b). Results were computed conditional on method convergence. *P*-curve converged in 95.5% of effect size estimations and tests, trim and fill in 99.8% of effect size estimations. All other methods converged in 100% of cases.

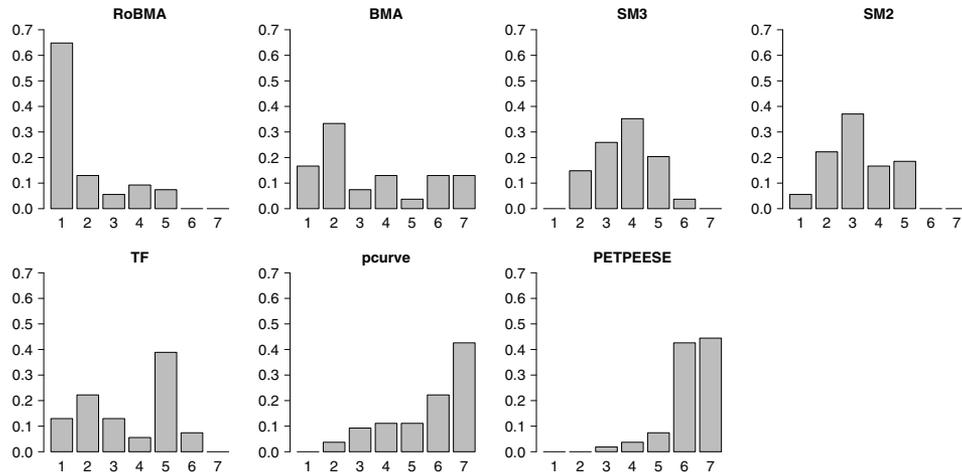
While presenting all individual conditions is beyond the scope of this paper (they can be found at <https://osf.io/buk8g/>), we show the results from three conditions that we find of particular interest to psychologists in Figure 9 because they seem to cover a range of values typical for meta-analysis in psychology (e.g., Stanley et al., 2018). These are  $K = 30$ ,  $\mu = 0$ ,  $\tau = 0$ ,  $\omega = (.2, .5, 1)$ ;  $K = 30$ ,  $\mu = .2$ ,  $\tau = .2$ ,  $\omega = (1, 1, 1)$ ; and  $K = 30$ ,  $\mu = .2$ ,  $\tau = .2$ ,  $\omega = (.2, .5, 1)$ .

As can be seen in Panel A, RoBMA estimates the pooled mean usually precise, although sometimes with slight underestimation due to model-averaging involving the null hypotheses (A2) or a slight overestimation due to the involvement of models assuming no bias (A3). For the effect size test (panel B) RoBMA has high power to detect a null effect (B1) and few false positives. Panels B2 and B3 show that power to detect a true effect is also satisfactory. The test for publication bias in Panel C shows that only the test for excess significance seems to have higher power (C1 and C3), however, that is mitigated by a false positive rate of 30% under no publication bias (C2). RoBMA, on the other hand, shows few false positives as well as high power to detect publication bias. RoBMA’s test for the absence of publication bias has lower power (C1), however, the other simulation conditions (<https://osf.io/buk8g/>) show that the power increases with more primary studies. In addition, the conditions available at <https://osf.io/buk8g/> show that RoBMA sometimes overestimates or underestimates when there are only 10 primary studies in the meta-analysis, although performance is still good compared to other methods. In addition, Appendix C shows that RoBMA leads to substantial overestimation under strong *p*-hacking. Although this limitation is shared with most other methods, it should be noted that Vevea and Hedges (1995) selection models perform better than RoBMA under strong *p*-hacking.

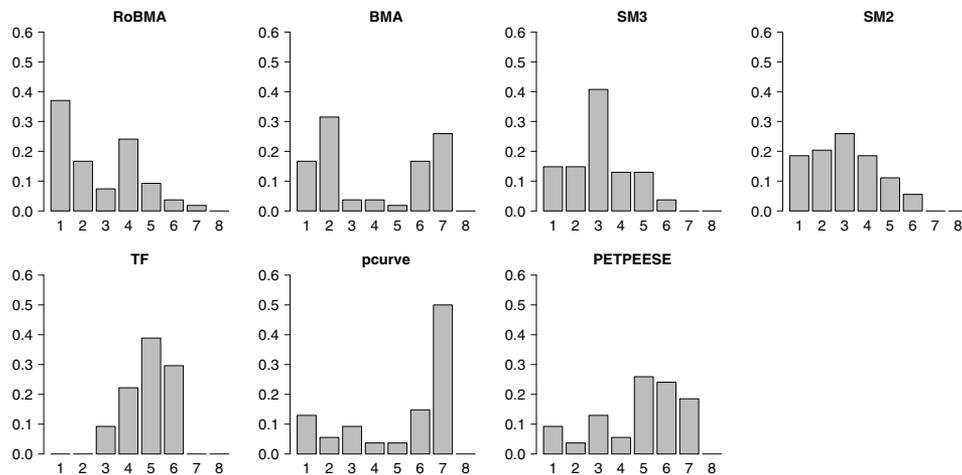
<sup>10</sup> Carter et al. (2019) used trim and fill with the default *L0* estimator that does not allow to test for publication bias.

**Figure 8**  
*Ranked RMSE and Bias of the Population Mean Estimates of Different Methods Across Conditions*

(a) *RMSE*



(b) *Bias*



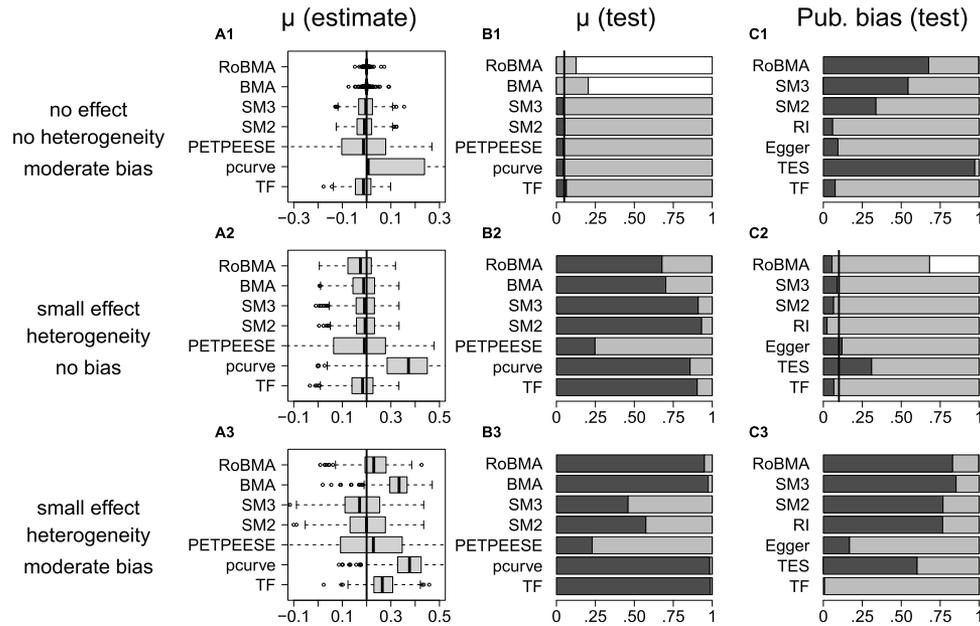
*Note.* SM3 = three-step selection model; SM2 = two-step selection model; TF = trim and fill; RoBMA = resulting robust Bayesian meta-analysis; PETPEESE = precision-effect test and precision-effect estimate with standard error; RMSE = root mean squared error. A rank of 1 means that a method performed best in a simulation condition while a rank of 7 indicates that a method performed worst. The y-axis indicates in which share of simulations a given method achieved a rank.

To summarize, RoBMA outperformed the other methods in most conditions examined here. In some conditions, RoBMA has lower power than other methods; however, these methods suffer from an inflated false positive rate in other conditions. In addition, RoBMA has a tendency to underestimate effect size whenever the alternative is true (due to the inclusion of the null models) and overestimate effect size under publication bias (due to the inclusion of the models assuming no publication bias). However, as can be seen in the online materials, these biases are slight and disappear under larger sample sizes, when almost all weight is given to the correct models. Finally, RoBMA fares noticeably worse than [Vevea and Hedges \(1995\)](#) selection model under strong  $p$ -hacking.

## Concluding Comments

In this article, we introduced a robust Bayesian meta-analysis that model-averages over selection models as well as fixed and random effects models. By applying a set of twelve models simultaneously our method respects the underlying uncertainty when deciding between different meta-analytical models and is comparatively robust to model misspecification. RoBMA also performs well in different simulation conditions and correctly finds support for the absence of publication bias in the Many Labs 2 example. Besides this ability to quantify the evidence for absence of publication bias, the Bayesian approach also allows updating evidence

**Figure 9**  
Results for Selected Scenarios With 30 Primary Studies



*Note.* For testing black lines indicate evidence for the alternative, white indicates evidence for the null (only applicable for Bayesian methods), and gray undecided. BMA = Bayesian meta-analysis; SM3 = selection model with three-steps; SM2 = selection model with two-steps; TF = trim and fill; RI = replicability index; TES = test for excess significance; RoBMA = resulting robust Bayesian meta-analysis; PETPEESE = precision-effect test and precision-effect estimate with standard error.

sequentially as studies accumulate, addressing recent concerns about accumulation bias (ter Schure & Grünwald, 2019).

As are all models, our model is based on assumptions. First, the test statistics follow a (weighted) normal distribution around the true effect size with additive heterogeneity. Second, we make assumptions about the publication process, specifically, that selection is based on *p*-values and that studies with smaller *p*-values are more likely to be published. Other methods, such as Egger regression (Egger et al., 1997) are based on different assumptions of the publication process. Therefore, these methods were disadvantaged in our simulation in which we operationalized publication bias based on significance. However, our model is supported by empirical findings regarding the publication process and *p*-value distributions (e.g., Masicampo & Lalande, 2012) and outperforms other methods on the real data example of Many Labs 2.

We are currently investigating the possibility of adding more models to the RoBMA ensemble such as different weight functions (Citkovicz & Vevea, 2017; Iyengar & Greenhouse, 1988; Patil & Taillie, 1989), “Copas” selection models (Copas, 1999; Copas & Li, 1997; Copas & Shi, 2001), or PET-PEESE models (Stanley & Doucouliagos, 2014). While these additions are beyond the scope of the current article, the RoBMA methodology is sufficiently flexible to include any plausible model for publication bias and unify different competing approaches under a single model-averaging umbrella.

As a general methodological point, we acknowledge that some researchers believe that testing null hypotheses is not meaningful or interesting; instead, these researchers wish to focus on estimation (e.g., Cumming, 2013, 2014; Gelman et al., 2014; Kruschke, 2015). This

philosophy can be accommodated within the framework of RoBMA. To do so, researchers can simply remove the null models (i.e., the model assuming no effect, the fixed-effects model, and the model assuming no publication bias) and then proceed with estimating RoBMA as usual (for example, see the vignettes for the R package on <https://cran.r-project.org/web/packages/RoBMA/>). Also, the R package allows researchers to retrieve the conditional parameter estimates. In other words, after fitting RoBMA—including the null models for testing—researchers may decide to engage in estimation using only those models that assume the effect of interest to be present.

To conclude, this work offers applied researchers a new, conceptually straightforward method to conduct meta-analysis. Instead of basing conclusions on a single model, our method is based on keeping all models in play, with data determining model importance according to predictive success. The simulations and the example suggest that RoBMA is a promising new method in the toolbox of various approaches to test and adjust for publication bias in meta-analysis.

## References

- Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., Rothstein, H. R., & Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: A meta-analytic review. *Psychological Bulletin, 136*(2), 151–173. <https://doi.org/10.1037/a0018251>
- Bai, R., Lin, L., Boland, M. R., & Chen, Y. (2020). A robust Bayesian Copas selection model for quantifying and correcting publication bias. *arXiv*. <https://arxiv.org/abs/2005.02930>

- Bartoš, F., & Maier, M. (2019). All mighty power? Clarifying the Relationship between Power and False Positives. <https://doi.org/10.31234/osf.io/ev29a>
- Bartoš, F., & Maier, M. (2020). RoBMA: An R package for robust Bayesian meta-analyses (R package version 1.0.0). <https://cran.r-project.org/web/packages/RoBMA/>
- Bartoš, F., Maier, M., & Wagenmakers, E.-J. (2020). *Adjusting for publication bias in JASP – Selection models and robust Bayesian meta-analysis* [Manuscript submitted for publication]. <https://doi.org/10.31234/osf.io/75bqn>
- Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle* (2nd ed.). Institute of Mathematical Statistics.
- Borenstein, M. (2019). *Common mistakes in meta-analysis and how to avoid them*. Biostat, Incorporated.
- Card, N. A. (2015). *Applied meta-analysis for social science research*. Guilford Press Publications.
- 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, Article 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>
- Chambers, C. D. (2013). Registered reports: A new publishing initiative at cortex. *Cortex*, 49(3), 609–610. <https://doi.org/10.1016/j.cortex.2012.12.016>
- Chambers, C. D. (2019). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.
- Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex*, 66, A1–A2. <https://doi.org/10.1016/j.cortex.2015.03.022>
- Citkowitz, M., & Vevea, J. L. (2017). A parsimonious weight function for modeling publication bias. *Psychological Methods*, 22(1), 28. <https://doi.org/10.1037/met0000119>
- Coburn, K. M., Vevea, J. L., & Coburn, M. K. M. (2019). Package ‘weightr’ (R package version 2.0.2). <https://cran.r-project.org/web/packages/weightr/weightr.pdf>
- 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/10.1177/0962282020101000402>
- Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29. <https://doi.org/10.1177/0956797613504966>
- Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, 71(9), 1–25. <https://doi.org/10.18637/jss.v071.i09>
- 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(2), 313–329. <https://doi.org/10.1214/16-STS599>
- Ferguson, C. J., & Kilburn, J. (2010). Much ado about nothing: The misestimation and overinterpretation of violent video game effects in Eastern and Western nations: Comment on Anderson et al. (2010). *Psychological Bulletin*, 136(2), 174–178. <https://doi.org/10.1037/a0018566>
- Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 63(3), 665–694. <https://doi.org/10.1348/000711010X502733>
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, 116(2), 439–453. <https://doi.org/10.1037/a0015251>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.
- Givens, G., Smith, D. D., & Tweedie, R. L. (1995). Estimating and adjusting for publication bias using data augmentation in Bayesian meta-analysis (tech. rep.). <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.7777>
- Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2020). A primer on Bayesian model-averaged meta-analysis. <https://doi.org/10.31234/osf.io/97qup>
- Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81, 80–97. <https://doi.org/10.1016/j.jmp.2017.09.005>
- Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10), 1–29. <https://doi.org/10.18637/jss.v092.i10>
- 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., Hoogeveen, S., Berkhout, S., Gronau, Q. F., & Wagenmakers, E.-J. (2020). A Bayesian multiverse analysis of Many Labs 4: Quantifying the evidence against mortality salience. <https://doi.org/10.31234/osf.io/cb9er>
- Heck, D., Gronau, Q., & Wagenmakers, E. (2017). metaBMA: Bayesian model averaging for random-and fixed-effects meta-analysis (R package). <https://cran.r-project.org/web/packages/metaBMA>
- Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, 7(2), 246–255. <https://doi.org/10.1214/ss/1177011364>
- Hilgard, J., Engelhardt, C. R., & Rouder, J. N. (2017). Overstated evidence for short-term effects of violent games on affect and behavior: A reanalysis of Anderson et al. (2010). *Psychological Bulletin*, 143(7), 757–774. <https://doi.org/10.1037/bul0000074>
- 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.
- Hoogeveen, S., Wagenmakers, E.-J., Kay, A. C., & Elk, M. V. (in press). Compensatory control and religious beliefs: A registered replication report across two countries. *Comprehensive Results in Social Psychology*. <https://doi.org/10.1080/23743603.2019.1684821>
- Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4(3), 245–253. <https://doi.org/10.1177/1740774507079441>
- Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 3(1), 109–117.
- JASP Team. (2020). JASP (Version 0.12) [Computer software]. <https://jasp-stats.org/>
- Jeffreys, H. (1938). The comparison of series of measures on different hypotheses concerning the standard errors. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 167(930), 367–384.
- Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press.

- 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>
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahnik, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., . . . Nosek, B. A. (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>
- Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. (2nd ed.). Academic Press/Elsevier.
- 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., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., . . . Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, 146(5), 451–479. <https://doi.org/10.1037/bul0000220>
- Larose, D. T., & Dey, D. K. (1998). Modeling publication bias using weighted distributions in a Bayesian framework. *Computational Statistics & Data Analysis*, 26(3), 279–302. [https://doi.org/10.1016/S0167-9473\(97\)00039-X](https://doi.org/10.1016/S0167-9473(97)00039-X)
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- 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>
- Mavridis, D., Sutton, A., Cipriani, A., & Salanti, G. (2013). A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. *Statistics in Medicine*, 32(1), 51–66. <https://doi.org/10.1002/sim.5494>
- 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>
- Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, 36(1), 97–131. <https://doi.org/10.1007/BF00485695>
- Patil, G., & Taillie, C. (1989). Probing encountered data, meta analysis and weighted distribution methods. In Y. Dodge (Ed.), *Statistical data analysis and inference* (pp. 317–345). Elsevier. <https://doi.org/10.1016/B978-0-444-88029-1.50035-6>
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.
- 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>
- Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in interpretation of levels of significance. *Psychological Reports*, 15(2), 570–570. <https://doi.org/10.2466/pr0.1964.15.2.570>
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis*. Wiley.
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21(2), 301–308. <https://doi.org/10.3758/s13423-014-0595-4>
- Rouder, J. N., & Morey, R. D. (2019). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, 73(2), 186–190. <https://doi.org/10.1080/00031305.2017.1341334>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237. <https://doi.org/10.3758/PBR.16.2.225>
- Rufibach, K. (2015). Package 'selectMeta' (R package). <https://cran.r-project.org/web/packages/selectMeta/index.html>
- Scheel, A. M., Schijen, M., & Lakens, D. (2020). An excess of positive results: Comparing the standard psychology literature with Registered Reports. <https://doi.org/10.31234/osf.io/p6e9c>
- Scheibehenne, B., Gronau, Q. F., Jamil, T., & Wagenmakers, E.-J. (2017). Fixed or random? A resolution through model-averaging. Reply to Carlsson, Schimmack, Williams, and Burkner. *Psychological Science*, 28(11), 1698–1701. <https://doi.org/10.1177/0956797617724426>
- Schimmack, U. D. (2020). *The replicability index is the most powerful tool to detect publication bias in meta-analyses*. <https://replicationindex.com/2020/01/01/the-replicability-index-is-the-most-powerful-tool-to-detect-publication-bias-in-meta-analyses/>
- Silliman, N. P. (1997). Hierarchical selection models with applications in meta-analysis. *Journal of the American Statistical Association*, 92(439), 926–936. <https://doi.org/10.1080/01621459.1997.10474047>
- 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), 534–547. <https://doi.org/10.1037/a0033242>
- Smith, D. D., Givens, G. H., & Tweedie, R. L. (2001). Adjustment for publication bias and quality bias in Bayesian meta-analysis. *Biostatistics Basel*, 4, 277–304.
- Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, 144(12), 1325–1346. <https://doi.org/10.1037/bul0000169>
- 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>
- ter Schure, J., & Grünwald, P. (2019). Accumulation bias in meta-analysis: the need to consider time in error control. *F1000Research*, 8, Article 962. <https://doi.org/10.12688/f1000research.19375.1>
- Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, 22(13), 2113–2126. <https://doi.org/10.1002/sim.1461>
- van Aert, R. C., & Mulder, J. (2020). Bayesian hypothesis testing and estimation under the marginalized random-effects meta-analysis model. <https://doi.org/10.31234/osf.io/ktc94>
- 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), 4. <https://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>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>
- Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, 25(3), 169–176. <https://doi.org/10.1177/0963721416643289>
- Wicherts, J. M. (2017). The weak spots in contemporary science (and how to fix them). *Animals*, 7(12), 90–119. <https://doi.org/10.1177/0963721416643289>
- Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, 42(249), 369–390.

(Appendices follow)

## Appendix A

### Specifications of RoBMA

In the previous version of our article and the RoBMA package (Bartoš & Maier, 2020; Versions < 1.2), we followed a description of the Bayesian version of selection models outlined in Equations 2 and 3 of Larose and Dey (1998). However, after the selection process has filtered out the observed effects that are relatively small, the true random effects will no longer be normally distributed. Especially under high heterogeneity, the Larose and Dey assumption may distort the model fit and unduly lower the marginal likelihood. Therefore, motivated by the frequentist specification of selection models (Vevea & Hedges, 1995) we adapted our model such that it does not estimate the true effects directly but rather marginalizes them out of the likelihood. A similar approach to a Bayesian random effects meta-analysis was recently described as a “marginalized random-effects meta-analysis” (van Aert & Mulder, 2020).

A Bayesian selection model with a three-step weight function, assuming a standard normal prior distributions for the meta-analytic effect size  $\mu$ , inverse-gamma distribution for between-study heterogeneity  $\tau$ , and a cumulative sum of Dirichlet distribution for the weights  $\omega$ , with  $y_k$  and  $\sigma_k$  denoting the observed effect sizes and standard errors for the  $k$  in 1:K studies, can be denoted as,

$$\begin{aligned} \mu &\sim \text{Normal}(0, 1) \\ \tau &\sim \text{Inverse-gamma}(1, 0.15) \\ \omega &\sim \text{Cumulative-Dirichlet}(1, 1, 1) \\ y_k &\sim \text{Weighted-normal}(\mu, \tau^2 + \sigma_k^2, \omega). \end{aligned} \quad (3)$$

*Weighted-normal* stands for a likelihood function of a weighted normal distribution, with mean  $\mu$ , variance  $\tau^2 + \sigma_k^2$ , weights  $\omega$ , and a cumulative probability function of a standard normal distribution  $\Phi$ ,

$$\begin{aligned} &\text{Weighted-normal}(y_k | \mu, \tau^2 + \sigma_k^2, \omega) \\ &= \frac{\text{Normal}(y_k | \mu, \tau^2 + \sigma_k^2) \times w(\omega, p_k)}{\int \text{Normal}(x | \mu, \tau^2 + \sigma_k^2) \times w(\omega, (1 - \Phi(x/\sigma_k)) * 2) dx}, \end{aligned} \quad (4)$$

where the weights  $\omega$  are assigned based on the two-sided  $p$ -values,  $p_k$ , through the weight function  $w$ ,

$$w(\omega, p_k) = \begin{cases} \omega_1, & \text{if } p_k > 0.10 \\ \omega_2, & \text{if } 0.05 < p_k \leq .10 \\ 1, & \text{if } p_k \leq .05 \end{cases} \quad (5)$$

The remaining models can be obtained either by restricting  $\mu = 0$  for models assuming no effect size, restricting  $\tau = 0$  for fixed effect models, by changing the weight functions  $w$  (and the prior on  $\omega$ ) to use only two steps for the two-steps weight function, or by using normal distribution for the observed effect sizes for the models assuming no publication bias.

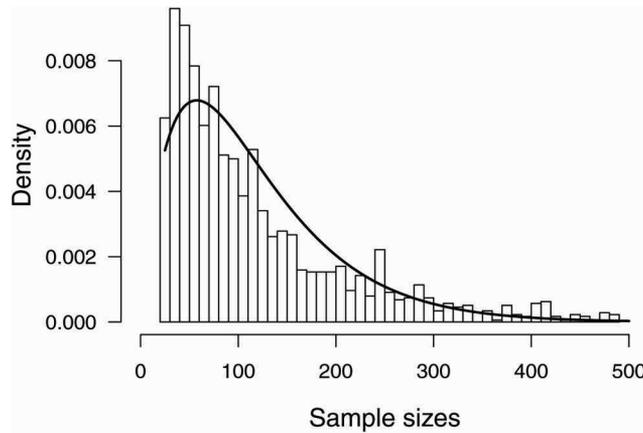
*(Appendices continue)*

## Appendix B

### Sample Sizes

Figure B1 shows sample sizes extracted from empirical meta-analyses in Stanley et al. (2018). The data can be found at <https://osf.io/2vfjy/files/>.

**Figure B1**  
*Distribution of Sample Sizes Used for the Simulation*



*Note.* The histogram is overlaid with Negative binomial distribution with shape  $r = 2$  and scale  $\theta = 58$ , truncated at 25 and 500 which was used for generating sample sizes for the simulation.

## Appendix C

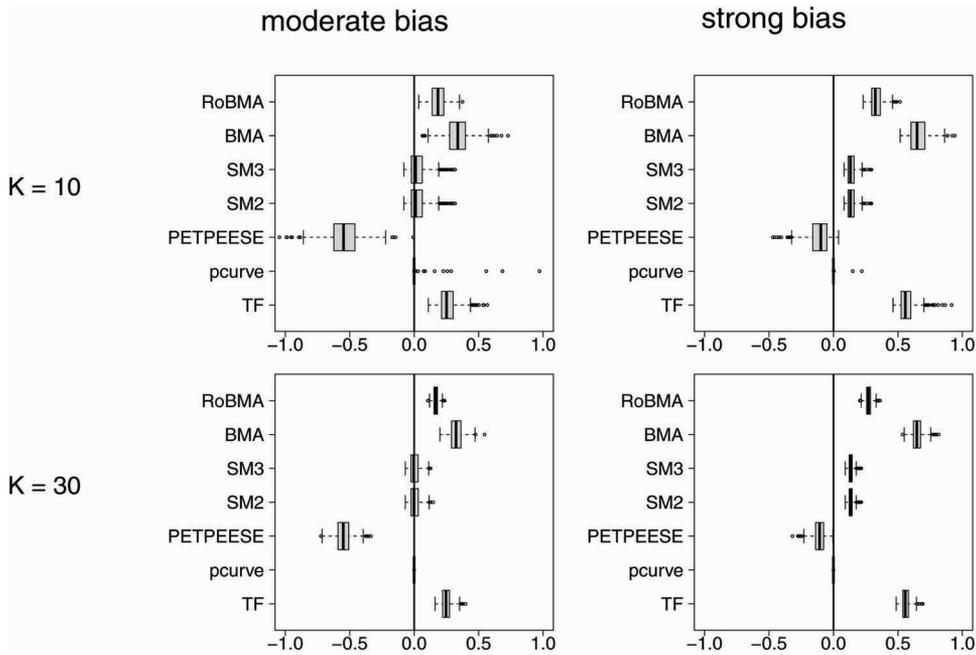
### *P*-Hacking Simulation

We operationalized strong *p*-hacking by starting with two groups of 10 participants, testing the difference after adding one additional participant to a condition, and stopping either upon reaching significance or collecting 100 participants overall. Then, the  $K = (10, 30)$  studies are presented as being in the expected direction and published under either moderate or severe publication bias. Trim and fill did not converge in .35%

of cases, other methods always converged. The simulations for intense *p*-hacking show that RoBMA overestimates the effect size by about .2. This is less than many other methods but more than frequentist selection models (Figure C1). Because *p*-hacking is common in psychological literature (e.g., Simmons et al., 2011) this shows that effect sizes might still be overestimated even after adjusting for publication bias.

*(Appendices continue)*

**Figure C1**  
*Effect Size Estimation Under Strong p-Hacking*



*Note.* BMA = Bayesian meta-analysis; SM2 = two-parameter selection model; SM3 = three-parameter selection model; TF = trim and fill; RoBMA = resulting robust Bayesian meta-analysis; PETPEESE = precision-effect test and precision-effect estimate with standard error.

Received July 8, 2020  
 Revision received January 22, 2021  
 Accepted March 19, 2021 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.