Adjusting for Publication Bias in JASP and R: Selection Models, PET-PEESE, and Robust Bayesian Meta-Analysis

Bartoš, F.; Maier, M.; Quintana, D.S.; Wagenmakers, E.-J.

DOI
10.31234/osf.io/75bqn
10.1177/25152459221109259

Publication date
2022

Document Version
Final published version

Published in
Advances in Methods and Practices in Psychological Science

License
CC BY-NC

Link to publication

Citation for published version (APA):
Bartoš, F., Maier, M., Quintana, D. S., & Wagenmakers, E-J. (2022). 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*, *5*(3). https://doi.org/10.31234/osf.io/75bqn, https://doi.org/10.1177/25152459221109259

Tutorial

Advances in Methods and Practices in Psychological Science
July-September 2022, Vol. 5, No. 3, pp. 1–19
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/25152459221109259
www.psychologicalscience.org/AMPPS

Adjusting for Publication Bias in JASP and R: Selection Models, PET-PEESE, and Robust Bayesian Meta-Analysis

František Bartoš[1,2], Maximilian Maier[1,3], Daniel S. Quintana[4,5,6,7], and Eric-Jan Wagenmakers[1]

[1]Department of Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands; [2]Institute of Computer Science, Czech Academy of Sciences, Prague, Czech Republic; [3]Department of Experimental Psychology, University College London, London, England; [4]Department of Psychology, University of Oslo, Oslo, Norway; [5]NevSom, Department of Rare Disorders, Oslo University Hospital, Oslo, Norway; [6]Norwegian Centre for Mental Disorders Research (NORMENT), University of Oslo, Oslo, Norway; and [7]KG Jebsen Centre for Neurodevelopmental Disorders, University of Oslo, Oslo, Norway

Abstract

Meta-analyses are essential for cumulative science, but their validity can be compromised by publication bias. To mitigate the impact of publication bias, one may apply publication-bias-adjustment techniques such as precision-effect test and precision-effect estimate with standard errors (PET-PEESE) and selection models. These methods, implemented in JASP and R, allow researchers without programming experience to conduct state-of-the-art publication-bias-adjusted meta-analysis. In this tutorial, we demonstrate how to conduct a publication-bias-adjusted meta-analysis in JASP and R and interpret the results. First, we explain two frequentist bias-correction methods: PET-PEESE and selection models. Second, we introduce robust Bayesian meta-analysis, a Bayesian approach that simultaneously considers both PET-PEESE and selection models. We illustrate the methodology on an example data set, provide an instructional video (https://bit.ly/pubbias) and an R-markdown script (https://osf.io/uhaew/), and discuss the interpretation of the results. Finally, we include concrete guidance on reporting the meta-analytic results in an academic article.

Keywords

selection models, PET-PEESE, robust Bayesian meta-analysis, model averaging, publication bias

Received 11/16/21; Revision accepted 4/6/22

Meta-analyses are a powerful tool for evidence synthesis. After a large body of literature has accumulated, researchers may want to conduct meta-analysis to assess the overall evidence for a claim. This might be because they wish to estimate the size of the effect more precisely or because they are interested in testing whether there is even an aggregate nonzero effect in this line of investigation.

However, these meta-analytic inferences can be frustrated by publication bias—the preferential publishing of statistically significant studies. This bias leads to an overestimation of effect sizes when evidence across a set of primary studies is accumulated (Kvarven et al., 2020; Rosenthal & Gaito, 1964). Some researchers have claimed that most research findings might never be published but instead languish in researchers' file drawers (e.g., Ioannidis, 2005; Rosenthal, 1979). Even if the true extent of publication bias were less severe than these researchers have suggested, it would remain a formidable threat to the validity of meta-analyses (Borenstein et al., 2009). Indeed, there have been cases in which

Corresponding Author:
František Bartoš, University of Amsterdam, Department of Psychological Methods, Amsterdam, The Netherlands
Email: fbartos96@gmail.com

Table 1. Summary of Publication-Bias-Adjustment Methods and Implementations

Method	R package/function	JASP implementation	Description[a]
Trim and fill	metafor/trimfill()	(Classical) Meta-analysis	Iteratively imputes studies until achieving a symmetric funnel plot
WAAP-WLS	base R/lm()[b]	WAAP-WLS	Iteratively removes studies with insufficient power to detect the meta-analytic effect size
PET-PEESE	base R/lm()	PET-PEESE	See description in the text
Selection models	weightr/weightfunct() metafor/selmodel() copas/copas() RobustBayesianCopas /RobustBayesianCopas()	Selection models	See description in the text
RoBMA	RoBMA/RoBMA()	Robust Bayesian meta-analysis	See description in the text
p-curve[c]			Estimates a fixed-effect version of a selection model using only statistically significant studies
p-uniform	puniform/puniform()		Estimates a fixed-effect version of a selection model using only statistically significant studies

[a] For an accessible overview and simulation studies, see Carter et al. (2019) and Hong and Reed (2020).
[b] The WAAP-WLS (a hybrid of weighted average of the adequately powered studies and weighted least squares) can be implemented using the a sequence of two lm() functions. The first lm() function estimates the unadjusted meta-analytic effect-size estimate using a weighted least square regression and the second lm() function reestimates the weighted least squares regression only with studies that have sufficient power for finding the unadjusted effect-size estimate.
[c] At the time of writing, CRAN did not feature an R package that implements *p*-curve. A web application from the original authors can be found at http://p-curve.com/.

entire paradigms were possibly based on spurious results, caused in part by publication bias (e.g., Bartoš et al., 2022; Carter & McCullough, 2014; Haaf et al., 2020; Klein et al., 2019; Maier et al., 2022).

In this tutorial, we first introduce two frequentist methods to adjust for publication bias in meta-analysis: precision-effect test and precision-effect estimate with standard errors (PET-PEESE) and selection models. First, PET-PEESE is a meta-analytic estimator that adjusts for the correlation between effect sizes and standard errors. Second, selection models form a set of meta-analytic estimators that correct for different publication probabilities across different *p*-value intervals (for other methods and their implementation, see Table 1).

Extensive simulation studies have shown that each of these methods often come to different conclusions depending on the data-generating process (Carter et al., 2019; Hong & Reed, 2020; McShane et al., 2016). A usual recommendation to accommodate the differences between methods is to apply multiple methods simultaneously (e.g., Carter et al., 2019; Hong & Reed, 2020; McShane et al., 2016). This can be done by fitting different methods and subjectively comparing their results. However, it is unclear how to combine the estimates across different methods or what to conclude if some methods find evidence for publication bias and others do not. The substantial differences between the estimates of different methods (see e.g., Meta Explore app by Carter et al., 2019; https://tellmi.psy.lmu.de/felix/metaExplorer/) make it difficult to derive robust conclusions from publication-bias-adjusted meta-analysis. In addition, researchers may unwittingly succumb to the temptation of “cherry-picking” a method that does not show publication bias in their specific setting.

Here, we outline a more formal way to combine inferences from different methods using Bayesian model averaging (Bartoš, Gronau, et al., 2021; Carter & McCullough, 2018; Gronau et al., 2021; Hinne et al., 2020; Hoeting et al., 1999). Bayesian model averaging is a technique that allows researchers to specify different models simultaneously and agnostically lets the data guide inferences using different models proportional to how well they predict the data. We combine PET-PEESE, selection models, and naive fixed- or random-effects meta-analysis in a model-averaging framework called *robust Bayesian meta-analysis* (RoBMA). We also implemented RoBMA in JASP (JASP Team, 2021; Ly et al., 2021), which is a free and open-source statistical-software package that uses a graphical user interface. In this tutorial, we explain how to use RoBMA in R and JASP.

In the next sections, we first briefly introduce the example data set, a meta-analysis on acculturation mismatch (Lui, 2015). Second, we provide an accessible explanation of PET-PEESE and selection models. Third, we introduce RoBMA, which combines selection models and PET-PEESE under one model-averaging umbrella. All of these methods have not previously been implemented in graphical-user-interface software, which has limited their accessibility to researchers without programming experience. Therefore, we provide guidance on using these methods in both R and JASP. We show how to apply these methods and interpret the results and provide an example report of a result section (Appendix A). We further accompany the tutorial with an R-markdown file (https://osf.io/uhaew/) and recorded tutorial videos (https://bit.ly/pubbias) to facilitate the application of the implemented methods. Detailed documentation describing options of the JASP analyses is accessible via the blue "i" icon in the analysis input headings.

Running Example: Acculturation Mismatch and Intergenerational Cultural Conflict

Lui (2015) studied how acculturation mismatch (i.e., the contrast between the collectivist cultures of Asian and Latin immigrant groups and the individualistic culture of the United States) correlates with intergenerational cultural conflict by meta-analyzing 18 independent studies that correlated acculturation mismatch with intergenerational cultural conflict. A standard random-effects reanalysis calculated with a restricted maximum likelihood estimator and using Fisher z values transformed from correlation coefficients (for more information about effect-size transformations, see Appendix C) indicates a significant relationship between acculturation mismatch on increased intergenerational cultural conflict, $\rho = 0.250$, 95% confidence interval (CI) = [0.172, 0.336], $p < .001$.[1] According to Cochran's Q test for residual heterogeneity, the true outcomes appear to be heterogeneous, Q(17) = 73.58, $p < .001$, $\tau^2 = 0.02$, $I^2 = 77.8\%$. Lui visually inspected a funnel plot and concluded that "the funnel plot . . . revealed a symmetrical 'funnel' shape, supporting that sample size bias [i.e., small study effects] was unlikely for this meta-analysis" (p. 430). Furthermore, Lui conducted a moderator analysis using publication status and found no evidence for a systematic difference in means between published and unpublished studies.

The data set for the following analyses can be downloaded from the OSF repository at https://osf.io/mgu7v/. The first part of the R-markdown file explains how to set up the R environment and packages, load the data set, and perform the effect-size transformation required for meta-analysis (handled automatically in JASP).

PET-PEESE

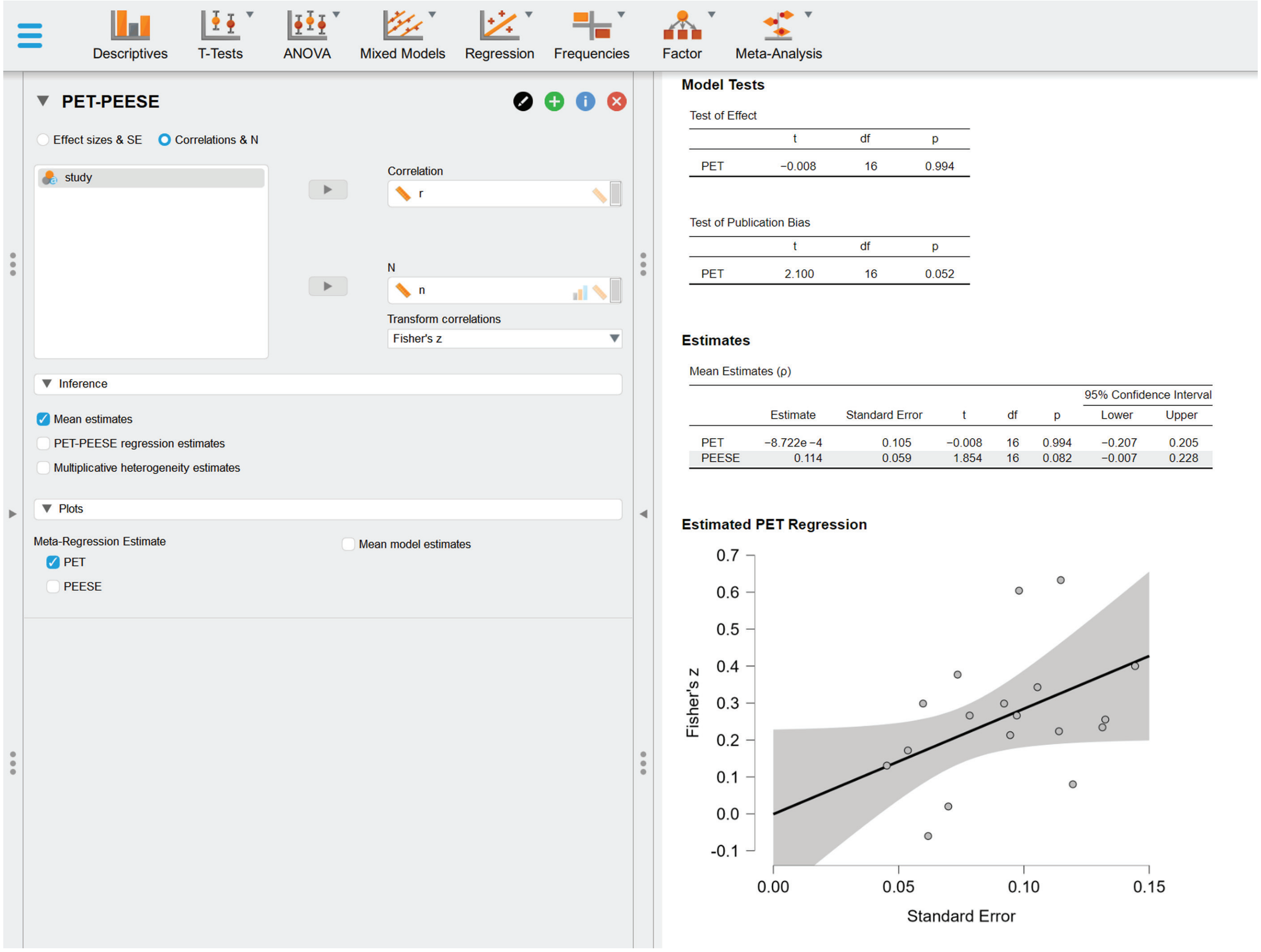
Theoretical background

PET-PEESE is a publication-bias-adjustment method that corrects for the correlation between effect size and standard errors or effect sizes and standard errors squared (Stanley & Doucouliagos, 2014). It can be considered one type of a broader class of funnel-plot-based methods that correct according to the relationship between effect sizes and standard errors (e.g., Duval & Tweedie, 2000; Egger et al., 1997). Because the standard error of standardized effect sizes depends on the sample size, the term "small-study effects" is often used to refer to the overestimation of the meta-analytic mean effect size that is due to less precise studies.

The general idea behind PET-PEESE, and the other funnel-plot-based methods, is that the effect sizes and standard errors ought to be unrelated in the absence of publication bias—information about the standard error of any given study should not inform about the effect-size estimate of the study. Publication bias can introduce a relationship between the effect sizes and standard errors; studies with large sample sizes will usually get published, whereas small studies will be published only if they reach statistical significance. Therefore, the presence of publication bias often results in a relative increase of imprecise studies with inflated effect-size estimates.

However, there might be explanations for a relationship between effect sizes and standard errors other than publication bias (Lau et al., 2006). For example, in the case of a heterogeneous population of effect sizes, researchers might conduct power analyses and target smaller effects with larger studies.

PET-PEESE corrects for the effect-size inflation by using a two-step procedure. In the first step, the PET model, specifying a weighted least square regression predicting the effect sizes with standard errors, is estimated and used to test for the presence of the effect with $\alpha = .10$ as recommended by Stanley and Doucouliagos (2014), Stanley (2017), and Renkewitz and Keiner (2019). If the PET effect-size estimate is significant, the PEESE model, specifying a weighted least square regression predicting the effect sizes with standard errors squared, is used for publication-bias adjustment because it provides a better effect-size approximation in the presence of an effect. If the PET effect-size estimate is not significant, the PET model and its effect-size estimate is used (Stanley, 2017).[2] PET-PEESE has been shown to provide appropriate corrections for publication bias in applied examples (Carter & McCullough, 2014; Kvarven et al., 2020). Moreover, simulation studies indicate low bias, although the variance can sometimes be considerable (Carter et al., 2019). In other words, the effect-size



PET-PEESE

○ Effect sizes & SE ● Correlations & N

study

Correlation: r

N: n

Transform correlations: Fisher's z

Inference

☑ Mean estimates

□ PET-PEESE regression estimates

□ Multiplicative heterogeneity estimates

Plots

Meta-Regression Estimate

☑ PET

□ PEESE

□ Mean model estimates

Model Tests

Test of Effect

	t	df	p
PET	−0.008	16	0.994

Test of Publication Bias

	t	df	p
PET	2.100	16	0.052

Estimates

Mean Estimates (ρ)

						95% Confidence Interval	
	Estimate	Standard Error	t	df	p	Lower	Upper
PET	−8.722e −4	0.105	−0.008	16	0.994	−0.207	0.205
PEESE	0.114	0.059	1.854	16	0.082	−0.007	0.228

Estimated PET Regression

Fig. 1. Results from Lui (2015) using the precision-effect test and precision-effect estimate with standard errors (PET-PEESE) analysis in JASP. Screenshot from JASP graphical user interface when we analyzed the data of Lui (2015). The analysis settings are specified in the left panel (click the blue "i" icon for a description of the controls) and the associated output is shown in the right panel. The shown output concerns (1) a test of effect size based on precision-effect test (PET), (2) a test of publication bias based on PET, (3) effect-size estimates from PET and PEESE, and (4) estimated PET regression model visualizing the relationship between standard errors and effect sizes.

estimates from PET-PEESE are close to the true effect size on average; however, in any particular case, the estimates can be imprecise (i.e., accompanied by large standard errors, e.g., when the number of studies is small or when all studies measure the effect with similar precision).

Application to the running example

The first part of the video (at the 5-min 30-s mark) shows how to perform the PET-PEESE analysis in JASP (Fig. 1). The corresponding analysis with R is outlined in the second part of the R-markdown file.

To interpret the results, we first focused on the test of effect size based on precision-effect test (PET) model under the "Test of Effect" table in the upper right part of the Figure 1. We found that the test of effect size is not significant, so we proceeded to interpret the effect-size estimate on the basis of the PET model under the "Estimates" table in the middle right part of Figure 1. We found that the adjusted mean-effect-size estimate is practically zero, $\rho = 0.000$, 95% CI = [−0.207, 0.205]. Note that the estimate provided by PET-PEESE is substantially wider than that of the naive random-effects meta-analysis because the model incorporating publication bias is more complex.

We can further visualize the PET metaregression estimate of the relationship between the effect sizes and standard errors displayed in the bottom right part of Figure 1. The figure illustrates the relationship between standard errors (x-axis) and effect sizes (y-axis). The adjusted estimate then corresponds to the intercept with the y-axis. Figure 1 is similar to a funnel plot except that the funnel plot visualizes the effect sizes on the x-axis and the standard errors on the y-axis. Showing the standard errors on the x-axis, as here, highlights the PET-PEESE signature of publication bias (i.e., that less precise studies show larger effect sizes).

Selection Models

Theoretical background

Selection models are publication-bias-adjustment methods that use weighted likelihood to account for studies that are missing because of publication bias. Selection models are well established among statisticians (e.g., Iyengar & Greenhouse, 1988; Larose & Dey, 1998; Vevea & Hedges, 1995) and can accommodate realistic assumptions regarding publication bias and heterogeneity (e.g., the chances of publication depend on reported p values; Citkowicz & Vevea, 2017).

Selection models offer multiple ways of defining the relationship between p values and the relative publication probabilities via the weight function. Parameters of the weight function are estimated simultaneously with the rest of the model, which allows selection models to correct for the missing studies. Here, we focus on the step-weight functions that specify distinct p-value intervals, each governed by a different relative publication probability (for information about other implementations of selection models, see Box 1). The step-weight-function selection models are the most popular, arguably because of their simplicity, accessibility, and good performance across multiple simulation studies (e.g., Carter et al., 2019; Hong & Reed, 2020; Maier, Bartoš, & Wagenmakers, 2022; McShane et al., 2016).

To apply selection models with step-weight functions, researchers specify p-value intervals with different publication probabilities, for example, "statistically significant" p values ($p < .05$) versus nonsignificant p values ($p > .05$). Selection models typically use maximum likelihood to obtain a publication-bias-adjusted pooled-effect-size estimate by accounting for the relative publication probabilities in each interval (called "weights") and using the weighted-likelihood function. Selection models can accommodate effect-size heterogeneity by extending random-effects models (Maier et al., 2021; McShane et al., 2016; Rothstein et al., 2005; Vevea & Hedges, 1995, pp. 145–174).

Step-weight-function selection models can be specified flexibly in several ways. First, researchers can decide between one-sided and two-sided selection. One-sided selection means that only significant effects in the expected direction are more likely to be published. Commonly, significant positive effects are more likely to be published, although in some cases, significant negative effect sizes might be more likely to be published as well. Researchers can specify the direction of selection flexibly. Two-sided selection means that the probability of publication does not depend on the direction of the effect; in other words, positive and negative effects have the same probability of being published given that they fall in the same p-value interval.

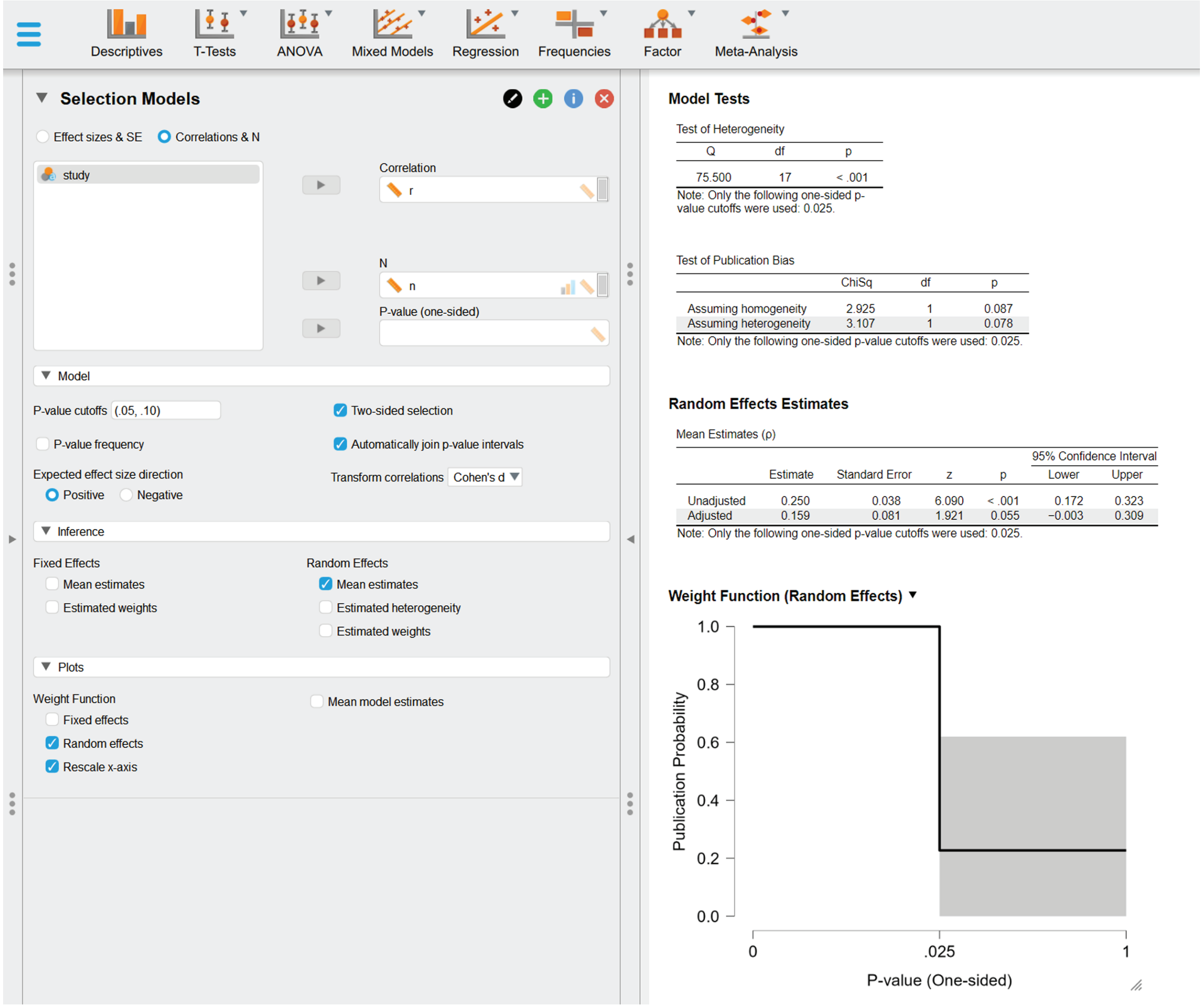
Second, researchers may also specify different intervals for different publication probabilities. For example, to account for the fact that marginally significant results ($.05 < p < .10$) are potentially more likely to be published than nonsignificant results, researchers could specify this as a third interval. Note that when the observed effect is in the predicted direction, a marginally significant result using a two-sided test is significant using a one-sided test. Therefore, a two-sided selection process with different publication probabilities for significant versus "marginally significant" findings accommodates a one-sided selection process with publication probabilities depending on whether the p value is statistically significant.

Box 1. Selection Models and Weight Functions

Throughout the article, we exclusively focus on selection models specified with a step-weight function that is estimated from data. This type of selection models allows researchers to adjust for publication bias operating on p-value thresholds, with the relative publication probabilities estimated simultaneously with the meta-analytic model.

However, there exist many other types of selection models and different use-cases. First, selection models offer a wide variety of weight functions that can be associated with p values, standard errors, or additional variables (for more details. see Citkowicz & Vevea, 2017; Iyengar & Greenhouse, 1988; McShane et al., 2016; Patil & Taillie, 1989; Preston et al., 2004). Second, selection models can be used with prespecified weight functions to perform a sensitivity analysis under different assumptions about the degree of publication bias (e.g., Mathur & VanderWeele, 2020; Vevea & Woods, 2005).

In R, many of the approaches above are implemented via the selmodel() function in the *metafor* package (Viechtbauer, 2010). Use ?metafor::selmodel in R for detailed documentation and examples.



Selection Models

○ Effect sizes & SE ● Correlations & N

study

Correlation: r

N: n

P-value (one-sided):

▼ Model

P-value cutoffs (.05, .10)

☑ Two-sided selection

□ P-value frequency

☑ Automatically join p-value intervals

Expected effect size direction: ● Positive ○ Negative

Transform correlations: Cohen's d

▼ Inference

Fixed Effects
□ Mean estimates
□ Estimated weights

Random Effects
☑ Mean estimates
□ Estimated heterogeneity
□ Estimated weights

▼ Plots

Weight Function
□ Fixed effects
☑ Random effects
☑ Rescale x-axis

□ Mean model estimates

Model Tests

Test of Heterogeneity

Q	df	p
75.500	17	< .001

Note: Only the following one-sided p-value cutoffs were used: 0.025.

Test of Publication Bias

	ChiSq	df	p
Assuming homogeneity	2.925	1	0.087
Assuming heterogeneity	3.107	1	0.078

Note: Only the following one-sided p-value cutoffs were used: 0.025.

Random Effects Estimates

Mean Estimates (ρ)

					95% Confidence Interval	
	Estimate	Standard Error	z	p	Lower	Upper
Unadjusted	0.250	0.038	6.090	< .001	0.172	0.323
Adjusted	0.159	0.081	1.921	0.055	−0.003	0.309

Note: Only the following one-sided p-value cutoffs were used: 0.025.

Weight Function (Random Effects)

Fig. 2. Results from Lui (2015) using selection models analysis in JASP. Screenshot from the JASP graphical user interface when analyzing the data of Lui. The analysis settings are specified in the left panel (use the blue "i" icon for description of the controls), and the associated output is shown in the right panel. The shown output concerns (1) a test of heterogeneity, (2) a test of publication bias, and (3) adjusted and unadjusted effect-size estimates for the random-effects models.

Application to the running example

The middle section of the video (at the 7-min mark) shows how to perform the selection-model analysis in JASP (Fig. 2), which uses the *weightr* R package (Coburn & Vevea, 2019). The corresponding analysis with R is then outlined in the third part of the R-markdown file.

To interpret the results, we first focused on the test of heterogeneity under the "Test of Heterogeneity" table in the upper right part of Figure 2.[3] We found that the test of heterogeneity was significant, so we proceeded to interpret the test for publication bias assuming heterogeneity in the "Test of Publication Bias" table, the second table on the right side of Figure 2. We found that the test for publication bias assuming heterogeneity is significant only when using $\alpha = .10$, as advocated by Renkewitz and Keiner (2019), $\chi^2(1) = 3.11$, $p = .078$. That led us to interpret the adjusted random-effects mean estimate under the "Random Effect Estimates" table, the third table on the right side of Figure 2. We found a nonsignificant adjusted mean effect-size estimate, $\rho = 0.159$, 95% CI = [−0.003, 0.309], that is considerably larger than the effect-size estimate obtained by PET-PEESE ($\rho = 0.000$, 95% CI = [−0.207, 0.205]). Note again the wider CIs compared with naive random-effects meta-analysis.

The above procedure involved an initial test for heterogeneity and an initial test for publication bias; on the basis of the outcomes of these tests, we then applied the random-effects selection model. However, some researchers have argued that random-effects models are to be preferred over fixed-effects models under almost all circumstances (Borenstein et al., 2010). Furthermore, the test for publication bias is often underpowered (Rothstein et al., 2005), and the presence of publication bias can greatly affect the heterogeneity estimate and its test (Augusteijn et al., 2019; Jackson, 2006). Therefore, one may argue that it is prudent to use the adjusted effect-size estimate from the random-effects selection model regardless of the tests for heterogeneity and publication bias (which would not change the result in our case).

We can further visualize the estimated-weight function on the basis of the random-effects model displayed.[4] The bottom right part of Figure 2 shows the estimated publication probabilities (y-axis) for the different p-value intervals (x-axis). The x-axis is rescaled to show equal distance between p-value cut points. This rescaling facilitates readability when the p-value cut points are defined to be relatively close. The first p-value interval $(0, 0.025)$ corresponds to statistically significant studies. Here, the weight function is fixed at the reference value of 1, which means that all these studies are being published. The second p-value interval $(0.025, 1)$ corresponds to statistically nonsignificant studies (with two-sided p values). The corresponding weight-function point estimate is $\omega_{0.025,1} = 0.305$ (black solid line), 95% CI = [0.000, 0.733] (gray area). This means that the statistically nonsignificant studies are less than one third as likely to get published than the statistically significant studies.

Limitations of PET-PEESE and Selection Models

Although the PET-PEESE and selection models provide powerful adjustment in several situations, the frequentist methods outlined above have several shortcomings.

The first limitation is that frequentist Neyman-Pearson point-null hypothesis significance tests (NHSTs) are based on binary accept/reject decisions.[5] When the number of primary studies is small, the methods might have insufficient power, compromising the reliability of the accept/reject decisions (cf. Robinson, 2019). Insufficient power is a considerable problem for the test of publication bias. From a frequentist point of view, the act of not rejecting the point-null hypothesis does not imply that there is evidence in its favor.[6] A single frequentist significance test against a point-null hypothesis cannot distinguish between absence of evidence (i.e., the data are uninformative) or evidence of absence (i.e., the data support the null hypothesis; Keysers et al., 2020; Wagenmakers et al., 2016). This problem was highlighted in the selection-models example—it was unclear whether nonsignificance at the .05 level indicates evidence of absence or absence of evidence regarding publication bias. A closely related limitation is that selection models cannot be estimated when there are insufficient p values in the specified p-value intervals, which is highly likely when the number of primary studies is relatively small.

The second limitation is accumulation bias (ter Schure & Grünwald, 2019). Consider meta-analyzing k primary studies with a frequentist method. At a later point in time, an additional study $k+1$ becomes available, and researchers would want to add this study to the set and update the analysis. For frequentist hypothesis testing, this introduces the problem of multiple testing. To avoid accumulation bias, the sampling plan would need to be known in advance. However, because researchers usually conduct meta-analyses on available data collected by others, accumulation bias is all but inevitable.

A third limitation is that one needs to decide between different methods. PET-PEESE and selection models will sometimes arrive at different results. Although it is advised to fit multiple adjustment methods that are suitable under the given conditions for sensitivity analysis (Carter et al., 2019; McShane et al., 2016), it is less clear what to conclude if the different methods disagree. Ideally, one would want to combine different models into a single method that bases the inference on multiple models simultaneously depending on how well they account for the data.

To overcome these limitations, we developed robust Bayesian meta-analysis (RoBMA; Bartoš, Maier, et al., 2021; Maier, Bartoš, & Wagenmakers, 2022), which combines selection models and PET-PEESE using Bayesian model averaging. In the next sections, we explain RoBMA conceptually and show how it alleviates the shortcomings of frequentist selection models. In addition, we illustrate the workings of the JASP and R implementation in practice.

RoBMA

Theoretical background

RoBMA is a meta-analytic framework that uses Bayesian model averaging to adjust for publication bias (Bartoš, Maier, et al., 2021; Maier, Bartoš, & Wagenmakers, 2022). RoBMA allows researchers to simultaneously estimate different models and base the results on a weighted combination of their estimates. The models can be generally divided into three different pairs:

1. models assuming the null hypothesis to be true versus models assuming the alternative hypothesis to be true (i.e., $\mathcal{H}_0$ vs. $\mathcal{H}_1$),
2. models assuming fixed effects versus models assuming random effects (i.e., $\mathcal{H}^f$ vs. $\mathcal{H}^r$),
3. models assuming publication bias and models assuming no publication bias (i.e., $\mathcal{H}^{pb}$ vs. $\mathcal{H}^{\overline{pb}}$).

The models assuming publication bias encompass the different publication-bias adjustments. We specify both the PET-PEESE publication-bias adjustment (however,

Box 2. Prior Distributions (Part I)

A core part of every Bayesian analysis is the specification of appropriate prior distributions. This Box outlines the default and alternative prior distributions. The R package internally transforms the specified prior distributions to the Fisher *z* scale that is used for estimating the models. Users can change the scale for setting the priors in both R and JASP. The suggested alternative prior distributions can be used for a robust Bayesian meta-analysis (RoBMA) sensitivity analysis, that is, an assessment of the degree to which the reported conclusions are robust to alternative specification of the prior distributions.

Prior Distributions on Effect Size

By default, we use a standard normal distribution on the effect size, $\text{Normal}(M = 0, SD = 1)$, that corresponds to a wide range of effect sizes expected in psychology. On the basis of the literature, we suggest the following plausible alternative priors on "Cohen's d" effect size $\delta = \mu/\sigma$:

$\delta \sim$ Cauchy(location = 0, scale = 0.707)—a default prior distribution in Bayes factor testing, appropriate when large effects cannot be ruled out (Morey & Rouder, 2015).

$\delta \sim$ Student$-t_{[0,\infty]}$(location = 0.35, scale = 0.102, *df* = 3)—an informed prior distribution for small- to medium-sized effects, called the "Oosterwijk prior distribution" after the expert from whom the distribution was elicited (Gronau et al., 2020).

$\delta \sim \text{Normal}_{[0,\infty]}(M = 0.30, SD = 0.15)$—another informed prior distribution for small- to medium-sized effects, called the "Vohs prior distribution" after the preregistered prior distribution used in a multiteam replication effort on the ego-depletion effect (Vohs et al., 2021).

Prior Distributions on Heterogeneity

We suggest the $\text{Inverse–Gamma}(\text{shape} = 1, \text{scale} = 0.15)$ empirical prior distribution for the heterogeneity parameter τ of the Cohen's *d* effect size based on heterogeneity estimates recorded from meta-analyses published in *Psychological Bulletin* (van Erp et al., 2017).

instead of selecting either PET or PEESE, we model averaged across both models) and selection-model adjustment. For the selection models, we specify the following weight functions:

1. Two-sided
   - (a) *p*-value cutoffs = .05
   - (b) *p*-value cutoffs = .05 and .10
2. One-sided
   - (a) *p*-value cutoffs = .05
   - (b) *p*-value cutoffs = .025 and .05
   - (c) *p*-value cutoffs = .05 and .50
   - (d) *p*-value cutoffs = .025, .05, and 0.50

Overall, RoBMA contains eight distinct ways of adjusting for publication bias (PET, PEESE, and six weight functions). The complete model ensemble is then constructed as a combination of all possible components, 2 (Effect vs. No Effect) × 2 (Heterogeneity vs. no Heterogeneity) × 9 (Publication Bias [8] vs. No Publication Bias), resulting in 36 different models. For further details see "Appendix A: Model Specifications" in Bartoš, Maier, et al. (2021).

Prior distributions

To complete the specification of RoBMA, we need to specify prior parameter distributions (see Boxes 2 and 3) and set the prior model probabilities. We use the default settings outlined and tested in a simulation study by Bartoš, Maier, et al. (2021). The simulation study verified that the prior specification performs well in terms of mean square error and bias of the estimates as well as the evidence in favor of the null and alternative hypotheses across a range of scenarios considered typical for psychology. Furthermore, the prior specification outperformed a variety of other publication-bias-adjustment methods on real data examples (Bartoš, Maier, et al., 2021).

We split the prior model probabilities equally across the different model pairs. In other words, we assign 50% prior model probability to models that assume the presence of an effect, 50% prior model probability to models that assume the presence of heterogeneity, and 50% prior model probability to models that assume the presence of publication bias. This division of prior model probabilities reflects a position of equipoise and puts the models on an equal footing before the arrival of the

Box 3. Prior Distributions (Part II)

Prior Distributions for Precision-Effect Test and Precision-Effect Estimate With Standard Errors

By default, we suggest half-Cauchy distributions on the PET, $\text{Cauchy}_+(\text{location} = 0, \text{scale} = 1)$, and PEESE, $\text{Cauchy}_+(\text{location} = 0, \text{scale} = 5)$, metaregression coefficients that ensure a positive correlation between the effect sizes and standard errors. We suggest the following alternative priors on the regression coefficient β, obtained from simulations (Bartoš, Maier, et al., 2021, Appendix B):

$\beta_{\text{PET}} \sim \text{Gamma}(\text{shape} = 2.84, \text{rate} = 2.19)$
$\beta_{\text{PEESE}} \sim \text{Gamma}(\text{shape} = 2.32, \text{rate} = 0.86)$

Prior Distributions for Publication Bias Weights

The default prior distribution for publication bias weights is unit cumulative Dirichlet prior distributions. This encodes the intuitive assumption that studies with statistically significant *p* values have higher relative publication probability than studies with marginally significant *p* values, and studies with marginally significant *p* values have a higher relative publication probability than studies with statistically nonsignificant *p* values (for an illustration supporting the assumption with a collection of over 1 million test statistics collected from Medline, see van Zwet & Cator, 2021, Fig. 1). This assumption allows a more efficient use of information about the publication process, which is especially relevant when the number of studies is small such that some *p*-value intervals contain only a few or no studies. We suggest the following alternative priors on the publication weight ω, obtained from simulations (Bartoš, Maier, et al., 2021, Appendix B):

$\omega_{\text{Two-sided}(.05)} \sim \text{CumDirichlet}(2.49, 0.83)$

$\omega_{\text{Two-sided}(.1,.05)} \sim \text{CumDirichlet}(2.88, 0.98, 0.99)$

$\omega_{\text{One-sided}(.05)} \sim \text{CumDirichlet}(2.61, 0.89)$

$\omega_{\text{One-sided}(.05,.025)} \sim \text{CumDirichlet}(2.92, 0.95, 0.75)$

$\omega_{\text{One-sided}(.5,.05)} \sim \text{CumDirichlet}(3.17, 0.80, 0.83)$

$\omega_{\text{One-sided}(.5,.05,.025)} \sim \text{CumDirichlet}(3.24, 1.02, 0.68, 0.66)$

data (e.g., Gronau et al., 2021; Hinne et al., 2020; Jeffreys, 1939; Madigan et al., 1994; Madigan & Raftery, 1994; Raftery, 1995; for alternatives, see Castillo et al., 2015; Scott & Berger, 2006, 2010; Wilson et al., 2010).

However, we point out that these are only the default settings, and researchers can specify different priors if they so desire. For instance, the prior distribution on the effect-size parameter of the null hypothesis can be modified to specify a test against a perinull hypothesis (i.e., the spike can be changed to a narrow "slab"; e.g., Berger & Delampady, 1987; Cornfield, 1966; George & McCulloch, 1993), and the prior distribution on the alternative hypothesis can be changed to be more informed or directional (Bartoš, Gronau, et al., 2021; Gronau et al., 2017, 2020; see Boxes 2 and 3). Prior knowledge can be also incorporated into the prior model probabilities. Researchers interested in effect-size estimation (e.g., McElreath, 2020) may remove models that assume the effect is absent (i.e., assign these models zero prior probability; but see van den Bergh et al., 2021). Other researchers may for theoretical reasons include only random-effects models and assign zero prior probability to fixed-effects models (e.g., Rothstein et al., 2005; but for empirical evidence from medicine showing that fixed-effects models show relatively good predictive performance, see Bartoš, Gronau, et al., 2021). In addition, it is sometimes argued that models based on the correlation between effect sizes and standard errors might find spurious evidence for publication bias, for example, when researchers take into account heterogeneity by studying small effects with larger samples (Lau et al., 2006). To test this possibility, one may omit the PET-PEESE models (i.e., assigning them zero prior probability) and assess the extent to which this affects the overall conclusions. Another reason for omitting some of the models from the ensemble is when they are clearly inappropriate (e.g., PET-PEESE publication-bias-adjustment models require variability in the standard errors/sample sizes of the original studies; Stanley, 2017)—if all conducted studies had the same standard error, the relationship between the standard errors and sample sizes cannot be estimated. Appendix B shows how RoBMA can be adjusted to compare a perinull hypothesis with an informed alternative hypothesis.

Bayesian model averaging

After we specified the prior distributions and estimated the individual models, we updated the individual models' posterior model probabilities using Bayes's rule. In other words, models that predict the data well receive a boost in posterior probability, whereas models that predict the data poorly suffer a decline (Wagenmakers, 2020; Wagenmakers et al., 2016). Comparing only two models, we can describe their relative predictive performance using Bayes's factors (BFs; Etz & Wagenmakers, 2017; Jeffreys, 1961; Kass & Raftery, 1995; Rouder & Morey, 2019; Wrinch & Jeffreys, 1921). The BF equals the change from prior to posterior odds. If both hypotheses are equally likely a priori, the posterior odds equal the BF. This relationship is illustrated in Equation 1 for two models that both assume the presence of heterogeneity and the absence of publication bias; however, one model assumes the presence of the effect, whereas the other assumes its absence:

$$\underbrace{\frac{p(\text{data}\,|\,\mathcal{H}_1^{r\overline{\text{pb}}})}{p(\text{data}\,|\,\mathcal{H}_0^{r\overline{\text{pb}}})}}_{\text{Bayes factor}} = \underbrace{\frac{p(\mathcal{H}_1^{r\overline{\text{pb}}}\,|\,\text{data})}{p(\mathcal{H}_0^{r\overline{\text{pb}}}\,|\,\text{data})}}_{\text{Posterior odds}} \Big/ \underbrace{\frac{p(\mathcal{H}_1^{r\overline{\text{pb}}})}{p(\mathcal{H}_0^{r\overline{\text{pb}}})}}_{\text{Prior odds}}, \qquad (1)$$

where $\mathcal{H}_1^{r\overline{\text{pb}}}$ denotes the random-effects model (superscript "r") assuming presence of the effect (subscript "1") and absence of publication bias (superscript "pb"), whereas $\mathcal{H}_0^{r\overline{\text{pb}}}$ denotes the random-effects model assuming absence of the effect and absence of publication bias.

More than two models can be compared using the "inclusion Bayes factor." This BF allows researchers to quantify the evidence for a meta-analytical effect, the evidence for heterogeneity, and the evidence for publication bias. When we compare the class of models assuming publication bias with the class of models assuming no publication bias, the inclusion BF can be calculated as in Equation 2:

$$\underbrace{\text{BF}^{pb\overline{pb}}}_{\text{Inclusion Bayes factor for publication bias}} =$$

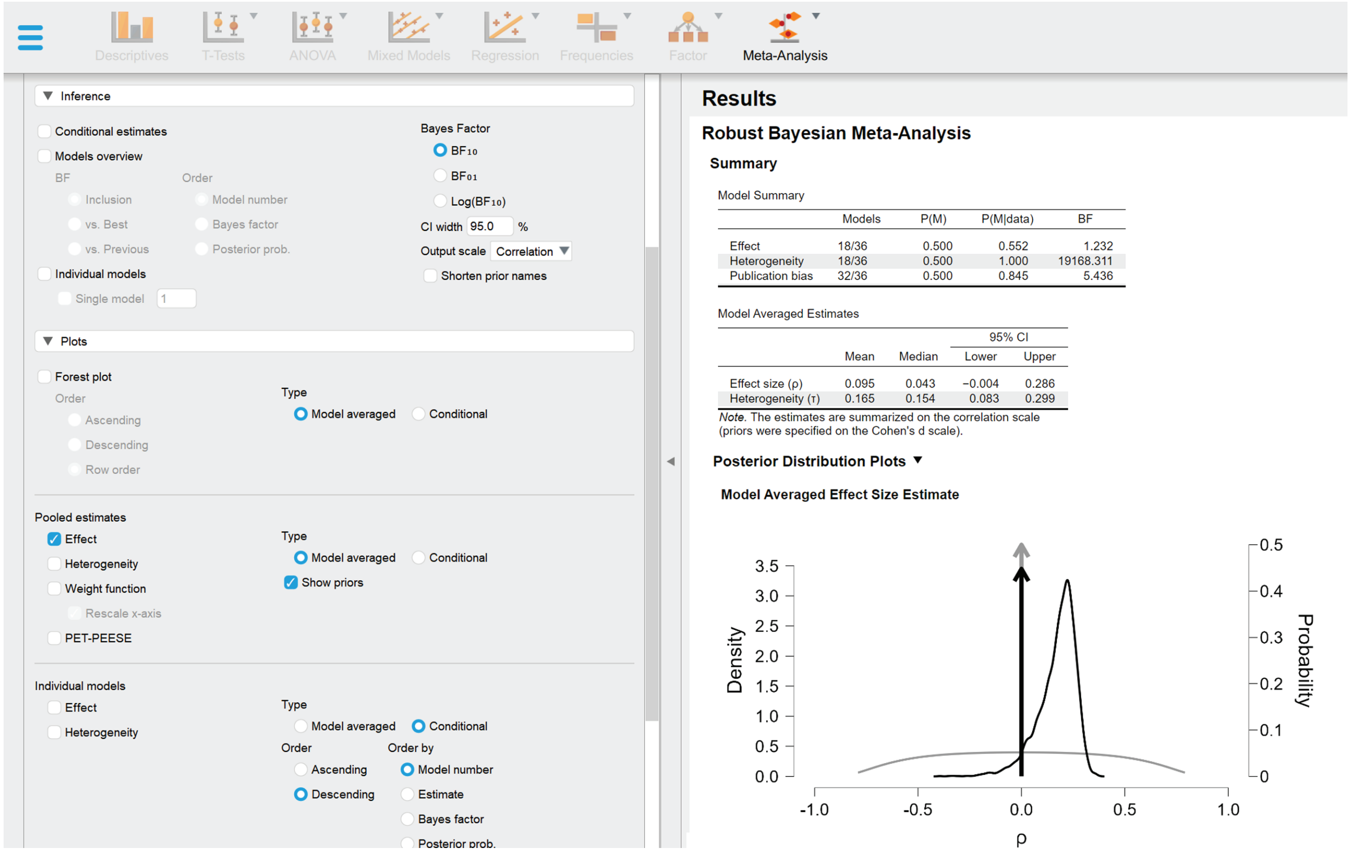
$$\underbrace{\frac{p(\mathcal{H}_1^{fpb}\,|\,\text{data}) + p(\mathcal{H}_1^{rpb}\,|\,\text{data}) + p(\mathcal{H}_0^{fpb}\,|\,\text{data}) + p(\mathcal{H}_0^{rpb}\,|\,\text{data})}{p(\mathcal{H}_1^{f\,\overline{pb}}\,|\,\text{data}) + p(\mathcal{H}_1^{r\,\overline{pb}}\,|\,\text{data}) + p(\mathcal{H}_0^{f\,\overline{pb}}\,|\,\text{data}) + p(\mathcal{H}_0^{r\,\overline{pb}}\,|\,\text{data})}}_{\text{Posterior inclusion odds for publication bias}}$$

$$/\underbrace{\frac{p(\mathcal{H}_1^{fpb}) + p(\mathcal{H}_1^{rpb}) + p(\mathcal{H}_0^{fpb}) + p(\mathcal{H}_0^{rpb})}{p(\mathcal{H}_1^{f\,\overline{pb}}) + p(\mathcal{H}_1^{r\,\overline{pb}}) + p(\mathcal{H}_0^{f\,\overline{pb}}) + p(\mathcal{H}_0^{r\,\overline{pb}})}}_{\text{Prior inclusion odds for publication bias}}. \qquad (2)$$

In other words, the inclusion BF for publication bias is obtained by contrasting the prediction accuracy of all models that assume publication bias to all models that assume no publication bias. The inclusion BF for effect size and heterogeneity can be calculated analogously. One advantage of BFs is that they can distinguish between absence of evidence and evidence of absence. In addition, they can quantify evidence on a continuous scale and can be updated sequentially as studies accumulate, which is not advisable when using a conventional NHST approach. The following rule of thumb can aid the interpretation of BFs: $1 < \text{BF} < 3$ corresponds to weak evidence, $3 < \text{BF} < 10$ corresponds to moderate evidence, and $\text{BF} > 10$ corresponds to strong evidence (e.g., Jeffreys, 1939; Lee & Wagenmakers, 2013).

After updating the models according to their posterior probability, the final effect-size estimate is obtained by Bayesian model averaging (e.g., Hinne et al., 2020; Hoeting et al., 1999). In Bayesian model averaging, the effect size from each individual model is weighted by its posterior probability. Because those models that predicted the data best have the highest posterior probability, the final estimate is based most strongly on the most appropriate models.

Bayesian model averaging is especially relevant in the context of publication-bias adjustment in meta-analyses. Carter et al. (2019) and Hong and Reed (2020) identified the conditions under which particular publication-bias methods perform best (e.g., when heterogeneity is high, selection models generally outperform most other methods); in practical application, however, it usually remains unclear what condition holds for the data set at hand. A similar logic applies to other assumptions—for instance, the degree of variability in the standard errors of the individual studies warranting the use of PET-PEESE publication-bias adjustment is hard to define (i.e., it is unclear what is the degree of variability below which PET-PEESE should no longer be used). Bayesian model averaging applies PET, PEESE, and selection models to the data simultaneously, weighting their relative impact with the extent to which the rival publication-bias methods predicted the observed data. If the variability of standard errors is too low, the PET-PEESE models will predict the data poorly and thus contribute little to inference. In other words, an assumption violation is often equivalent to a poor description of the data by a model, therefore, Bayesian model averaging makes models more robust to misspecification. Finally, whereas a large number of studies will often yield clear evidence for a single model, a low number of studies usually yields evidence that is less conclusive. In such cases, Bayesian model averaging allows the uncertainty about the most appropriate model to be incorporated in a coherent manner, providing optimal estimates that do not suffer from overconfidence (Hoeting et al., 1999).



▼ Inference

☐ Conditional estimates
☐ Models overview
BF: ○ Inclusion ○ vs. Best ○ vs. Previous
Order: ○ Model number ○ Bayes factor ○ Posterior prob.
☐ Individual models
☐ Single model 1

Bayes Factor: ● BF_{10} ○ BF_{01} ○ Log(BF_{10})
CI width 95.0 %
Output scale Correlation ▼
☐ Shorten prior names

▼ Plots

☐ Forest plot
Order: ○ Ascending ○ Descending ○ Row order
Type: ● Model averaged ○ Conditional

Pooled estimates
☑ Effect
☐ Heterogeneity
☐ Weight function
☑ Rescale x-axis
☐ PET-PEESE
Type: ● Model averaged ○ Conditional
☑ Show priors

Individual models
☐ Effect
☐ Heterogeneity
Type: ○ Model averaged ● Conditional
Order: ○ Ascending ● Descending
Order by: ● Model number ○ Estimate ○ Bayes factor ○ Posterior prob.

Results

Robust Bayesian Meta-Analysis

Summary

Model Summary

	Models	P(M)	P(M|data)	BF
Effect	18/36	0.500	0.552	1.232
Heterogeneity	18/36	0.500	1.000	19168.311
Publication bias	32/36	0.500	0.845	5.436

Model Averaged Estimates

			95% CI	
	Mean	Median	Lower	Upper
Effect size (ρ)	0.095	0.043	−0.004	0.286
Heterogeneity (τ)	0.165	0.154	0.083	0.299

Note. The estimates are summarized on the correlation scale (priors were specified on the Cohen's d scale).

Posterior Distribution Plots ▼

Model Averaged Effect Size Estimate

Fig. 3. Results from Lui (2015) using the robust Bayesian meta-analysis (RoBMA) in JASP. Screenshot from the JASP graphical user interface when we analyzed the data of Lui (2015). The analysis settings are specified in the left panel (use the blue "i" icon for description of the controls), and the associated output is shown in the right panel. The shown output displays (1) summary of the model components, (2) model-averaged estimates of the effect size and heterogeneity, and (3) prior (gray) and posterior (black) model-averaged distribution for the effect size estimate. The arrows centered at zero correspond to the point probability mass allocated to the null hypothesis (secondary *y*-axis), and the smooth densities correspond to the distributions under the alternative hypothesis.

RoBMA's benefits

RoBMA overcomes the limitations of frequentist selection models and PET-PEESE in several ways. First, the BF allows researchers to quantify relative evidence for the null hypothesis and thus distinguish between absence of evidence and evidence of absence.

Second, the model averaging obviates the need to select a single model in an all-or-none fashion. Therefore, if there is uncertainty regarding the presence of publication bias, RoBMA can base the inference on both the "normal" models and the publication-bias-adjusted models instead of needing to commit fully to a single model.

Third, the prior distributions allow the selection models to be estimated even in cases with few *p* values in some of the *p*-value intervals, which is a limitation of frequentist selection models. The method will not fail to converge under these conditions. However, especially in this context, it is important to specify the prior distributions carefully and check the robustness of the results to different specifications of the prior distributions. Concretely, we recommend using the distributions in Boxes 2 and 3 in addition to the default priors to check whether the conclusions are robust to the prior choice.

Fourth, BFs allow for sequential updating (Rouder, 2014; Rouder & Morey, 2011; Wagenmakers et al., 2016), meaning that new studies can be added to the set and the analysis can be updated without having to worry about accumulation bias. At every point in time, RoBMA quantifies evidence using the relative predictive performance of the rival models for the observed data.

Application to the running example

The last part of the video (at the 15-min 40-s mark) shows how to perform the robust Bayesian meta-analysis in JASP (Fig. 3), which uses the *RoBMA* R package

Table 2. Summary of Meta-Analytic Estimates for Lui (2015) Based on Different Models

Method	Effect size estimate ρ and 95% CI	Test against no effect
Random effects	0.249 [0.170,0.324]	$z = 6.06, p < .001$
PET-PEESE (PET)	−0.001 [−0.210,0.208]	$t(16) = -0.01, p = .994$
Selection model	0.159 [−0.003,0.309]	$z = 1.92, p = .055$
RoBMA-PSMA	0.095 [−0.004,0.286]	$\text{BF}_{10} = 1.23$

Note: PET = precision-effect test; PEESE = precision-effect estimate with standard errors; RoBMA-PSMA = robust Bayesian meta-analysis specified as in Bartoš, Maier et al. (in press).

(Bartoš & Maier, 2020). The corresponding analysis with R is outlined in the fourth part of the R-markdown file. In contrast to the previous methods, RoBMA is estimated via Markov chain Monte Carlo (MCMC), the convergence of which ought to be checked. Both JASP and R return automatic convergence warnings (further diagnostics can be obtained in the "MCMC Diagnostics" menu in JASP; the R-markdown file contains more details regarding the R version) that prompt the user to adjust the MCMC fitting process (for details, see the JASP or R help files).

To interpret the results, we first focused on model components summary under the "Model Summary" table in the upper right part of the Figure 3. We found absence of evidence for the presence of the effect, $\text{BF}_{10} = 1.23$; extreme evidence for presence of the heterogeneity, $\text{BF}^{\text{rf}} = 19{,}168$; and a moderate evidence for the presence of publication bias, $\text{BF}_{\text{pb}} = 5.44$. The posterior model-averaged estimates are summarized under the "Model Averaged Estimates" table in the middle right part of Figure 3. We found that the model-averaged mean effect-size estimate is between the PET-PEESE and selection models estimates from the previous sections, $\rho = 0.095$, 95% CI = $[-0.004,\ 0.286]$. See Table 2 for a comparison of effect-size estimates and tests against no effect based on the different methods.

Furthermore, the bottom right part of Figure 3 visualizes the model-averaged prior (gray lines) and posterior (black lines) distributions of the effect-size estimate (on the Cohen's *d* scale). The black vertical arrow is slightly lower than the gray vertical arrow, reflecting a slight decrease in probability for models that assume the effect is absent—note that the secondary *y*-axis indexes the prior and posterior probability mass allocated to the null hypothesis. The continuous prior and posterior distributions are associated with models that assume the effect is present.

This example highlights the Bayesian benefit of taking all uncertainty into account. In the frequentist framework, it was unclear whether to adjust for publication bias and how to adjust for publication bias. In contrast, RoBMA does not require an all-or-none decision on the presence of publication bias. Instead, all models are taken into account simultaneously, and the effect-size estimate is based on a weighted average across the various models. The weights are determined according to the support that each model receives from the data. Taking all models into account, we still found the absence of evidence for an effect. In other words, more primary studies are needed to learn about the relationship between intergenerational cultural conflict and acculturation mismatch. When the new studies are conducted, RoBMA allows researchers to continuously update the evidence.

Concluding Comments

In this article, we introduced three approaches to adjust for publication bias in meta-analysis, all implemented in JASP and R. First, we discussed PET-PEESE, a regression-based estimator with low bias that has been shown to perform well on empirical examples. Second, we discussed frequentist selection models, which have been demonstrated to work well even under high heterogeneity. Third, we discussed RoBMA, a Bayesian approach for combining complementary publication-bias-adjustment methods according to how well they describe the data at hand. RoBMA allows researchers to move beyond single-model inference and incorporate model-selection uncertainty into the meta-analytic estimates; in addition, RoBMA allows researchers to conduct multimodel tests for the presence of the effect, for heterogeneity, and for publication bias.

The RoBMA ensemble is highly modifiable; prior parameter distributions may be adjusted to reflect different background knowledge, prior model probabilities can be set to reflect theoretical preferences or expectations (e.g., entire classes of models can be excluded when deemed inappropriate on theoretical grounds), and more generally, researchers who do not wish to engage in model averaging may inspect the parameter estimates and posterior model probabilities for each individual model separately.

In cases in which the data-generating process and type of publication bias are known, the Meta Explorer app (https://tellmi.psy.lmu.de/felix/metaExplorer/) by Carter et al. (2019) can be used to select the most appropriate method in a given situation. In cases in which uncertainty about the data-generating process and the presence or type of publication bias exists, RoBMA allows researchers to combine the adjustment approaches according to their predictive performance for the observed data.

However, we note that RoBMA also has several limitations. First, whereas averaging over a set of models alleviates problems because of model misspecification, the meta-analytic estimate might still suffer from over- or underestimation if none of the models approximate the data-generating process well. Second, whereas the RoBMA's performance was demonstrated across multiple simulation environments and empirical examples, it can lead to overcorrection of the effect-size estimates under moderate and strong questionable research practices, as simulated by Carter et al. (2019). Third, RoBMA has considerably longer fitting time compared with frequentist approaches. However, for educational purposes or when distributing results with colleagues, one can share a .JASP file with models that have already been fitted to illustrate interpretation of the results.

To conclude, the publication-bias-adjusted meta-analyses in JASP allows researchers without programming experience to conduct state-of-the-art, publication-bias-corrected meta-analysis in an intuitive and user-friendly way. We hope that this methodology will improve the inferences researchers make when conducting meta-analysis.

Appendix A

Example report

In the previous sections, we illustrated the various publication-bias-adjustment methods implemented in the JASP "Meta-Analysis" module. Here, we briefly demonstrate how to report the results of PET-PEESE, selection models, and RoBMA using the example of acculturation mismatch and intergenerational cultural conflict (Lui, 2015) that we followed throughout the article. For more general reporting guidelines, see van Doorn et al. (2021). We also emphasize that reporting results is the last step of conducting a meta-analysis; for guides for planning, selecting articles, describing methods, and so on, see, for instance, Borenstein et al. (2009), Higgins et al. (2019), and Quintana (2015). Please also see the Meta-Analysis Reporting Standards and Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for how the results of a meta-analysis need to be reported (Cooper, 2016, p. 287; Moher et al., 2009). We conducted the analysis with the JASP statistical software (JASP Team, 2021; Ly et al., 2021) that relies on *metafor*, *weightr*, and *RoBMA* R packages (Bartoš & Maier, 2020; Coburn & Vevea, 2019; Viechtbauer, 2010).

A random-effects meta-analysis was performed using Fisher *r*-to-*z* transformed correlation coefficients; heterogeneity (i.e., τ^2) was calculated using restricted maximum likelihood estimator. Cochran's Q test for heterogeneity was also be performed, and I^2 was calculated. Fisher's *z* was back-transformed to *r* for reporting the results. Reanalysis confirmed a significant relationship between acculturation mismatch on increased intergenerational cultural conflict, $\rho = 0.250$, 95% CI = [0.172, 0.336], $p < .001$. According to Cochran's Q test for residual heterogeneity, the true outcomes appear to be heterogeneous, Q(17) = 73.58, $p < .001$, $\tau^2 = 0.02$, $I^2 = 77.8\%$.

We first adjusted for publication bias using PET-PEESE. The PET model did not find a significant effect on the $\alpha = .10$ level, $p = .994$. Consequently, we interpreted the effect-size estimate using the PET model, $\rho = 0.000$, 95% CI = [−0.207, 0.205].

Then we proceeded to using selection models. Before data analysis, we decided to use significance level $\alpha = .10$ for publication bias (Renkewitz & Keiner, 2019) and $\alpha = .05$ for heterogeneity and effect sizes. We estimated the two-sided selection models with *p*-value cutoffs set to (.05, .10) and automatically joined *p*-value intervals. The models were estimated using correlations and sample sizes with Cohen's *d* effect-size transformation. The *p*-value intervals were automatically reduced to two intervals separated by a .025 cutoff corresponding to a one-sided *p* value. The test for heterogeneity was significant, $Q(17) = 75.5$, $p < .001$. Therefore, we applied the test for publication bias assuming heterogeneity, which was significant as well, $\chi^2(1) = 3.11$, $p = .078$. Consequently, we interpreted the bias-adjusted effect-size estimate from a random-effects model. The effect size was not statistically significant, $\rho = 0.159$, 95% CI = [−0.003, 0.309], $p = .055$; heterogeneity estimate, τ (on Cohen's *d* scale) = 0.339, 95% CI = [0.042, 0.477]. Altogether, both PET-PEESE and selection models did not find a statistically significant effect, however, they provided notably different adjusted mean effect-size estimates.

Third, we reanalyzed the same data set using robust Bayesian meta-analysis. Before the analysis, we decided to use the default prior settings (i.e., standard normal distribution on effect sizes, inverse gamma distribution with $\alpha = 1$ and $\beta = 0.15$ on heterogeneity, six weight functions and PET-PEESE publication-bias adjustment as specified in Bartoš, Maier, et al., 2021). We set the prior hypothesis probability to 0.50 for the effect size, heterogeneity, and publication bias. The results showed absence of evidence for the presence of the effect, $\text{BF}_{10} = 1.23$; extreme evidence for heterogeneity, $\text{BF}^{\text{rf}} = 19{,}168$; and moderate evidence for the presence of publication bias, $\mathrm{BF}_{\mathrm{pb}} = 5.44$. The resulting model-averaged effect-size estimate was $\rho = 0.095$, 95% CI = [–0.004, 0.286]; heterogeneity estimate, $\tau = 0.165$, 95% CI = [0.083, 0.299]. The MCMC diagnostics were good; all R-hat values were below 1.01, and all effective sample size (ESS) were above 500; the commonly used rule of thumb for a good R-hat is < 1.05, and for ESS, it is > 500 (Gelman & Rubin, 1992; McElreath, 2020). RoBMA extends the findings of PET-PEESE and selection models by finding moderate evidence against the presence of the effect and by providing a model-averaged effect-size estimate that incorporates the uncertainty about the best publication-bias-adjustment model. The resulting JASP file can be found at https://osf.io/rpkhw/.

Finally, we assessed the sensitivity of our conclusions with respect to prior distributions for the publication-bias-adjustment part of RoBMA. We changed prior distributions for the PET-PEESE publication-bias adjustment and the weight functions from the default prior distributions to the specification outlined in Boxes 2 and 3. The model-averaged effect-size estimate, $\rho = 0.079$, 95% CI = [–0.013, 0.280], and the heterogeneity estimate, $\tau = 0.173$, 95% CI = [0.084, 0.310], stayed essentially the same, as did the absence of evidence for the presence of the effect, $\mathrm{BF}_{10} = 0.980$. Furthermore, there was a slight, albeit inconsequential, increase in the evidence in favor of the heterogeneity, $\mathrm{BF}^{\mathrm{rf}} = 22{,}467$, and increase in the evidence in favor of publication bias, $\mathrm{BF}_{\mathrm{pb}} = 7.08$.

Appendix B

Specifying different priors

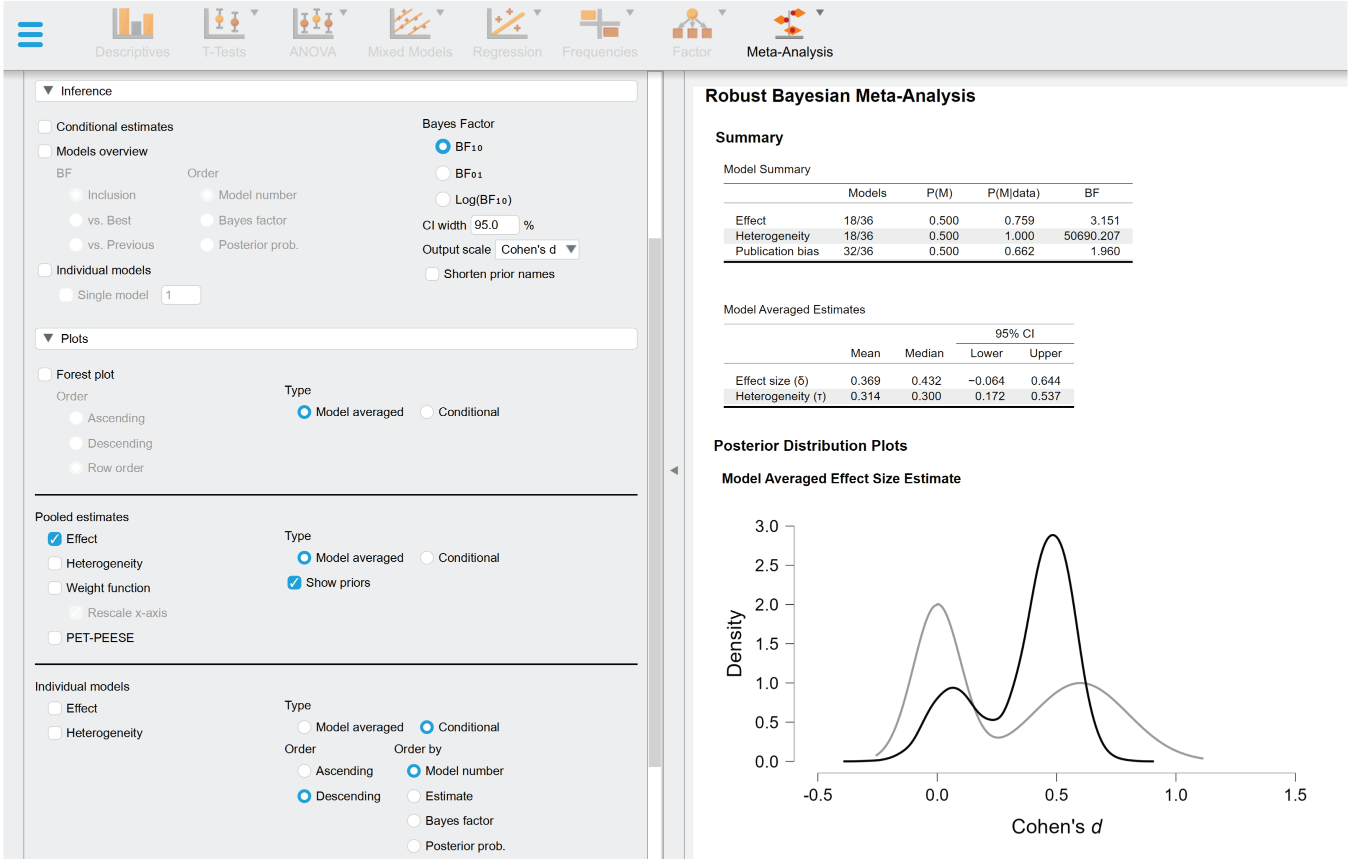
In the previous example, RoBMA revealed compelling evidence against the point-null hypothesis. However, it has been repeatedly argued that point-null hypotheses are not realistic and therefore not meaningful to test (e.g., Gelman & Carlin, 2014; Good, 1967; Meehl, 1978; Orben & Lakens, 2020). RoBMA overcomes this objection by allowing the specification of "perinull" hypotheses, that is, hypotheses with prior distributions tightly centered around an effect size of zero (e.g., Berger & Delampady, 1987; Cornfield, 1966; George & McCulloch, 1993). Adjustments may also be desired for the prior distribution on effect size that is postulated under the alternative hypothesis. By default, the most plausible value for this prior distribution is zero (i.e., under $\mathcal{H}_1$, the prior distribution is centered on zero), and this may not reflect the information at hand (e.g., Gronau et al., 2020). A more diagnostic test requires an "informed prior," one that is centered around a nonzero value for the effect size.

Here, we use the example of acculturation mismatch and intergenerational cultural conflict from the main text (Lui, 2015) and demonstrate how RoBMA allows researchers to specify both a perinull hypothesis and an informed hypothesis and compare their predictive performance for the observed data. First, we specified a perinull hypothesis by assigning effect size a zero-centered normal distribution with a standard deviation .10 on the Cohen's *d* scale; propagated to the correlation scale, this yields a prior distribution with approximately 95% probability mass in the interval $r \in [-0.10, 0.10]$. For a perinull hypothesis, this range may be considered relatively wide. If the goal of the perinull distribution is primarily to counter the objection that "the null hypothesis is never true exactly," a much narrower 95% interval could be specified, such as one ranging from −.01 to .01, for instance. Second, we specified the informed alternative hypothesis by assigning the effect size a normal distribution centered at .60 with standard deviation .20 on the Cohen's *d* scale. Translated to the correlation scale, this results in a prior distribution with most probability on correlations higher than .10, with the prior median at a correlation slightly lower than .30.

The model specification implemented in JASP allows researchers to specify any desired combination of hypotheses using different prior distributions[7] (see the JASP help file accessible under the "i" icon). The specified prior distribution under each hypothesis is then used to generate a combination of all possible models. These models are automatically used to draw inference using the inclusion BF and model averaged to obtain model estimates.

In contrast to the default analysis, we found weak evidence for the presence of the effect, $\mathrm{BF}_{10} = 3.15$; even stronger evidence for presence of the heterogeneity, $\mathrm{BF}^{\mathrm{rf}} = 50{,}690$; and considerably weaker evidence for the presence of publication bias, $\mathrm{BF}_{\mathrm{pb}} = 1.96$. The posterior model-averaged estimates then reflect the higher inclination toward the informed alternative hypothesis with the model-averaged mean effect-size estimate, $\rho = 0.179$, 95% CI = [–0.032, 0.306].

We can, again, visualize the model-averaged prior (gray line) and posterior (black line) distributions of the effect-size estimate (on Cohen's *d* scale) displayed in the bottom right part of Figure B1. The figure now shows a single continuous prior and posterior distribution that model averages across all models. Most of the posterior distribution mass is concentrated at the effect sizes specified under the informed alternative hypothesis.



▼ Inference

□ Conditional estimates

□ Models overview

BF
○ Inclusion
○ vs. Best
○ vs. Previous

Order
○ Model number
○ Bayes factor
○ Posterior prob.

□ Individual models

□ Single model 1

Bayes Factor
● BF_{10}
○ BF_{01}
○ $Log(BF_{10})$

CI width 95.0 %

Output scale Cohen's d ▼

□ Shorten prior names

▼ Plots

□ Forest plot

Order
○ Ascending
○ Descending
○ Row order

Type
● Model averaged ○ Conditional

Pooled estimates

☑ Effect

□ Heterogeneity

□ Weight function

□ Rescale x-axis

□ PET-PEESE

Type
● Model averaged ○ Conditional

☑ Show priors

Individual models

□ Effect

□ Heterogeneity

Type
○ Model averaged ● Conditional

Order
○ Ascending
● Descending

Order by
● Model number
○ Estimate
○ Bayes factor
○ Posterior prob.

Robust Bayesian Meta-Analysis

Summary

Model Summary

	Models	P(M)	P(M|data)	BF
Effect	18/36	0.500	0.759	3.151
Heterogeneity	18/36	0.500	1.000	50690.207
Publication bias	32/36	0.500	0.662	1.960

Model Averaged Estimates

			95% CI	
	Mean	Median	Lower	Upper
Effect size (δ)	0.369	0.432	−0.064	0.644
Heterogeneity (τ)	0.314	0.300	0.172	0.537

Posterior Distribution Plots

Model Averaged Effect Size Estimate

Fig. B1. Comparison between a perinull hypothesis and an informed alternative hypothesis as applied to the data from Lui (2015) using the robust Bayesian meta-analysis (RoBMA) in JASP. Screenshot from the JASP graphical user interface from when we analyzed the data of Lui. The analysis settings are specified in the left panel (use the blue "i" icon for description of the controls), and the associated output is shown in the right panel. The shown output concerns (1) summary of the model components, (2) model-averaged estimates of the effect size and heterogeneity, and (3) prior and posterior model-averaged distribution for the effect size estimate.

Appendix C

Effect-size transformations

There are two main reasons for transforming effect sizes before performing a meta-analysis:

1. Studies with different designs often report different standardized effect-size measures. For example, an experiment comparing mean differences between two groups will report a Cohen's *d*, whereas an observational study assessing the relationship between two continuous variables will report a correlation coefficient. If researchers believe that the different designs are comparable and measure the same underlying effect, they can, under certain assumptions, transform the different effect sizes into a common scale to combine them with a meta-analysis (for more detail, see Borenstein et al., 2009).
2. Certain effect-size measures are better suited for a meta-analysis than others. For example, correlation coefficients are bounded to (–1, 1) range, and standard errors of correlation coefficients, used for weighting in the meta-analysis, are heavily influenced by the effect size itself, $SE(r) \approx (1 - r^2)/\sqrt{(N-1)}$. This is problematic because the likelihood of most meta-analytic models assumes (unbounded) normally distributed standard errors. Furthermore, the inherent relationship between effect sizes and standard errors in many standardized effect-size measures (e.g., correlation coefficients, Cohen's *d*, Hedges's *g*, log odds ratios) might conflict with regression-based publication-bias-adjustment methods (e.g., PET-PEESE) that assume the effect sizes and standard errors are unrelated under the absence of publication bias. Performing the meta-analysis on effect sizes transformed to a different effect-size scale can then mitigate these issues.

When the standard errors ought to be independent of the effect sizes by design (e.g., PET-PEESE), the Fisher's z transformation is an ideal solution because standard errors are dependent only on the sample sizes, $SE(z) \approx 1/\sqrt{N-3}$. Otherwise, Cohen's d is another suitable option because researchers are often familiar with the scale and it has unrestricted range. After the meta-analytic model is estimated, the effect-size estimates can be transformed and interpreted on any of the standardized effect-size scales.

Transparency

Action Editor: Daniel J. Simons
Editor: Daniel J. Simons
Author Contributions

F. Bartoš and M. Maier are joint first authors. All of the authors approved the final manuscript for submission.

Declaration of Conflicting Interests

F. Bartoš, M. Maier, and E.-J. Wagenmakers declare that they are or were involved in the development of the open-source software package JASP (https://jasp-stats.org), a noncommercial, publicly funded effort to make Bayesian and non-Bayesian statistics accessible to a broader group of researchers and students.

Funding

This project was supported in part by a Vici grant (No. 016.Vici.170.083) to E.-J. Wagenmakers.

Open Practices

Open Data: not applicable
Open Materials: https://osf.io/jxf82
Preregistration: not applicable

All materials have been made publicly available via OSF and can be accessed at https://osf.io/jxf82/. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

František Bartoš https://orcid.org/0000-0002-0018-5573
Maximilian Maier https://orcid.org/0000-0002-9873-6096
Daniel S. Quintana https://orcid.org/0000-0003-2876-0004
Eric-Jan Wagenmakers https://orcid.org/0000-0003-1596-1034

Acknowledgments

The data and annotated JASP files that were used in this tutorial are openly available on OSF at http://doi.org/10.17605/OSF.IO/6HF7R.

Notes

1. This result is close to that reported by Lui (2015): $\rho = 0.23$. For the reanalysis, we used the data set as recoded by Stanley et al. (2018), obtained from https://osf.io/2vfyj/files/. The correlation coefficient was calculated by back-transforming the Fisher z summary effect-size estimate of 0.254, 95% CI = [0.170, 0.324]. The Fisher z transformation "unwinds" the correlation coefficients to an unrestricted range and makes the standard errors of the effect-size estimates independent of the effect sizes. Effect sizes measured as correlation coefficients can be transformed into Fisher z as $z = 0.5 \log((1+r)/(1-r))$ with standard error $SE(z) = 1/\sqrt{N-3}$, and the meta-analytic effect-size estimate can be transformed back into correlation coefficient as $r = (\exp(2z)-1)/(1+\exp(2z))$.

2. In the PET model, the regression coefficient of the standard error corresponds to the Egger's test (Egger et al., 1997).

3. Under the output tables, the note "Only the following one-sided p-value cutoffs were used: 0.025" informs readers that some of the specified p-value intervals did not contain enough p values for estimation and were therefore collapsed.

4. Note that the relative publication probabilities might be estimated imprecisely (Hedges & Vevea, 1996).

5. This is not the case for the NHST framework developed by Sir Ronald Fisher, who proposed the p value as a continuous measure of evidence against the point-null hypothesis (e.g., Hubbard & Bayarri, 2003).

6. Frequentist equivalence tests (e.g., Hodges & Lehmann, 1954; Lakens et al., 2018; Schuirmann, 1987) to interval null hypotheses instead of to point-null hypotheses and generally suffer from low power (Linde et al., 2021).

7. For specifying more complex models, see the RoBMA R package manual (Bartoš & Maier, 2020) and the "Fitting Custom Meta-Analytic Eensembles Vignette" (https://fbartos.github.io/RoBMA/articles/CustomEnsembles.html).

References

Augusteijn, H. E., van Aert, R., & van Assen, M. A. (2019). The effect of publication bias on the Q test and assessment of heterogeneity. *Psychological Methods*, *24*(1), 116–134. https://doi.org/10.1037/met0000197

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*, *40*, 6743–6761. https://doi.org/10.1002/sim.9170

Bartoš, F., & Maier, M. (2020). *RoBMA: An R package for robust Bayesian meta-analyses* [R package version 2.1.0]. https://CRAN.R-project.org/package=RoBMA

Bartoš, F., Maier, M., Stanley, T., & Wagenmakers, E.-J. (2022). *Adjusting for publication bias reveals mixed evidence for the impact of cash transfers on subjective well-being and mental health*. PsyArXiv. https://doi.org/10.31234/osf.io/d9vcg

Bartoš, F., Maier, M., Wagenmakers, E.J., Doucouliagos, H., & Stanley, T. D. (2021). Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods*. Advance online publication. https://doi.org/10.1002/jrsm.1594

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*(3), 317–335. https://doi.org/10.1214/ss/1177013238

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

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

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., & McCullough, M. E. (2018). A simple, principled approach to combining evidence from meta-analysis and high-quality replications. *Advances in Methods and Practices in Psychological Science*, *1*(2), 174–185. https://doi.org/10.1177/2515245918756858

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

Castillo, I., Schmidt-Hieber, J., & Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, *43*(5), 1986–2018. https://doi.org/10.1214/15-AOS1334

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

Coburn, K. M., & Vevea, J. L. (2019). *weightr: Estimating weight-function models for publication bias* [R package version 2.0.2]. https://CRAN.R-project.org/package=weightr

Cooper, H. M. (2016). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). Sage.

Cornfield, J. (1966). A Bayesian test of some classical hypotheses—With applications to sequential clinical trials. *Journal of the American Statistical Association*, *61*(315), 577–594. https://doi.org/10.1080/01621459.1966.10480890

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

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.1080/01621459.1993.10476353

Good, I. J. (1967). A Bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society B: Methodological*, *29*(3), 399–418. https://doi.org/10.1111/j.2517-6161.1967.tb00705.x

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/10.1177/25152459211031256

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian t-tests. *The American Statistician*, *74*, 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

Haaf, J. M., Hoogeveen, S., Berkhout, S. W., Gronau, Q. F., & Wagenmakers, E.-J. (2020). *A Bayesian multiverse analysis of Many Labs 4: Quantifying the evidence against mortality salience*. PsyArXiv. https://doi.org/10.31234/osf.io/cb9er

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, *21*(4), 299–332. https://doi.org/10.3102/10769986021004299

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

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

Hodges, J., Jr., & Lehmann, E. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society B: Methodological*, *16*(2), 261–268. https://doi.org/10.1111/j.2517-6161.1954.tb00169.x

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–401. https://doi.org/10.1214/SS%5C%2F1009212519

Hong, S., & Reed, W. R. (2020). Using Monte Carlo experiments to select meta-analytic estimators. *Research Synthesis Methods*, *12*(2), 192–215. https://doi.org/10.1002/jrsm.1467

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. *The American Statistician*, *57*(3), 171–182. https://doi.org/10.1198/0003130031856

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, *3*(1), 109–117. https://doi.org/10.1214/ss/1177013012

Jackson, D. (2006). The implications of publication bias for meta-analysis' other parameter. *Statistics in Medicine*, *25*(17), 2911–2921. https://doi.org/10.1002/sim.2293

JASP Team. (2021). *JASP* (Version 0.15.1) [Computer software]. https://jasp-stats.org/

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

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

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

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J., Croma, R., Gardiner, G., Gosnell, C., Grahe, J., Hall, C., Howard, I., Joy-Gaba, J., Kolb, M., Legg, A. M., . . . Ratliff, K. (2019). *Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement*. PsyArXiv. https://doi.org/10.31234/osf.io/vef2c

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

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

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

Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *BMJ*, *333*(7568), 597–600. https://doi.org/10.1136/bmj.333.7568.597

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

Linde, M., Tendeiro, J. N., Selker, R., Wagenmakers, E.-J., & van Ravenzwaaij, D. (2021). Decisions about equivalence: A comparison of TOST, HDI-ROPE, and the Bayes factor. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000402

Lui, P. P. (2015). Intergenerational cultural conflict, mental health, and educational outcomes among Asian and Latino/a Americans: Qualitative and meta-analytic review. *Psychological Bulletin*, *141*(2), 404–446. https://doi.org/10.1037/a0038449

Ly, A., van den Bergh, D., Bartoš, F., & Wagenmakers, E.-J. (2021). Bayesian inference with JASP. *The ISBA Bulletin*, *28*, 7–15.

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*(428), 1535–1546. https://doi.org/10.1080/01621459.1994.10476894

Madigan, D., Raftery, A. E., York, J. C., Bradshaw, J. M., & Almond, R. G. (1994). Strategies for graphical model selection. In P. Cheeseman & R. W. Oldford (Eds.), *Selecting models from data. Lecture notes in statistics* (Vol, 89, pp. 91–100). Springer. https://doi.org/10.1007/978-1-4612-2660-4_10

Maier, M., Bartoš, F., Stanley, T. D., Shanks, D., Harris, A. J., & Wagenmakers, E. J. (2022). No evidence for nudging after adjusting for publication bias. *PNAS*, *119*(31), Article e2200300119. 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*. Advance online publication. https://doi.org/10.31234/osf.io/u4cns

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), Article e1256. https://doi.org/10.1002/cl2.1256

Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society C: Applied Statistics*, *69*(5), 1091–1119. https://doi.org/10.1111/rssc.12440

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC Press.

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

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & the PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes factors for common designs* [R package version 0.9.12-4.3]. http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor

Orben, A., & Lakens, D. (2020). Crud (re)defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238–247. https://doi.org/10.1177/2515245920917961

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

Preston, C., Ashby, D., & Smyth, R. (2004). Adjusting for publication bias: Modelling the selection process. *Journal of Evaluation in Clinical Practice*, *10*(2), 313–322. https://doi.org/10.1111/j.1365-2753.2003.00457.x

Quintana, D. S. (2015). From pre-registration to publication: A non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in Psychology*, *6*, Article 1549. https://doi.org/10.3389/fpsyg.2015.01549

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–196). Blackwells.

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

Robinson, G. K. (2019). What properties might statistical inferences reasonably be expected to have?—Crisis and resolution in statistical inference. *The American Statistician*, *73*(3), 243–252. https://doi.org/10.1080/00031305.2017.1415971

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

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. (2005). *Publication bias in meta-analysis*. John Wiley & Sons.

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. (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. (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

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. https://doi.org/10.1007/BF01068419

Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, *136*(7), 2144–2162. https://doi.org/10.1016/j.jspi.2005.08.031

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*(5), 2587–2619. https://doi.org/10.1214/10-AOS792

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., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi/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

van den Bergh, D., Haaf, J. M., Ly, A., Rouder, J. N., & Wagenmakers, E.-J. (2021). A cautionary note on estimating effect size. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245921992035

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*(3), 813–826. https://doi.org/10.3758/s13423-020-01798-5

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

van Zwet, E. W., & Cator, E. A. (2021). The significance filter, the winner's curse and the need to shrink. *Statistica Neerlandica*, *75*(4), 437–452. https://doi.org/10.1111/stan.12241

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), 428–443. https://doi.org/10.1037/1082-989X.10.4.428

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

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A., Wagenmakers, E.-J., & Albarracín, D. (2021). A multi–site preregistered paradigmatic test of the ego depletion effect. *Psychological Science*, *32*, 1566–1581. https://doi.org/10.1177/0956797621989733

Wagenmakers, E.-J. (2020). *Bayesian thinking for toddlers*. JASP Publishing.

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

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C., & Schildkraut, J. M. (2010). Bayesian model search and multilevel inference for SNP association studies. *The Annals of Applied Statistics*, *4*(3), 1342–1364. https://doi.org/10.1214/09-AOAS322

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