

# A Bayesian approach to mitigation of publication bias

Maime Guan and Joachim Vandekerckhove<sup>†</sup>  
University of California, Irvine

## Abstract

The reliability of published research findings in psychology has been a topic of rising concern. Publication bias, or treating positive findings differently from negative findings, is a contributing factor to this “crisis of confidence,” in that it likely inflates the number of false positive effects in the literature. We demonstrate a Bayesian model averaging approach that takes into account the possibility of publication bias and allows for a better estimate of true underlying effect size. Accounting for the possibility of bias leads to a more conservative interpretation of published studies as well as meta-analyses. We provide mathematical details of the method and examples.

## Introduction

It is reasonably well established that publication bias—a differential publishing rate between positive and negative results—permeates the scientific literature (Franco, Malhotra, & Simonovits, 2014). Some of the strongest such claims are made by Ioannidis (2005), who famously concluded that “most claimed research findings are false” (p. 696). Decades earlier, Rosenthal (1979) had pointed out the theoretical possibility that all research findings are false, and that file drawers around the world conceal the 95% of studies whose luck of the draw did not lead to publication. The prevalence of publication bias has been a growing concern in psychology, especially in recent years, and is often mentioned as one of the causes of the “crisis of confidence” (Pashler & Wagenmakers, 2012). As Young, Ioannidis, and Al-Ubaydli (2008) point out, “the small proportion of results chosen for publication are unrepresentative of scientists’ repeated samplings of the real world” (p. 1418) It is no wonder, then, that successful replication attempts are scarce (e.g., Francis, 2012a).

The favorable results preferred by systemic publication bias are often defined by statistical significance. Results are said to be ‘significant’ if the probability  $p$  of observing data that are *at least as extreme as the observed* is smaller than some criterion  $\alpha$  if the null hypothesis is true. The conventional criterion (for psychological scientists) in null hypothesis significance testing (NHST),  $\alpha = .05$ , sets a convenient bar for consideration by researchers

---

JV was supported by NSF grant #1230118 from the Methods, Measurements, and Statistics panel and John Templeton Foundation grant #48192. MG was supported by NSF-GRFP grant #DGE-1321846.  
<sup>†</sup>Corresponding author.

and reviewers alike. However, the presence of an arbitrary threshold may contribute to the number of positive findings reported in the literature. Masicampo and Lalande (2012) surveyed the distributions of  $p$ -values in three prominent psychological journals and found that  $p$ -values were conspicuously more common than expected just below  $\alpha$ .

To tackle this important and widespread issue of publication bias, the field can endeavor to prevent future cases, as well as take action to alleviate the current state of psychological science. Our focus is on the latter goal of *mitigation*, by postulating a set of possible bias mechanisms and using these to improve the estimation of effect sizes in the presumed presence of publication bias.

The idea of publication bias *detection* is not new—there exist tests for it in the literature (e.g., the test for “excess significance” based on Ioannidis & Trikalinos, 2007), and these have been applied on numerous occasions (e.g., Francis, 2012a, 2012b, 2012c). Neither are we the first to propose behavioral models of the publication process (Givens, Smith, & Tweedie, 1997; Greenwald, 1975; Hedges, 1992), and there are several recently-developed approaches to deal with publication bias.

One such new method, based on the  $p$ -curve, is due to Simonsohn, Nelson, and Simmons (2014). The  $p$ -curve is the distribution of statistically significant  $p$ -values for a set of independent findings. The exact shape of the  $p$ -curve is used to determine the evidential value of that particular set of findings, or the set’s ability to rule out selective reporting as the sole explanation of those findings. The central intuition is that if the null hypothesis is true,<sup>1</sup>  $p$ -curves are expected to be uniform, but if the null is false, only right-skewed  $p$ -curves with more low than high significant  $p$ -values are diagnostic of evidential values. The  $p$ -curve method assesses if the observed  $p$  distribution is significantly right-skewed. Inference from  $p$ -curve is analogous to null hypothesis significance testing; a right-skewed  $p$ -curve does not imply all studies have evidential value, similar to how significance does not imply all observations were influenced by the experimental manipulation. Therefore,  $p$ -curve can only determine whether selective reporting can be ruled out as an explanation for a set of significant findings and cannot make conclusions about estimates of the true underlying population effect size.

Stanley and Doucouliagos (2013) propose another approach consisting of a set of meta-regression approximations that are designed to be a practical solution to the publication bias issue. This method is developed from Taylor polynomial approximations to the conditional mean of a truncated distribution, because results are selected to be statistically significant in the desirable direction. The observed effects will consequently depend on the population’s true effect, plus another term that reflects selection bias. A critical limitation of this method is that it requires a relatively large number of estimates on the same empirical phenomenon because they are based on regression analysis.

Lastly,  $p$ -uniform (van Assen, van Aert, & Wicherts, in press) is again based on the distribution of  $p$ -values conditional on a certain population effect size  $\mu$ . For example, if the true population effect size is zero, then the conditional  $p$ -value distribution should be close to the uniform distribution. Therefore,  $p$ -uniform tests the null hypothesis by

---

<sup>1</sup>That is, if the effect size is truly zero, and all the ancillary assumptions of the significance test are met. It is important to keep in mind that there are many ways for a null hypothesis to be violated. In the case of a  $t$  test,  $p$  is not uniformly distributed if there is heteroskedasticity or if the conditional distributions are not normal. The  $p$ -curve is also not uniform in the case of compound hypotheses such as a one-tailed  $t$  test.

testing whether the observed conditional  $p$ -value distribution deviates from the uniform. This method performs three tasks simultaneously: (a) the testing of publication bias; (b) effect size estimation, and (c) the testing of the null hypothesis of no effect.  $p$ -uniform only considers studies that report statistically significant results, so those containing non-significant effects are discarded.  $p$ -uniform makes two assumptions: (1) homogeneity of population effect size across all studies and (2) that all studies with statistically significant results are equally likely to be published and included in the meta-analysis. According to the authors,  $p$ -uniform is the first method to possess the aforementioned three qualities, without making sophisticated assumptions or choices.

Our proposed Bayesian mitigation method encompasses these characteristics as well, in addition to being a Bayesian approach to detect and mitigate the effects of publication bias in the field. We believe we are the first to attempt mitigation of publication bias by—as we will demonstrate—averaging over a set of plausible behavioral models. Our interest is foremost in recovering the true underlying effect size as accurately as possible, and not necessarily to decide whether a particular field exhibits more or less publication bias.

### Behavioral models for publication bias

We postulate a set of four generative behavioral models for the publication process. While this set is not exhaustive, we believe it covers a reasonably large range of possible processes. The first model is a *no-bias model* under which all results, regardless of statistical significance, are published. This would be the optimal scenario for the scientific literature, but it is unlikely to hold in all cases. The second model is an *extreme-bias model* in which nonsignificant results are never published. The presence of occasional nonsignificant results in the literature implies that this model cannot hold in all cases either. The third model is one inspired by Greenwald (1975), in which nonsignificant results are published with some constant (but unknown) probability. Finally, the fourth model is one inspired by Givens et al. (1997), in which the publication probability is an exponentially decreasing function of the  $p$ -value if  $p > \alpha$  (i.e., it diminishes as the observed  $p$ -value departs from  $\alpha = .05$ ).<sup>2</sup>

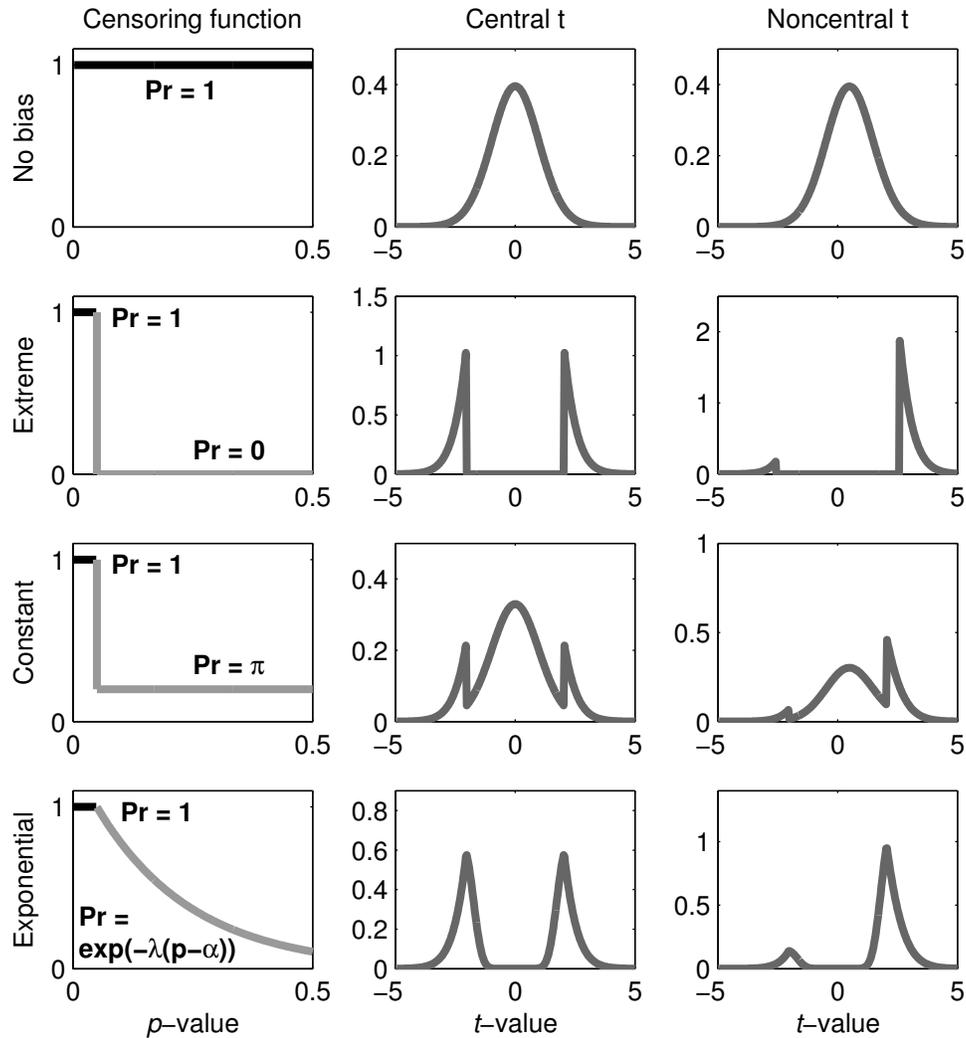
For each of the four models, we consider two possible states of nature: either (a) there exists a true effect of some unknown magnitude; or (b) the true effect size is 0. While the behavioral processes are conveniently described through a censoring function that operates on the  $p$ -value, the state-of-nature component of these models depends on the specific experiment to which our method is applied (i.e., it depends on whether the experiment is comparing two means with a  $t$  or a  $z$  test, or interactions with an  $F$  or a  $\chi^2$  test, etc.). Each of these models has, as a parameter, the size of the experimental effect, and fitting these models to the observed data (i.e., the reported effect size) will yield a new estimate of the effect size conditional on the censoring process.

The general formulation of the likelihood function associated with these models is:

$$p\left(\delta^{obs} \mid \eta^{true}, \mathcal{M}, \Theta\right) \propto p\left(\delta^{obs} \mid \eta^{true}\right) C_{\mathcal{M}}\left(\delta^{obs}, \Theta\right),$$

where  $\delta^{obs}$  is an observed test statistic,  $\eta^{true}$  is the true effect size (which may be 0 for some

<sup>2</sup>There is a fifth possible model, in which very high  $p$ -values are also more likely to be published, because they are erroneously seen as providing evidence for the null. We describe a model with this property in the Discussion but do not add it to our set of behavioral models here.



*Figure 1.* Predictions made by the eight proposed models. Left column: the censoring functions for each of the four biasing processes, with on the horizontal axis the observed  $p$ -value and on the vertical axis the probability of publication  $P_{pub}$ . Middle and right column: the resulting truncated central (center) and noncentral (right)  $t$  distributions after applying the censoring functions. The top row is the no-bias model, where the resulting  $t$  distributions are simply the central and noncentral  $t$  distributions untouched. For the remaining three rows, the  $t$  distributions are truncated to the region of significance or downweighted in the complementary region, as determined by the corresponding biasing functions. The figure illustrates that each behavioral model generates  $t$  distributions with a unique shape.

models), and  $C_{\mathcal{M}}$  is a model-specific censoring function that describes the probability of an observation of size  $\delta^{obs}$  being published, given the parameters  $\Theta$  of the censoring process. Figure 1 summarizes the eight models, and a comprehensive mathematical treatment of the models is given in Appendix A. Models of this form are known in the statistical literature as *selection models* (e.g., Bayarri & DeGroot, 1987, 1991, Iyengar & Greenhouse, 1988).

### Bayesian model averaging

A critical issue is that, while we may define any number of behavioral processes that censor the published literature, we do not know *which* process, if any, was in play for a given report. However, there exist standard statistical tools for dealing with the unknown. The most common such tool is *marginalizing* over (a.k.a., “integrating out”) an unknown:

$$p\left(\eta^{true} \mid \delta^{obs}\right) = \sum p\left(\eta^{true} \mid \mathcal{M}, \delta^{obs}\right) P\left(\mathcal{M} \mid \delta^{obs}\right),$$

where  $p\left(\eta^{true} \mid \mathcal{M}, \delta^{obs}\right)$  is the posterior distribution of the effect size under a censoring process  $\mathcal{M}$  and given the observed test statistic  $\delta^{obs}$ .  $P\left(\mathcal{M} \mid \delta^{obs}\right)$  is the probability that  $\mathcal{M}$  happened given the observed data. Crucially, we treat the identity of the censoring process  $\mathcal{M}$  as merely another unknown and compute the likelihood of each possible value for  $\eta^{true}$  using the probability of each model’s truth as a weight (equivalently, we can compute the posterior distribution of a *mitigated test statistic*  $\delta^{mit}$  as a simple transformation of  $\eta^{true}$ ). Computing those weights—the *posterior probabilities*—for each model relies on the same principle, combined with Bayes’ theorem:  $P\left(\mathcal{M} \mid \delta^{obs}\right) = p\left(\delta^{obs} \mid \mathcal{M}\right) P\left(\mathcal{M}\right) / p\left(\delta^{obs}\right)$ .  $p\left(\delta^{obs}\right)$ , in turn, is obtained through marginalization:  $p\left(\delta^{obs}\right) = \sum p\left(\delta^{obs} \mid \mathcal{M}\right) P\left(\mathcal{M}\right)$ .

Taken together, this yields the *posterior distribution* of  $\delta^{mit}$  given  $\delta^{obs}$ , which we may then use to make inferences about the existence or non-existence of an effect. In particular, we can now quantify the degree to which the observation  $\delta^{obs}$  changes how likely we consider it that the true effect size is 0. The degree of change from prior to posterior information is known as the *Bayes factor*, and integrating over models in this way is commonly referred to as *Bayesian model averaging* (for an introduction, see Hoeting, Madigan, Raftery, & Volinsky, 1999).

Finally, in the case of multiple independent studies  $s = 1, \dots, S$ , and hence a set of multiple observed test statistics  $\left\{\delta_s^{obs}\right\}_{s=1}^S$ , we can compute an aggregated posterior distribution  $p\left(\delta^{mit} \mid \left\{\delta_s^{obs}\right\}_{s=1}^S\right)$ . The details of this computation are in Appendix B.

### Simulation studies

To demonstrate the effects of statistical mitigation, we performed a series of Monte Carlo studies, of which we report four. All studies were based on 500 simulations of each case.

**(1) Single result re-analysis.** In this Monte Carlo study, we simulated a single empirical result regarding some hypothetical phenomenon of a given effect size. In the simulation, the true effect size ranged from 0 (no effect) to 0.3 (a large effect) in increments of 0.05. The result of a one-sided paired  $t$  test (for positive effects) was then either “published”

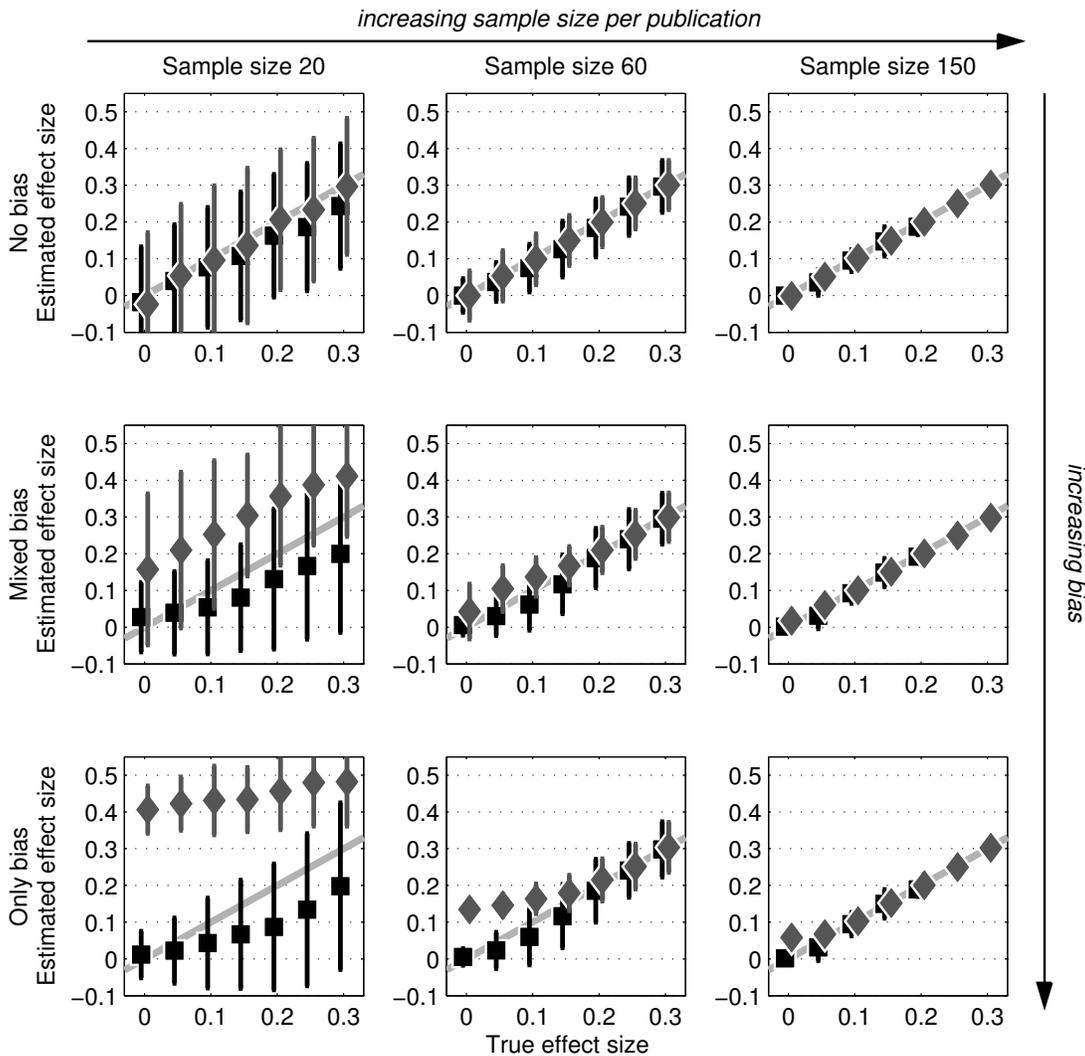


Figure 2. Recovery of the mitigation method (**squares**) for a single paper, evaluated under three different bias conditions, with three different sample sizes, and a range of effect sizes. The classical effect size measure Hedges'  $g$  (**diamonds**) is given for comparison. Vertical lines extend one standard deviation in each direction. In almost all scenarios, the mitigation method either outperforms the classical method, or recovers the same effect size.

or not, with the publication decision based on one of three *regimes*: Either the finding was subjected to no publication bias at all (“no bias”), or was subjected to extreme publication bias (“only bias”; only significant effects published), or to a combination of all four censoring functions (with equal probability; “mixture”). This implies that oftentimes no result was published, so there was nothing to compute. In the cases where a result *was* published, we applied our mitigation method to improve the estimate of the underlying effect size. In addition to the three biasing *regimes*, we also manipulated the simulated *sample size*  $n$ ,

which was either small ( $n = 20$ ), medium ( $n = 60$ ), or large ( $n = 150$ ). These values were inspired by Marszalek, Barber, Kohlhart, and Holmes (2011)’s review of typical sample sizes in experimental psychology.

Each panel in Figure 2 represents one *regime*  $\times$  *sample size* combination, with the horizontal axis indicating the true value of the effect size and the vertical axis representing the estimated effect size. Squares indicate our mitigated effect size, diamonds indicate the classical effect size measure Hedges’  $g$  (also computed in each case on the basis of a single “published” effect size). Unbiased estimates fall on the indicated diagonal.

In almost all cases, the mitigated effect size is closer to the true effect size than the classical estimator. As the degree of publication bias increases, both measures perform worse, but the effect of publication bias on the classical estimator is dramatic while the effect on the new estimator is comparatively modest. Both methods converge to the true value as the sample size  $n$  in the simulated papers increases,<sup>3</sup> and in the no-bias scenario Hedges’  $g$  is exact (on average).

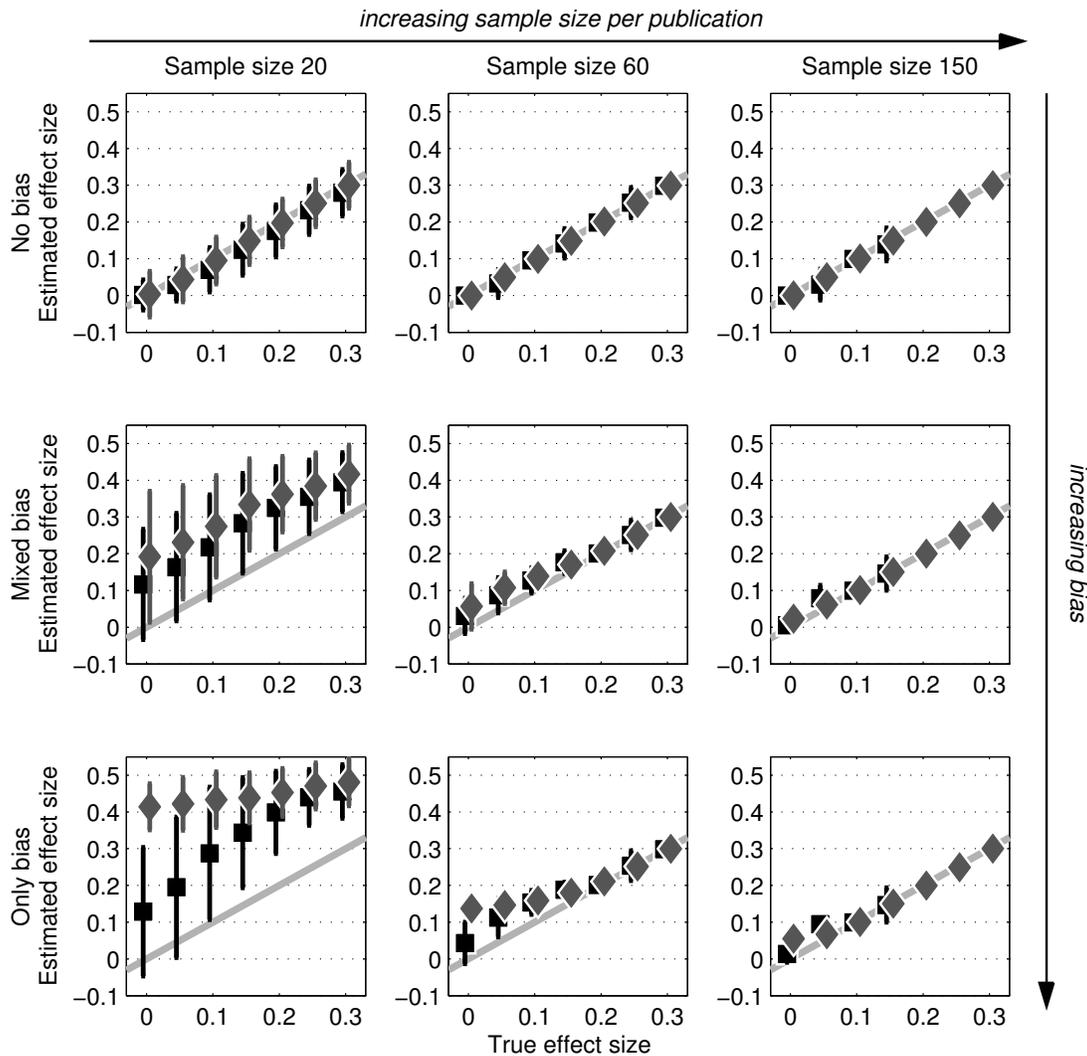
**(2) Meta-analysis.** In the second Monte Carlo study, we simulated a situation in which ten independent studies were conducted on the same hypothetical phenomenon of a given effect size. Again, our effect sizes ranged from 0 to 0.3 in increments of 0.05. For each of the ten studies, the result of a one-sided paired  $t$  test was either published or not, under the same three regimes as above, so that some proportion of the results was visible in the literature. We also again manipulated the simulated sample size  $n$ , which was either small ( $n = 20$ ), medium ( $n = 60$ ), or large ( $n = 150$ ), and which was the same for all studies in each set. The “unpublished” results were discarded, and meta-analytic effect sizes computed, either using our new method, or using the aggregated Hedges’  $g$ .

Figure 3 is set up in the same way as Figure 2, with each panel representing one *regime*  $\times$  *sample size* combination; horizontal axes indicating the true value of the effect size and vertical axes representing the estimated effect size; Squares indicate our mitigated effect size and diamonds indicate the classical aggregated effect size Hedges’  $g$  (also computed only on the basis of the “published” effect sizes).

It is again true that in almost all cases, the mitigated effect size is closer to the true effect size, and that the effect of increasing publication bias is greater for the classical estimator than for the new estimator.

**(3,4) Meta-analysis with different priors.** The final two Monte Carlo studies served to illustrate the sensitivity of our analyses to the prior distribution over models. For this purpose, we repeated the meta-analytical design of the second Monte Carlo study with two alternative prior distributions. Our default prior was  $(20, 20, 1, 1, 1, 1, 1, 1) / 46$  (see Appendix B for a rationale), but in these simulations we changed it first to  $(20, 0, 1, 0, 1, 0, 1, 0) / 23$ , eliminating the null hypothesis, and then to  $(0, 0, 0, 0, 1, 1, 1, 1) / 4$ , eliminating all but the two more nuanced biasing processes. The qualitative effect of these changes on the recovery performance was negligible, and we do not discuss them here.

<sup>3</sup>For large effect sizes and large sample sizes, the numerical integration required to compute the mitigated effect size becomes somewhat time-consuming, so we omit those conditions from the studies reported in Figures 2 and 3 while we work to make the software more efficient.



*Figure 3.* Recovery of the mitigation method (**squares**) for a simulated literature of 10 papers, of which only a biased subset is visible to the method, evaluated under three different bias conditions, with three different sample sizes, and a range of effect sizes. The classical effect size measure Hedges’  $g$  (**diamonds**) is given for comparison. In almost all scenarios, the mitigation method either outperforms the classical method, or recovers the same effect size.

### Example applications

**Wishful seeing.** We now consider a single test within a published paper. Balcetis and Dunning (2010) reported evidence for “wishful seeing,” a phenomenon in which desirable objects are perceived to be physically closer than other objects that are less desirable. This effect, the authors suggest, serves the function of energizing the perceiver to engage in

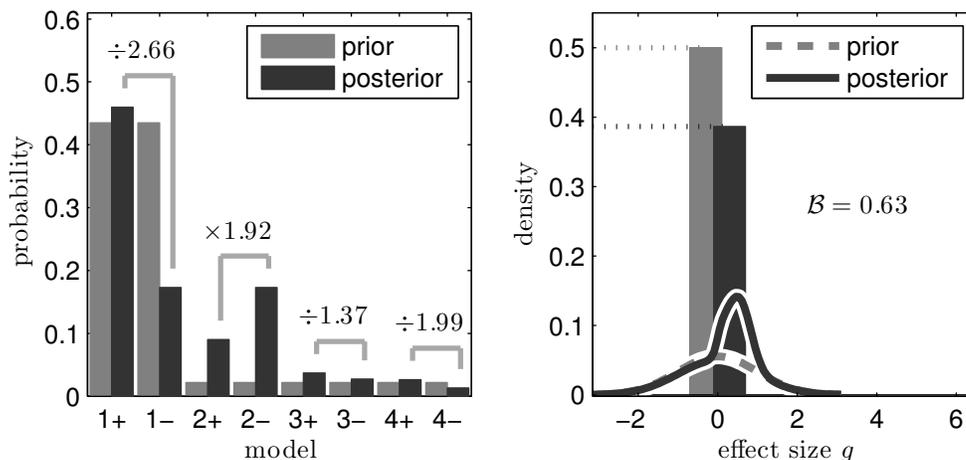


Figure 4. Model results for the “wishful seeing” example. *Left panel:* Prior and posterior probabilities over models. Numbers in the horizontal axis labels indicate the model, + and – indicate the states of nature (true effect vs. no true effect, resp.). Above the bars we indicate the posterior ratio between the two states of nature within each biasing model. For example, if we assume there is no bias at all (i.e.,  $\mathcal{M}_1$  is true), then we have 2.66 times more evidence for the effect than we do for the null, but if we assume bias is extreme (i.e.,  $\mathcal{M}_2$  is true), then we have 1.92 times more evidence for the null than we do for the effect. *Right panel:* Prior and posterior distributions over the true value of the effect size Hedges’  $g$ . Both distributions have point masses at  $g = 0$  (i.e., a probability that the effect size is exactly zero owing to the inclusion of the four no-effect models), which are displayed as bars. The posterior distribution at 0 is lower than the prior, indicating that the data support the effect slightly more than the null. Aggregated over all models, there is 1.59 times more evidence for the effect than there is for the null ( $1.59 = \frac{1}{0.63}$ ).

actions that lead to the goal of obtaining the desirable object. In one study, participants were asked to position themselves at a certain distance relative to an object. The authors hypothesized that participants would stand further from a desirable object (chocolates) than from an undesirable one (feces), because desirable chocolate would be misperceived as physically closer. In line with their prediction, the authors found a significant effect on perceived distance depending on whether participants saw chocolates or feces (unpaired  $t_{50} = 2.29$ ,  $p = .026$ , Hedges’  $g = 0.64$ ).

Using only the reported  $t$  value and sample size, and making no assumptions beyond those requisite for the  $t$  test and a prior on effect size Hedges’  $g = t\sqrt{(n_f + n_c)/(n_f n_c)}$ , we are able to compute probabilities for each of the eight models (using formulas given in Appendix B). These probabilities can then be used as weights to compute a new, mitigated effect size (or, equivalently,  $t$  value) that takes into account the possibility and likelihood of biasing mechanisms.

The result of this exercise is displayed in Figure 4. The left panel depicts prior (grey bars; see Appendix B for a rationale of the choice of prior model probabilities) and posterior

(black bars) probabilities of the eight models. The largest change in probability from prior to posterior is seen in model  $\mathcal{M}_{2-}$ : the extreme-bias model under the assumption that the null hypothesis is true becomes almost 8 times more likely after taking into account the data. Model  $\mathcal{M}_{1+}$ , the model of no bias with the null hypothesis being false, received the highest weight. The panel shows no compelling reason to prefer the two more complex biasing processes  $\mathcal{M}_3$  and  $\mathcal{M}_4$ .

The posterior distribution (black bars) quantifies what we know intuitively: a significant  $t$  value results either from a true effect reported in an unbiased world, or no true effect and an active biasing process. From the posterior distribution, we also learn that if we assume there is no publication bias, then there is about 2.66 times as much evidence for the alternative hypothesis as there is for the null (by comparing the two leftmost black bars). If we assume there is extreme bias, there is about 1.92 times as much evidence for the null as there is for the alternative (by comparing the next two black bars).

The right panel of Figure 4 shows the prior and posterior distributions of the effect size parameter. Since half of the models under consideration predict an effect size of exactly zero, the prior and posterior probabilities of this null hypothesis are shown as bars in the figure. The panel shows that the data reported by Balcetis and Dunning (2010) do not deliver convincing evidence regarding the effect. The probability that  $g = 0$  is slightly lower after observing the data than it was before (the Bayes factor  $\mathcal{B}$  in favor of the null hypothesis is 0.63).

In order to evaluate the sensitivity of our method to the prior assumptions, we repeated the analysis under three scenarios that differ in the prior. In alternative scenario 1, we give prior weight only to  $\mathcal{M}_1$ . This is equivalent to a default Bayesian  $t$ -test using a unit-information Bayes factor (as in Rouder, Speckman, Sun, Morey, & Iverson, 2009), and yields a Bayes factor  $\mathcal{B} = 2.81$  against the null. In alternative scenario 2, we distribute prior weight over  $\mathcal{M}_1$  and  $\mathcal{M}_2$ . This leads to a very similar inference but with a spike-and-slab prior on the effect size that has a spike at  $g = 0$ . Now  $\mathcal{B} = 1.34$  against the null. In scenario 3, we consider only the models that allow for an effect size ( $\mathcal{M}_{1+}$ ,  $\mathcal{M}_{2+}$ ,  $\mathcal{M}_{3+}$ , and  $\mathcal{M}_{4+}$ ). Including the possibility of bias turns the balance of evidence:  $\mathcal{B} = 0.96$ .

Our primary analysis described above allows for both true null effects and publication bias, and there the evidence for the null was even stronger. It is clear from this example that the outcome of revisiting a single empirical result depends to some extent on whether or not we are willing to consider that the effect may be truly zero, and whether or not we are willing to consider the possibility of a biased publication process. However, the analysis as presented here (in Fig. 4) allows for nuanced conclusions to accommodate different prior assumptions.

**Meta-analysis: Depression.** Using the statistical mitigation approach, we can also mitigate the effects of publication bias in meta-analyses. In this section, we consider Bolier et al.’s 2013 meta-analysis of the effectiveness of positive psychology interventions. Bolier et al. focused on the effects of positive psychology interventions on subjective well-being, psychological well-being, and depressive symptoms. We will focus on the effects on depressive symptoms. The authors included a total of 14 studies that looked at positive psychology interventions on depression. We extracted the focal statistics ( $z$ -values) directly from Figure 4 in the original article, and used those  $z$ -values for our mitigation analysis.

The results are depicted in Figure 5. The left panel shows the prior and posterior

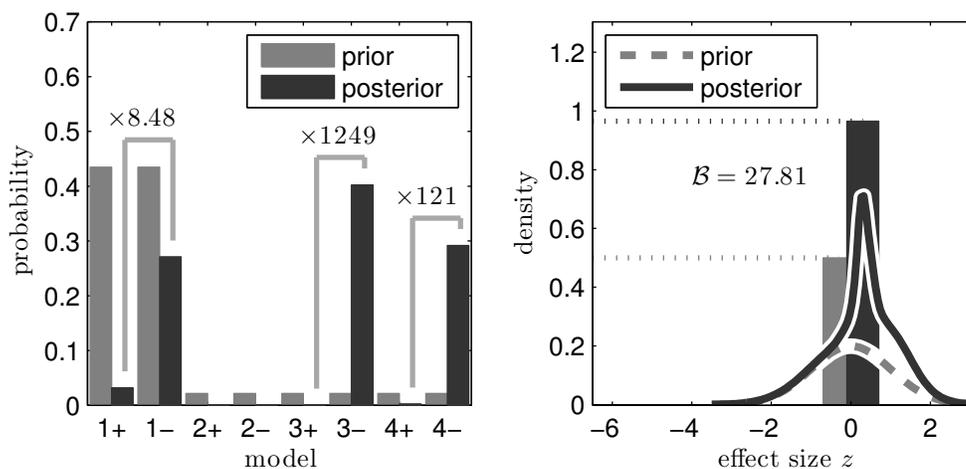


Figure 5. Results for the depression meta-analysis. *Left panel:* Prior and posterior probabilities over models. The most weight (and largest shift) is seen in the models that suppose constant or exponential bias, with true effect size zero. The extreme-bias model ( $\mathcal{M}_2$ ) is falsified by the presence of non-significant results. Within each biasing model, the posterior probability of the null model is much higher than that of the effect model. *Right panel:* Prior and posterior distributions over the true value of the test statistic  $z$ . The point mass at  $z = 0$  is approximately 62.2% higher under the posterior than under the prior. Given that the prior point mass on  $z = 0$  was 50% and the maximum shift is therefore by a factor of 2, this indicates a sizeable shift in the weight of evidence towards the null value.

probabilities over models. We observe large increases in probability from prior to posterior in both models  $\mathcal{M}_{3-}$  and  $\mathcal{M}_{4-}$ , the two more complex biasing models under the assumption that the null hypothesis is true. In addition, model  $\mathcal{M}_{1-}$  with no bias under the assumption that the null is true also receives considerable weight: 8.48 times more than the no-bias model with an effect ( $\mathcal{M}_{1+}$ ). Results show next to no evidence in support of any model that assumes the null hypothesis is false. The right panel shows the prior and posterior probability distributions of the effect size parameter  $z$ . The large weight of the various null-effect models is seen here in the increased posterior density at  $z = 0$ . The Bayes factor in favor of the null hypothesis is 27.81, making the posterior probability of the null approximately 96.5%.

**Meta-analysis: Sleep quality.** In this section, we consider another meta-analysis. Casement and Swanson (2012) conducted a meta-analysis of imagery rehearsal therapy for post-trauma nightmares; aggregating results from 9 papers. The original analysis evaluated the effect of imagery rehearsal on three measures: nightmare frequency, sleep quality, and posttraumatic stress. We will revisit the effect on sleep quality (as measured by the Pittsburgh Sleep Quality Index or PSQI). We obtained the focal statistics from each of the nine individual studies listed in Table 2 (p. 571) of Casement and Swanson (2012).<sup>4</sup>

The Bayesian mitigation results of this meta-analysis are summarized in Figure 6.

<sup>4</sup>We thank Melynda Casement for kindly providing data pertaining to the meta-analysis.

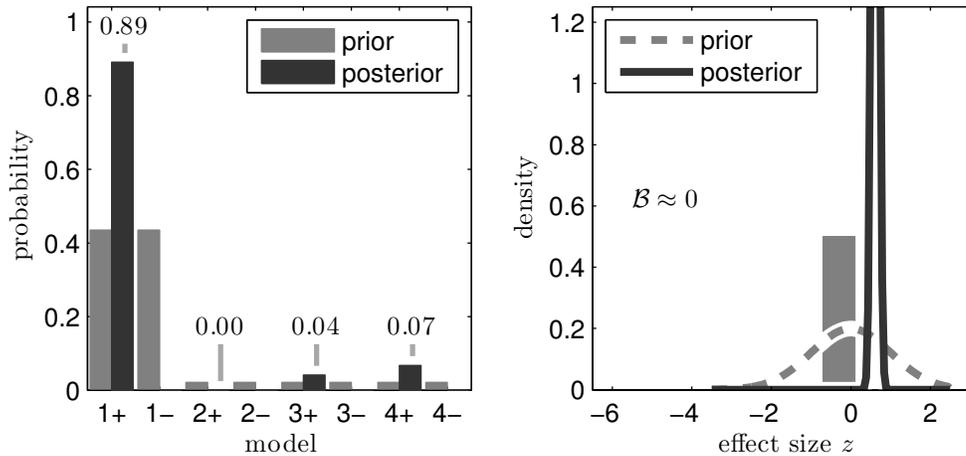


Figure 6. Results for the sleep quality meta-analysis. *Left panel:* Prior and posterior probabilities over models. The exact posterior probabilities are printed. *Right panel:* Prior and posterior distributions over the true value of  $z$ . The point mass at  $z = 0$  is negligible under the posterior, indicating a strong shift in the weight of evidence towards the hypothesis that there is a non-zero effect.

From the left panel, it is clear that the most posterior weight is given to  $\mathcal{M}_{1+}$ , the model under which there is no bias and the null hypothesis is false (i.e., there is an effect of imagery rehearsal on sleep quality). The rest of the models receive negligible weight. In the right panel, we show the prior and posterior distributions of the effect parameter measured in terms of  $z$  values. It is evident here that the meta-analysis data from Casement and Swanson (2012) indeed changes the information about the effect. The posterior mean  $z$  is approximately 0.614, and the probability density at  $z = 0$  is essentially zero—it is trillions of times lower after observing the data than before (Bayes factor against the null:  $7 \times 10^{15}$ ). Overall, our results show that after taking into account the possibility of various forms of publication bias, this meta-analysis indicates that imagery rehearsal is indeed effective in improving sleep quality.

## Discussion

We have outlined a novel method for inference regarding published effects. Our method is based on the assumption that there may be processes biasing the visibility of empirical studies, and relies on probability calculus to perform inference conditional on this possibility. The method has many desirable qualities that resonate well with intuitions researchers have about science and statistics. All that is required to apply the method is the published effect sizes. Sets of studies that consist exclusively of just-significant findings—which indicates that bias is likely, and non-significant findings are disproportionately hidden from the literature—are rendered inconclusive or indicative of only very small effects, while sets that contain effects of various sizes (including non-significant ones) provide more evidence for relatively larger effects. Unlike classical null hypothesis significance testing, our

Bayesian method allows for the affirmation of the null hypothesis as well as its falsification, in addition to allowing for the inclusion of prior information.

Throughout, we focus on the empirical literature on a topic as the population of interest: For the purposes of effect size estimation across papers, all included studies should report on the same effect so that our mitigated effect size has a clear interpretation. However, we do not see the model probabilities as parameters to be estimated for any population. Instead, we think of the entire biasing process as an inhomogeneous mixture of various censoring mechanisms. Any given paper, under a given set of fleeting circumstances, may fit the assumptions of any model or some combination of them.

There are some potential caveats with the approach we present here. First, there is the issue that assuming the possibility of publication bias makes the analysis overly conservative in the case where publication bias does not occur. However, we believe that in most common use cases, this is not a concern as we know publication bias to be a widespread issue. More generally, our method requires the explicit inclusion of domain knowledge both in the formulation of the biasing processes, their relative probabilities, and the prior distribution of the effect sizes. We believe the assumptions we have chosen are (a) both sensible and in line with what is known of the field and (b) superior to the assumption of no publication bias that is implicit in all classical analyses.

Second, the censoring processes that we propose all assume that significant results are always published. Clearly this is not the case—papers submitted for publication are often rejected even bearing statistically significant results. However, we do not believe that there exists any systematic suppression of results because they were statistically significant to such an extent that it is skewing the literature at a scale comparable to the bias against the null. Furthermore, it is reasonable to assume that those papers with significant results that are not published often have issues unrelated to the outcome, for example with the experimental design (which would preclude use in meta-analysis in its own right). Finally, if indeed a nonignorable fraction of significant results goes unpublished, randomly, even though the applied methods are sound, then a valid interpretation of our censoring functions is simply that they describe the probability of publication of a nonsignificant result *relative to the probability of publication of a significant result*, with as only restriction that the former is no greater than the latter, which we consider a very weak assumption.

Third, there may be other biasing mechanisms in addition to the ones we propose here. However, we believe that the set of processes we consider spans a reasonably large range of possible mechanisms—that is, we struggle to think of biasing mechanisms that will yield fundamentally different censoring functions. One possible exception is a practice known as “hypothesizing after the results are known” (HARKing), which may in fact cause an overrepresentation of results that are *clearly* nonsignificant (e.g.,  $p > .50$ ), because these results are declared non-pivotal after they are known, and are then reported parenthetically in a paper with a different focal test. We decided not to include a HARKing model here because such a model would apply to a different set of observations (such as “non-pivotal statistics,” which are published statistics that are not used to support inference regarding a new finding or that are otherwise irrelevant for the main narrative of a paper; e.g., manipulation checks and failed secondary manipulations). A possible HARKing model is formally very similar to  $\mathcal{M}_4$ , but predicts that the probability of publishing nonsignificant findings increases as  $e^{-\gamma(1-p)}$ . Another biasing mechanism for which we do not currently

account is academic fraud.

Fourth, our meta-analyses as we have implemented them here make the assumption of independence between studies. In practice, this assumption may be wrong and biases may tend to be correlated (positively in the case of labs repeatedly publishing on the same topic, but negatively in the case of adversarial replication attempts).<sup>5</sup> Our Bayesian model averaging framework allows us to account for correlated bias in the same way we deal with other unknowns: marginalization over the unknown quantity. We provide the equations for such an exercise in Appendix B, but leave the implementation of this analysis aside until a computationally efficient treatment for this case is developed.

Finally, we wish to emphasize again the subjectivity of some of the assumptions we have made, and the fact that reasonable people can reasonably disagree in the case of differing prior assumptions. The method that we have outlined allows for the inclusion of a wide variety of prior assumptions, and it may be improved by further psychological research in the behavioral processes of publication bias.

#### References

- Abramowitz, M., Stegun, I. A. et al. (1972). *Handbook of mathematical functions*. Dover New York.
- Balcetis, E. & Dunning, D. (2010). Wishful seeing: Desired objects are seen as closer. *Psychological Science*, *21*, 147–152.
- Bayarri, M. & DeGroot, M. (1987). Bayesian analysis of selection models. *The Statistician*, 137–146.
- Bayarri, M. & DeGroot, M. (1991). *The analysis of published significant results*. Purdue University. Department of Statistics.
- Bolier, L., Haverman, M., Westerhof, G. J., Riper, H., Smit, F., & Bohlmeijer, E. (2013). Positive psychology interventions: a meta-analysis of randomized controlled studies. *BMC Public Health*, *13*(1), 119.
- Casement, M. D. & Swanson, L. M. (2012). A meta-analysis of imagery rehearsal for post-trauma nightmares: effects on nightmare frequency, sleep quality, and posttraumatic stress. *Clinical Psychology Review*, *32*(6), 566–574.
- Ellis, P. D. (2010). *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.
- Francis, G. (2012a). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991.
- Francis, G. (2012b). The same old new look: publication bias in a study of wishful seeing. *i-Perception*, *3*(3), 176–178.
- Francis, G. (2012c). Too good to be true: publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*(2), 151–156.
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: unlocking the file drawer. *Science*, *345*(6203), 1502–1505.
- Givens, G., Smith, D. D., & Tweedie, R. L. (1997). Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science*, *12*, 221–250.

<sup>5</sup>We thank Richard Shiffrin for pointing out this issue with meta-analysis in general.

- Greenwald, A. G. (1975). Significance, nonsignificance, and interpretation of an ESP experiment. *Journal Of Experimental Social Psychology*, *11*, 180–191.
- Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, *7*(2), 246–255.
- Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.
- Ioannidis, J. P. & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245–253.
- Iyengar, S. & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 109–117.
- Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.
- Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 377–395.
- Marszalek, J., Barber, C., Kohlhart, J., & Holmes, C. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*(2), 331–348.
- Masicampo, E. & Lalande, D. R. (2012). A peculiar prevalence of  $p$  values just below .05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279.
- Pashler, H. & Wagenmakers, E.-J. (2012). Editor’s Introduction to the Special Section on Replicability in Psychological Science: a crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian  $t$  tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014).  $p$ -curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534.
- Stanley, T. D. & Doucouliagos, H. (2013). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78.
- van Assen, M. A., van Aert, R., & Wicherts, J. M. (in press). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*.
- Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, J. T. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–317). Oxford, UK: Oxford University Press.
- Verhagen, A. J. & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475.
- Young, N. S., Ioannidis, J. P., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS medicine*, *5*(10), e201.

## Appendix A

## Behavioral models of publication bias

Here we detail the behavioral models of publication bias that we employ to mitigate the effects of bias on reported effect sizes, and we describe the details of the Bayesian inference methods that we apply to these models. For convenience of exposition, we will focus on unpaired  $t$  tests, but the method extends to other parametric tests based on  $p$ -values.

In the description of the models that follows,  $p(\cdot)$  will indicate a likelihood function,  $t(\cdot | \nu)$  will indicate the probability density function of a central  $t$  distribution with  $\nu$  degrees of freedom and  $t'(\cdot | x, \nu)$  of the noncentral  $t$  distribution with noncentrality parameter  $x$ .  $T(\cdot | \nu)$  and  $T'(\cdot | x, \nu)$  indicate their respective cumulative distribution functions.  $n_1$  and  $n_2$  will be the sample sizes in two groups.  $\delta^{obs}$  will indicate a test statistic (TS) obtained from the sample. A ‘true’ effect size (ES) will be indicated  $\eta^{true}$ . The mitigated TS  $\delta^{mit}$  is our estimate of the true TS.

Throughout, it is important to note that, given a sample size and a design, any TS uniquely maps to an ES. For example, for the unpaired  $t$  test,  $\eta^{true} = \varphi \delta^{true}$  with  $\varphi = \sqrt{\frac{n_1+n_2}{n_1 n_2}}$ . Transformations for other inferential tests are well known (e.g., Ellis, 2010; or succinctly in Footnote 4 of Verhagen & Wagenmakers, 2014). To define a prior for the TS, we propose a unit information prior on the ES  $\eta^{true}$ :  $\eta^{true} \sim N(0, 1)$ . We can readily transform this prior to the scale of the TS:  $(\delta^{true} | \mathcal{M}_+) \sim N(0, \varphi^2)$ , where  $\mathcal{M}_+$  indicates any model under which the null hypothesis  $\eta^{true} = 0$  is false. The prior might also be generalized to  $(\delta^{true} | \mathcal{M}_+) \sim N(\varphi\mu, \varphi^2\sigma^2)$ , to either reduce the slightly assumptive nature of the unit information prior or to incorporate genuine prior knowledge.

The available data are  $n_1$  and  $n_2$  and the observed TS  $\delta^{obs}$  of at least one experiment. Associated with  $\delta^{obs}$  is an observed  $p$ -value  $p^{obs} = 2 \times T(-|\delta^{obs}| | \nu)$ . Below, we give the likelihood functions associated with each of the four behavioral models of publication bias, under each of the two states of nature ( $H_0$  false or true).

 **$\mathcal{M}_1$ : A no-bias model**

Suppose  $\mathcal{M}_1$ : There is no publication bias. The probability of publishing,  $P_{pub}$ , is 1.

**Case  $\mathcal{M}_{1+}$ :  $H_0$  false.**  $p(\delta^{obs} | \mathcal{M}_{1+}, \eta^{true}) = t'(\delta^{obs} | \eta^{true}, \nu)$ , with true ES  $\eta^{true}$ .

**Case  $\mathcal{M}_{1-}$ :  $H_0$  true.**  $p(\delta^{obs} | \mathcal{M}_{1-}, \eta^{true}) = t(\delta^{obs} | \nu)$ , with  $\nu$  degrees of freedom.

 **$\mathcal{M}_2$ : An extreme-bias model**

Suppose  $\mathcal{M}_2$ : Publication bias is extreme—publication only happens if significance is found. The probability of publishing,  $P_{pub}$ , is a step function of  $p^{obs}$ :

$$P_{pub}(p^{obs}) = \begin{cases} 0 & \text{if } p^{obs} \geq \alpha \\ 1 & \text{otherwise} \end{cases}.$$

where  $p^{obs}$  is a function of the observed TS  $\delta^{obs}$ , as above, and  $\alpha$  is conventionally set to .05. Associated with  $\alpha$  is the critical TS  $\delta^{crit}$ , which is the smallest TS (in absolute value) to be considered significant at level  $\alpha$  (i.e., any TS under  $-\delta^{crit}$  or above  $\delta^{crit}$  would be considered statistically significant).

**Case  $\mathcal{M}_{2+}$ :  $H_0$  false.**

$$p\left(\delta^{obs} \mid \mathcal{M}_{2+}, \eta^{true}\right) = \begin{cases} 0 & \text{if } p^{obs} \geq \alpha \\ \frac{1}{B_2} t'(\delta^{obs} \mid \eta^{true}, \nu) & \text{otherwise} \end{cases},$$

or the  $t'$  distribution truncated to the significance region. Note that the denominator  $B_2$  can be computed without the need for expensive numerical methods, using the direct calculation of the cumulative noncentral  $t$  distribution  $T'$ :

$$\begin{aligned} B_2 &= \int_{-\infty}^{-\delta^{crit}} t'(x \mid \eta^{true}, \nu) dx + \int_{\delta^{crit}}^{+\infty} t'(x \mid \eta^{true}, \nu) dx \\ &= T'(-\delta^{crit} \mid \eta^{true}, \nu) + 1 - T'(\delta^{crit} \mid \eta^{true}, \nu). \end{aligned}$$

**Case  $\mathcal{M}_{2-}$ :  $H_0$  true.**

$$p\left(\delta^{obs} \mid \mathcal{M}_{2-}\right) = \begin{cases} 0 & \text{if } p^{obs} \geq \alpha \\ \frac{t(\delta^{obs} \mid \nu)}{\int_{-\infty}^{-\delta^{crit}} t(x \mid \nu) dx + \int_{\delta^{crit}}^{+\infty} t(x \mid \nu) dx} = \frac{1}{\alpha} t(\delta^{obs} \mid \nu) & \text{otherwise} \end{cases},$$

that is, the  $t$  distribution truncated to the region that yields statistical significance.

### $\mathcal{M}_3$ : A constant-bias model inspired by Greenwald

Greenwald (1975) proposed a model of the research-publication process in which a number of parameters characterize the various steps of publishing in psychology, such as the investigators' probability of reporting research and the editors' probability of publishing manuscripts reporting significant or nonsignificant results. We simplify this model by summarizing the probabilities in each step into one single constant probability of publishing nonsignificant results.

Suppose  $\mathcal{M}_3$ : Publication occurs with certainty if a significant effect is found, but also with some constant probability  $\pi$  if no significant effect is found. The probability of publishing,  $P_{pub}$ , is

$$P_{pub}(p^{obs}, \pi) = \begin{cases} \pi & \text{if } p^{obs} \geq \alpha \\ 1 & \text{otherwise} \end{cases}.$$

**Case  $\mathcal{M}_{3+}$ :  $H_0$  false.**

$$p\left(\delta^{obs} \mid \mathcal{M}_{3+}, \eta^{true}, \pi\right) = \begin{cases} \pi \frac{1}{B_3} t'(\delta^{obs} \mid \eta^{true}, \nu) & \text{if } p^{obs} \geq \alpha \\ \frac{1}{B_3} t'(\delta^{obs} \mid \eta^{true}, \nu) & \text{otherwise} \end{cases},$$

where  $0 \leq \pi \leq 1$  with prior  $\pi \sim U(0, 1)$ , and

$$\begin{aligned} B_3 &= \pi \int_{-\delta^{crit}}^{\delta^{crit}} t'(x \mid \eta^{true}, \nu) dx + \int_{\delta^{crit}}^{+\infty} t'(x \mid \eta^{true}, \nu) dx + \int_{-\infty}^{-\delta^{crit}} t'(x \mid \eta^{true}, \nu) dx \\ &= 1 - (1 - \pi) [T'(\delta^{crit} \mid \eta^{true}, \nu) - T'(-\delta^{crit} \mid \eta^{true}, \nu)]. \end{aligned}$$

**Case  $\mathcal{M}_{3-}$ :  $H_0$  true.**

$$p\left(\delta^{obs} \mid \mathcal{M}_{3-}, \pi\right) = \begin{cases} \pi \frac{1}{A_3} t(\delta^{obs} \mid \nu) & \text{if } p^{obs} \geq \alpha \\ \frac{1}{A_3} t(\delta^{obs} \mid \nu) & \text{otherwise} \end{cases},$$

where  $A_3 = \pi \int_{-\delta^{crit}}^{\delta^{crit}} t(x \mid \nu) dx + 2 \int_{\delta^{crit}}^{+\infty} t(x \mid \nu) dx = \pi(1 - \alpha) + \alpha$ .

**$\mathcal{M}_4$ : An exponential-bias model inspired by Givens**

In Givens et al. (1997), another approach is introduced to estimate and adjust for publication bias, by partitioning the unit interval into segments so that a  $p$ -value from any given study falls into one of these regions. Each interval region is assigned a corresponding probability of publication, with probabilities decreasing as the region departs from significance. In  $\mathcal{M}_4$ , we capture this same concept with a probability of publishing that decreases exponentially as a function of the difference between the observed  $p$ -value  $p^{obs}$  and  $\alpha$ . The rate of exponential decay as  $p^{obs}$  departs from  $\alpha$  is determined by a strictly positive rate parameter  $\lambda$ . A-priori, we suppose  $\lambda \sim \text{Exp}(5)$ .

Suppose  $\mathcal{M}_4$ : Publication occurs with certainty if a significant effect is found, but also with some nonzero probability if no significant effect is found.

$$P_{pub}(p^{obs}, \lambda) = \begin{cases} e^{-\lambda(p^{obs}-\alpha)} & \text{if } p^{obs} \geq \alpha \\ 1 & \text{otherwise} \end{cases}.$$

**Case  $\mathcal{M}_{4+}$ :  $H_0$  false.**

$$p(\delta^{obs} | \mathcal{M}_{4+}, \eta^{true}, \lambda) = \begin{cases} e^{-\lambda(p^{obs}-\alpha)} \frac{1}{B_4} t'(\delta^{obs} | \eta^{true}, \nu) & \text{if } p^{obs} \geq \alpha \\ \frac{1}{B_4} t'(\delta^{obs} | \eta^{true}, \nu) & \text{otherwise} \end{cases},$$

where

$$\begin{aligned} B_4 &= \int_{-\delta^{crit}}^{\delta^{crit}} e^{-\lambda(p(x,\nu)-\alpha)} t'(x | \eta^{true}, \nu) dx \\ &\quad + \int_{\delta^{crit}}^{+\infty} t'(x | \eta^{true}, \nu) dx + \int_{-\infty}^{-\delta^{crit}} t'(x | \eta^{true}, \nu) dx \\ &= \int_{-\delta^{crit}}^{\delta^{crit}} e^{-\lambda(p(x,\nu)-\alpha)} t'(x | \eta^{true}, \nu) dx \\ &\quad + [1 - T'(\delta^{crit} | \eta^{true}, \nu)] + T'(-\delta^{crit} | \eta^{true}, \nu) dx. \end{aligned}$$

with  $p(x, \nu) = 2 \times T(-|x| | \nu)$ , the  $p$ -value associated with a particular observation given  $\nu$  degrees of freedom.

**Case  $\mathcal{M}_{4-}$ :  $H_0$  true.**

$$p(\delta^{obs} | \mathcal{M}_{4-}, \lambda) = \begin{cases} e^{-\lambda(p^{obs}-\alpha)} \frac{1}{A_4} t(\delta^{obs} | \nu) & \text{if } p^{obs} \geq \alpha \\ \frac{1}{A_4} t(\delta^{obs} | \nu) & \text{otherwise} \end{cases},$$

where

$$\begin{aligned} A_4 &= \int_{-\delta^{crit}}^{\delta^{crit}} e^{-\lambda(p(x,\nu)-\alpha)} t(x | \nu) dx + 2 \int_{\delta^{crit}}^{+\infty} t(x | \nu) dx \\ &= \int_{-\delta^{crit}}^{\delta^{crit}} e^{-\lambda(p(x,\nu)-\alpha)} t(x | \nu) dx + \alpha. \end{aligned}$$

## Appendix B

### Bayesian inference details

#### Computation of Jeffreys weights

A standard Bayesian approach to model comparison is the Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995). Bayes factors summarize the evidence provided by the observed data in favor of one model over another, and implicitly control for goodness-of-fit as well as model complexity. The Bayes factor  $\mathcal{B}$  between two models is simply the ratio of Bayesian

evidence, also known as the marginal likelihood, for each model. The Bayesian evidence  $\mathcal{E}$  is computed by integrating the likelihood over the prior parameter space, so that the Bayes factor between two models,  $\mathcal{M}_{1+}$  and  $\mathcal{M}_{2-}$ , denoted  $\mathcal{B}(\mathcal{M}_{1+} : \mathcal{M}_{2-})$  is:

$$\mathcal{B}(\mathcal{M}_{1+} : \mathcal{M}_{2-}) = \frac{\mathcal{E}_{1+}}{\mathcal{E}_{2-}} = \frac{\int p(x | \theta_1, \mathcal{M}_{1+}) p(\theta_1 | \mathcal{M}_{1+}) d\theta_1}{\int p(x | \theta_2, \mathcal{M}_{2-}) p(\theta_2 | \mathcal{M}_{2-}) d\theta_2}$$

where  $p(x | \theta_2, \mathcal{M}_{2-})$  is the data-dependent likelihood for  $\mathcal{M}_{2-}$ ,  $p(\theta_1 | \mathcal{M}_{1+})$  is the prior distribution over the parameters of  $\mathcal{M}_{1+}$ , and the conditioning of the evidences and Bayes factor on the data  $x$  is implicit. In our case, the Bayes factors between these two models is nontrivial but not prohibitive. If  $|\delta^{obs}| < |\delta^{crit}|$ ,  $\mathcal{B}(\mathcal{M}_{1+} : \mathcal{M}_{2-}) = +\infty$ , because this is an impossible occurrence under  $\mathcal{M}_{2-}$ . Otherwise,

$$\mathcal{B}(\mathcal{M}_{1+} : \mathcal{M}_{2-}) = \alpha \frac{\int t' \left( \delta^{obs} \mid \eta^{true}, \nu \right) \mathcal{N}(\eta^{true} \mid 0, \sqrt{n}) d\eta^{true}}{t(\delta^{obs} \mid \nu)}.$$

Note that the integral in the denominator disappears because model  $\mathcal{M}_{2-}$  has no random parameters. The integral in the numerator can be efficiently approximated with numerical methods such as Gaussian quadrature (Abramowitz, Stegun, et al., 1972).

The Bayes factor is the ratio of the posterior odds of one model to its prior odds, regardless of the actual value of the prior odds. Because it is an odds ratio, it has a convenient and intuitive interpretation. If  $\mathcal{B}(\mathcal{M}_{1+} : \mathcal{M}_{2-})$  is greater than 1, then  $\mathcal{M}_{1+}$  is more strongly supported by the observed data than  $\mathcal{M}_{2-}$ . Similarly, if  $\mathcal{B}(\mathcal{M}_{1+} : \mathcal{M}_{2-})$  is less than 1, then  $\mathcal{M}_{2-}$  is more strongly supported by the observed data than  $\mathcal{M}_{1+}$ . Intuitively, it quantifies how much one should update one's prior beliefs about the models under consideration, given evidence from the observed data. The Bayes factor also operates nicely on a continuum, as opposed to classical NHST in which a  $p$  value is judged against a single arbitrary cut-off criterion  $\alpha$ . Jeffreys (1961) suggests some interpretative boundaries, with values of 3, 10, and 30 corresponding to strengths of evidence that are “barely worth a mention,” “substantial,” and “strong,” respectively.

*Jeffreys weights* are a multi-alternative extension of Bayes factors, taking the Bayesian evidence for each model and normalizing by the sum of evidences over all models under consideration (Vandekerckhove, Matzke, & Wagenmakers, 2015). For our eight models, if the evidence for  $\mathcal{M}_k$  is denoted  $\mathcal{E}_k$ , then the corresponding Jeffreys weight  $\mathcal{J}_k$  is:

$$\mathcal{J}_k = \frac{\mathcal{E}_k}{\sum_{j=1}^8 \mathcal{E}_j}.$$

Jeffreys weights can be multiplied with model priors to arrive at model posterior probabilities:

$$P(\mathcal{M}_k | x) = \frac{\mathcal{J}_k P(\mathcal{M}_k)}{\sum_{j=1}^8 \mathcal{J}_j P(\mathcal{M}_j)}.$$

## Model priors

We need to define a prior distribution over the eight possible models  $\mathcal{M}_{1+}, \dots, \mathcal{M}_{4-}$ . In selecting our prior over models, we followed three main desiderata. First, we hold that observing a single significant TS should not allow us to differentiate between a

no-bias model and an extreme-bias model with the same true effect size. Given that  $\mathcal{B}(\mathcal{M}_{1-} : \mathcal{M}_{2-}) = \alpha$ , we decide that the prior for these models must reflect that ratio, so that  $P(\mathcal{M}_{1-}) = P(\mathcal{M}_{2-})/\alpha$ . Not scaling  $P(\mathcal{M}_{2-})$  by at least  $1/\alpha$  would cause a single significant observation to lead us to conclude that publication bias occurred.<sup>6</sup> Additionally, we do not want the prior to prefer either of the two states of nature:  $\forall i : p(\mathcal{M}_{i-}) = p(\mathcal{M}_{i+})$ . Finally, we want the prior to express prior equiprobability among all the biasing mechanisms (not including the no-bias mechanism). Taking all these desiderata together, the prior over models is defined as  $(20, 20, 1, 1, 1, 1, 1, 1)/46$ , and this is the prior pictured in Figures 4, 5, and 6.

### Aggregation of information across studies

Given multiple independent studies  $s$ , each with a unique observed TS  $\delta_s^{obs}$ , the aggregated posterior distribution of  $\delta^{mit}$  is obtained by first applying Bayes' theorem, then implementing the independence assumption, and then marginalizing over models:

$$\begin{aligned} p\left(\delta^{mit} \mid \left\{\delta_s^{obs}\right\}_1^S\right) &\propto p(\delta^{mit}) p\left(\left\{\delta_s^{obs}\right\}_1^S \mid \delta^{mit}\right) \\ &\propto p(\delta^{mit}) \prod_{s=1}^S p\left(\delta_s^{obs} \mid \delta^{mit}\right) \\ &\propto p(\delta^{mit}) \prod_{s=1}^S \sum_{\mathcal{M}} p\left(\delta_s^{obs} \mid \delta^{mit}, \mathcal{M}\right) P\left(\mathcal{M} \mid \left\{\delta_s^{obs}\right\}_1^S\right), \end{aligned}$$

and finally, computing the normalizing constant, so that

$$p\left(\delta^{mit} \mid \left\{\delta_s^{obs}\right\}_1^S\right) = \frac{p(\delta^{mit}) \prod_{s=1}^S \sum_{\mathcal{M}} p\left(\delta_s^{obs} \mid \delta^{mit}, \mathcal{M}\right) P\left(\mathcal{M} \mid \left\{\delta_s^{obs}\right\}_1^S\right)}{\int_{-\infty}^{\infty} p(x) \prod_{s=1}^S \sum_{\mathcal{M}} p\left(\delta_s^{obs} \mid x, \mathcal{M}\right) P\left(\mathcal{M} \mid \left\{\delta_s^{obs}\right\}_1^S\right) dx}. \quad (1)$$

In cases where the independence assumption is violated or not desired, a correlated error structure can instead be explicitly modeled, so that the effect size factor in the likelihood function becomes  $N\left(\left\{\varphi_s \delta_s^{obs}\right\}_1^S \mid \left\{\varphi_s \delta^{mit}\right\}_1^S, \Sigma\right)$ , with  $\Sigma \in R_{S \times S}$ . Unfortunately, the evaluation of the associated posterior distribution would require the repeated computation of an integral over the space of all covariance matrices, which is currently prohibitively expensive.

<sup>6</sup>A reviewer remarked that 20:1 prior odds (if  $\alpha = .05$ ) in favor of a no-bias model seems excessive given that we know bias for significant results to be widespread. Our argument for this model prior is admittedly subjective. We elicited our prior starting from the assumption that a single significant observation should not be discriminating between the bias and no-bias models.