Received: 26 February 2020 | Accepted: 26 August 2021

DOI: 10.1111/biom.13560

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Modelling publication bias and *p*-hacking

Jonas Moss | **Riccardo De Bin**

Department of Mathematics, University of Oslo, Oslo, Norway

Correspondence
Jonas Moss, Department of Mathematics, University of Oslo, Moltke Moes vei 35, 0851 Oslo, Norway.
Email: jonas.moss.statistics@gmail.com

Abstract

Publication bias and *p*-hacking are two well-known phenomena that strongly affect the scientific literature and cause severe problems in meta-analyses. Due to these phenomena, the assumptions of meta-analyses are seriously violated and the results of the studies cannot be trusted. While publication bias is very often captured well by the weighting function selection model, *p*-hacking is much harder to model and no definitive solution has been found yet. In this paper, we advocate the selection model approach to model publication bias and propose a mixture model for *p*-hacking. We derive some properties for these models, and we compare them formally and through simulations. Finally, two real data examples are used to show how the models work in practice.

KEYWORDS

file drawer problem, fishing for significance, meta-analysis, questionable research practices, selection bias

1 | INTRODUCTION

Meta-analysis is the quantitative combination of information from different studies. Aggregating information from multiple studies brings higher statistical power, higher accuracy in estimation and greater reproducibility. Unfortunately, it is not always possible to believe in the results of meta-analyses, as some model assumptions may be seriously violated. In particular, a meta-analysis must not be based on a biased selection of studies. Publication bias (Sterling, 1959) and *p*-hacking (Simmons *et al.*, 2011) are the most common phenomena that violate these assumptions. Depending on their magnitude, they may have a substantial effect on the cumulative evidence (see, e.g., Friese and Frankenbach, 2020).

Publication bias, also known as the *file drawer problem* (Rosenthal, 1979), denotes the phenomenon when a study with a smaller *p*-value is more likely to be published than a study with a higher *p*-value. Publication bias is a well-known issue, and several approaches have been proposed to tackle it. Two famous examples are the trim-and-fill (Duval and Tweedie, 2000) and fail-safe *N* (Becker, 2005) methods, but neither of them explicitly model the publication selection mechanism. From a statistical point of view, the most important class of models used to deal with publication bias are the selection models. They were first studied by Hedges (1984) for F-distributed variables with a cutoff at 0.05, and extended to the setting of t-values by Iyengar and Greenhouse (1988). Hedges (1992) proposed a random effects publication bias model with more than one cutoff, while Citkowicz and Vevea (2017) used beta distributed weights. Other examples of selection models include the non-parametric approach of Dear and-Begg (1992), the sensitivity analysis of Copas and Shi (2000) and the regression methods of Vevea and Hedges (1995). McShane *et al.* (2016) is an accessible overview of selection models in publication bias.

Publication bias is a well-known problem in several research areas, and therefore various approaches to solve the issue have been also proposed outside the statistical literature. Hailing from economics, PET, PEESE and PET-PEESE (Stanley and Doucouliagos, 2014) are models based on linear regression and an approximation of the selection mechanism based on the inverse Mill's ratio. From psychology, the *p*-curve of Simonsohn *et al.* (2014a) is a method that only looks at significant *p*-values and judges

whether their distribution shows sign of being produced by studies with insufficient power. The *p*-curve for estimation (Simonsohn *et al.*, 2014b) is a fixed effect selection model with a significance cutoff at 0.05 estimated by minimizing the Kolmogorov–Smirnov distance (McShane *et al.*, 2016). Another method from the psychology literature is *p*-uniform (van Assen *et al.*, 2015), which is similar to the *p*-curve. A recent study by Carter *et al.* (2019) compared several approaches and showed that the selection model works better than the others. However, not even the best method works well in every considered scenario. For more information on publication bias, we refer to Rothstein *et al.* (2006) and Marks-Anglin and Chen (2020).

In contrast, *p*-hacking, sometimes also called *questionable research practices* (Sijtsma, 2016) and *fishing for significance* (Boulesteix, 2009), occurs when the authors of a study manipulate results into statistical significance. *p*-hacking can be done at the experimental stage, using, for example, optional stopping, or at the analysis stage, for instance, by changing models or dropping out participants. Examples of *p*-hacking can be found in Simmons *et al.* (2011). While publication bias, at least that based on *p*-values, has been shown to be captured well by selection models such as that of Hedges (1992), *p*-hacking is harder to model (Carter *et al.*, 2019). The aforementioned *p*-curve approach by Simonsohn *et al.* (2014a) has been used for *p*-hacking as well, but it has been shown to be not reliable (Bruns and Ioannidis, 2016). Here, we advocate the selection model approach to model publication bias and propose a mixture model for *p*-hacking. We derive some properties for these models and argue they are best handled by Bayesian methods.

The paper is organized as follows: In Section 2, we define the framework and introduce the models, which are also theoretically compared. Further comparisons are presented through simulations in Section 3 and examples in Section 4. We conclude with some remarks and possible extensions in Section 5.

2 | MODELS

2.1 | Framework

The main ingredient of a meta-analysis is a collection of exchangeable statistics x_i. Each statistic x_i has density $f^\star(x_i \mid \theta_i, \eta_i)$, where η_i is a known or unknown nuisance parameter and θ_i is an unknown parameter we wish to make inference on. This paper corrects the bias due to the transformation of the true data-generation model $f^\star(x_i \mid \theta_i, \eta_i)$ into a new model $f(x_i \mid \theta_i, \eta_i)$ by publication bias and *p*-hacking. Our goal is to understand the latter model and correctly estimate θ_i. Although our methodology only requires the dependencies on θ_i and that statistical inference on θ_i is the goal of the analysis, here we mainly focus on the most common case of Gaussian densities (i.e. $f^\star(x_i \mid \theta_i, \eta_i) = \phi(x_i \mid \theta_i, \sigma_i^2)$). Moreover, as usual in meta-analysis studies (van Houwelingen *et al.*, 2002), we assume σ_i^2 as known.

The parameter θ_i is typically an effect size, such as a standardized mean difference. In a fixed effects meta-analysis, $\theta_i = \theta$ for all i. In a random effects meta-analysis, θ_i is drawn from an effect size distribution $p(\theta)$ common to all i, and the goal of the study is often to make inference on the parameters of the effect size distribution, usually on the mean θ_0 and the variance τ^2 when $\theta_i \sim N(\theta_0, \tau^2)$. If we marginalize away θ_i we will end up with a density on the form $N(\theta_0, \sigma_i^2 + \tau^2)$. This is possible in our framework, but it turns out that an important property of the publication bias model gets lost, as marginalizing out the θ_is can mask the fact that the selection mechanism in the publication bias has an effect both on the effect size distribution and the individual densities $\phi(x_i \mid \theta_i, \sigma_i^2)$.

An overview of all quantities used in this paper can be found in the Web Appendix A.

2.2 | The selection model

Before introducing the publication bias and the *p*-hacking models, let us define the *selection model* of which both models are instances. Consider the statistic x_i and its density $\phi(x_i)$, for the moment without dependencies on θ_i and σ_i^2. Let the *selection variable* s be a binary stochastic variable that equals 1 if and only if x_i is observed, for instance, if the paper containing x_i has been accepted by an editor. When the selection only depends on x_i, the density of our observed statistic is $f(x_i) = p(s = 1 \mid x_i)\phi(x_i)/p(s = 1)$. This is also known as a *weighted distribution* (Rao, 1985, eq. 3.1), and can be interpreted as a rejection sampling model (von Neumann, 1951).

The selection mechanism may depend on other quantities, such as the study-specific parameter θ_i and the study-specific nuisance parameter σ_i^2. We will see that the selection mechanism can be changed by conditioning on θ_i in the denominator, which is what we do with the *p*-hacking model in Section 2.4.

2.3 | The publication bias model

Imagine the publication bias scenario:

> Alice is an editor who receives a study with a *p*-value u_i. She knows her journals will suffer if she publishes many null-results, so she

> is disinclined to publish studies with large p-values. Still, she will publish any result with some p-value-dependent probability $w(u_i)$. Every study you will ever read in Alice's journal has survived this selection mechanism, the rest are lost forever.

In this story, the underlying model $\phi(x_i \mid \theta_i, \sigma_i^2)$ is transformed into a publication bias model

$$f(x_i \mid \theta_i, \sigma_i^2) \propto \phi(x_i \mid \theta_i, \sigma_i^2) w(u_i) \tag{1}$$

by the selection probability $w(u_i) \in [0, 1]$, which is a probability for each u_i. Here, u_i is a p-value that depends on x_i and maybe on something else, such as the standard deviation of x_i, but does not depend on θ_i. We can write the model using the selection variable s, as $w(u_i) = w(u_i(x_i, \sigma_i^2)) = p(s = 1 \mid x_i, \sigma_i^2)$. Note that $w(u_i)$ cannot depend on θ_i since the editor has no way of knowing the parameter θ_i; if she did, she would not have to look at the p-values at all. The normalizing constant of model (1) is finite for any probability $w(u_i)$, hence f is a bona fide density.

An argument against the publication bias scenario is that publication bias does not act only through p-values, but also through other features of the study such as language (Egger and Smith, 1998) and originality (Callaham *et al.*, 1998). While this is true, the publication bias scenario seems to completely capture the idea of p-value-based publication bias. Even if other sources of publication bias exist, maybe acting through x_i but not its p-value, publication bias based on p-values is a universally recognized problem, and a good place to start. The kind of model sketched here is almost the same as the one of Hedges (1992), with the sole exception that Hedges (1992) does not require $w(u_i)$ to be a probability. We demand it, as otherwise the intuitive publication bias scenario interpretation of the model disappears.

Even if we know the underlying $\phi(x_i \mid \theta_i, \sigma_i^2)$ of model (1), we will need to decide on what p-value to use. Usually, the p-value will be approximately a one-sided normal p-value, because most hypotheses have just one direction that is interesting. For instance, the effect of an antidepressant must be positive for the study to be publishable. A one-sided p-value can also be used if the researchers reported a two-sided value, since $p = 0.05$ for a two-sided hypothesis corresponds to $p = 0.025$ for a one-sided hypothesis. We will use the one-sided normal p-value in all examples in this paper.

Provided we know the underlying distributions and p-values u_i, we only need to decide on the selection probability to have a fully specified model. Hedges (1992) proposes the discrete selection probability

$$w(u_i \mid \rho, \alpha) = \sum_{j=1}^{J} \rho_j 1_{[\alpha_{j-1}, \alpha_j)}(u_i), \tag{2}$$

where α is a vector of cutoffs satisfying $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_J = 1$ and ρ is a non-negative vector with $\rho_1 = 1$. The interpretation of this selection probability is simple: When Alice reads the p-value u_i, she finds the j that makes u_i an element of $[\alpha_{j-1}, \alpha_j)$ and accepts the study with probability ρ_j. Related to this view, Hedges (1992) proposed to use the cutoffs 0.001, 0.005, 0.01 and 0.05, as these 'have particular salience for interpretation' (Hedges, 1992). In fact, a publication decision often depends on whether a p-value crosses the 0.05-threshold. Considering the bias-variance trade-off heuristic, we would prefer to only use one split point at 0.05, as done by Iyengar and Greenhouse (1988) in their second weight function. Other reasons to prefer one split are ease of interpretation and presentation. Despite this, only using 0.05 as a threshold for one-sided p-values is problematic, as many published results are calculated using a two-sided p-value instead. For this reason, it is useful to add a splitting point at 0.025. In our examples, we will use a two-step function selection probability $w(u_i \mid \rho) = 1_{[0,0.025)}(u_i) + \rho_2 1_{[0.025,0.05)}(u_i) + \rho_3 1_{[0.05,1]}(u_i)$, where the selection probability when $u_i \in [0, 0.025)$ is normalized to 1 to make the model identifiable. Nevertheless we present models in broad generality in order to allow for an arbitrary number of cutoffs J. This possibility is already implemented in the `R` package associated with this paper, `publipha` (Moss, 2020b).

The following proposition shows the densities of the one-sided normal step function selection probability publication bias models, with fixed effects and with normal random effects, respectively. Here, the notation $\phi_{[a,b)}(x \mid \theta, \sigma_i^2)$ indicates a normal truncated to $[a, b)$.

Proposition 1. *The density of an observation from a fixed effects one-sided normal step function selection probability publication bias model and parameters θ, σ_i^2, is*

$$f(x_i \mid \theta, \sigma_i^2) = \sum_{j=1}^{J} \pi_j^\star \phi_{(c_j, c_{j-1}]}(x_i \mid \theta, \sigma_i^2), \tag{3}$$

where $\pi_j^\star = \rho_j \frac{\Phi(c_{j-1} \mid \theta, \sigma_i^2) - \Phi(c_j \mid \theta, \sigma_i^2)}{\sum_{j=1}^{N} \rho_j [\Phi(c_{j-1} \mid \theta, \sigma_i^2) - \Phi(c_j \mid \theta, \sigma_i^2)]}$ is the probability that $x_i \in (c_j, c_{j-1}]$ and $c_j = \theta + \sigma_i \Phi^{-1}(1 - \alpha_j)$. Here, $u_i = 1 - \Phi(\frac{x_i - \theta}{\sigma_i})$, so $w(u_i \mid \rho) = \rho_j$ when $\alpha_{j-1} \le 1 - \Phi(\frac{x_i - \theta}{\sigma_i}) < \alpha_j$.

The density of an observation from the one-sided normal step function selection probability publication bias model

with normal random effects and parameters $\sigma_i^2, \theta_0, \tau$, is

$$f(x_i \mid \theta_0, \tau, \sigma_i^2) = \sum_{j=1}^{J} \pi_j^\star(\theta_0, \tau, \sigma_i^2)\phi_{(c_j, c_{j-1}]}(x \mid \theta_0, \tau^2 + \sigma_i^2), \tag{4}$$

where

$$\pi_j^\star(\theta_0, \tau, \sigma_i^2) = \rho_j \frac{\Phi(c_{j-1} \mid \theta_0, \tau^2+\sigma_i^2) - \Phi(c_j \mid \theta_0, \tau^2+\sigma_i^2)}{\sum_{j=1}^{J} \rho_j [\Phi(c_{j-1} \mid \theta_0, \tau^2+\sigma_i^2) - \Phi(c_j \mid \theta_0, \tau^2+\sigma_i^2)]}.$$

Proof. Consider the fixed effects model. By definition ,

$$f(x_i \mid \theta, \sigma_i^2) \propto \sum_{j=1}^{J} \rho_j 1_{[\alpha_{j-1}, \alpha_j)}(u_i)\phi(x_i \mid \theta_i, \sigma_i^2). \tag{5}$$

The normalizing constant is

$$\begin{aligned} &\sum_{j=1}^{J} \rho_j \int 1_{[\alpha_{j-1}, \alpha_j)}(u_i)\phi(x_i \mid \theta_i, \sigma_i^2)dx_i \\ &= \sum_{j=1}^{J} \rho_j \left[\Phi(c_{j-1} \mid \theta, \sigma_i^2) - \Phi(c_j \mid \theta, \sigma_i^2)\right]. \end{aligned} \tag{6}$$

Rewriting

$$\frac{\phi(x_i \mid \theta_i, \sigma_i^2)1_{[\alpha_{j-1}, \alpha_j)}(u_i)}{\Phi(c_{j-1} \mid \theta, \sigma_i^2) - \Phi(c_j \mid \theta, \sigma_i^2)} = \phi_{(c_j, c_{j-1}]}(x_i \mid \theta, \sigma_i^2), \tag{7}$$

we get Equation (3).

For the random effect model, we proceed similarly, see Web Appendix B. Note that $f(x_i \mid \theta_0, \tau, \sigma_i^2)$ does not equal $\int f(x_i \mid \theta_i, \sigma_i^2)\phi(\theta_i \mid \theta_0, \tau^2)d\theta_i$, as might have been expected. □

2.4 | The *p*-hacking model

Imagine the *p*-hacking scenario:

> Bob is an astute researcher who is able to *p*-hack any study to whatever level of significance he wishes. Whenever Bob does his research, he decides on a significance level to reach by drawing an α from a distribution ω. Then he *p*-hacks his study to this α-level, for example by excluding particular observations, collecting new data *ex-post*, or selectively excluding covariates.

In this scenario, the original density $\phi(x_i \mid \theta_i, \sigma_i^2)$ is transformed into the *p*-hacked density

$$f(x_i \mid \theta_i, \sigma_i^2) = \int_{[0,1]} \phi_\alpha(x_i \mid \theta_i, \sigma_i^2)\omega(\alpha)d\alpha, \tag{8}$$

where ϕ_α is ϕ truncated so that the *p*-value u_i lies inside $[0, \alpha]$, with $\alpha \in [0, 1]$. For instance, when using a one-sided *p*-value, we get that $\phi_\alpha(x_i \mid \theta_i, \sigma_i^2) = \phi_{[c_j, \infty)}(x_i \mid \theta_i, \sigma_i^2)$, where $c_j = \theta_i + \sigma_i\Phi^{-1}(1 - \alpha_j)$. As described in the *p*-hacking scenario, the *p*-hacking level α is drawn from a density $\omega(\alpha)$, which might depend on covariates. On the other hand, it should not depend on θ_i, as the researcher cannot know the true effect size of his study. While publication bias model (1) is a selection model, the *p*-hacking model (8) is a mixture model. Mathematically, the publication bias model could be written as a mixture model on the same form as the *p*-hacking model, but then ω would incorrectly depend on θ_i, see Web Appendix C. In our current setting, where the original distributions are Gaussian, it means that the publication bias model and the *p*-hacking model differs in the random effect case, while are equal in the fixed effect case (see Web Appendix D). Correspondingly, the *p*-hacking model could be written as a selection model (1), but the publication probability would then in general depend on the true effect size, which violates an obvious condition for a model to be considered a publication bias model. We stress therefore that the model (8) is not a publication bias model.

Just as the publication bias model requires a choice of w, the *p*-hacking model requires a choice of ω. A *p*-hacking scientist is motivated to *p*-hack to the 0.05 level, or may be to the levels 0.01 or 0.025, but never to a level such as 0.07 or 0.37. This motivates the discrete *p*-hacking probability distribution

$$\omega(\alpha \mid \pi) = \sum_{j=1}^{J} \pi_j 1(\alpha = \alpha_j) \tag{9}$$

for some j-ary vector of cutoffs α satisfying $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_J = 1$, and j-ary vector of probabilities π. In our example, it means that Bob will *p*-hack at a level α_1 with probability π_1, α_2 with probability π_2 and so on. The resulting density is

$$f(x_i \mid \theta_i, \sigma_i^2) = \sum_{j=1}^{J} \pi_j \phi_\alpha(x_i \mid \theta_i, \sigma_i^2)1(\alpha = \alpha_j). \tag{10}$$

Using reasoning analogous to that of Section 2.3, we suggest to use an ω only based on the two splitting points 0.025 and 0.05, that is, $\omega(\alpha \mid \pi) = \pi_1 1(\alpha = 0.025) + \pi_2 1(\alpha = 0.05) + \pi_3 1(\alpha = 1)$, but the model below is again presented using the more general J cutoffs.

The density of an observation from a fixed effects one-sided normal discrete probability *p*-hacking model is

$$f(x_i \mid \theta, \sigma_i^2) = \sum_{j=1}^{J} \pi_j \phi_{[c_j,\infty)}(x_i \mid \theta, \sigma_i^2), \tag{11}$$

where $c_j = \theta + \sigma_i \Phi^{-1}(1 - \alpha_j)$. Compared to the corresponding fixed effects publication bias model (3), where $f(x_i \mid \theta, \sigma_i^2) = \sum_{j=1}^{J} \pi_j^\star \phi_{(c_j, c_{j-1}]}(x_i \mid \theta, \sigma_i^2)$, the main difference are the truncation intervals: $[c_j, \infty)$ for the *p*-hacking model and $(c_j, c_{j-1}]$ for the publication bias model. We illustrate the fixed effects models further in Web Appendix D, where we also show that models are reparametrizations of each other. However, as already mentioned, the two models differ in the random effect case. In contrast to the publication bias model, there is no closed form for the density of the random effects variant of the one-sided normal discrete probability *p*-hacking model.

2.5 | The difference between the models in the case of random effects

In the random effects publication bias model, a completely new study is done whenever the last one failed to be published. In the event that $s = 0$ and the study fails to be published, a new effect size θ_i is sampled from the original effect size distribution $p(\theta_i)$, and then a new x_i from $N(\theta_i, \sigma_i^2)$. As a consequence, the modified effect size distributed $p^\star(\theta_i \mid \sigma_i^2)$ will generally not equal the original effect size distribution, as

$$p^\star(\theta_i \mid \sigma_i^2) = \int \frac{p(s = 1 \mid x_i, \sigma_i^2)}{p(s = 1 \mid \sigma_i^2)} f(x_i \mid \theta_i, \sigma_i^2) p(\theta_i) dx_i \neq p(\theta_i). \tag{12}$$

The dependence on σ_i^2 in $p^\star(\theta_i \mid \sigma_i^2)$ cannot be removed. In practice, the modified effect size distribution will be skewed towards favourable θ_is, as the selection mechanism of the publication bias model penalizes studies for which the effect sizes θ_is come from the least favourable part of the support of $p(\theta_i)$. Even if we somehow knew all the θ_is corresponding to our sample of x_is, the mean of these θ_is would be larger than the mean of the underlying effect size distribution. This implies that the modified effect size distribution $p^\star(\theta_i \mid \sigma_i^2)$ cannot be used directly to predict the value of a new draw from the true effect size distribution.

The *p*-hacking model does not modify the effect size distribution. The *p*-hacker will hack his study all the way to significance, regardless of θ_i. In this case, there will not be a new θ_i when $s = 0$: The *p*-hacker will modify the study until success ($s = 1$) given the sampled θ_i. The modified effect size distribution equals the original effect size distribution, that is,

$$p^\star(\theta_i \mid \sigma_i^2) = \int \frac{p(s = 1 \mid x_i, \sigma_i^2)}{p(s = 1 \mid \theta_i, \sigma_i^2)} f(x_i \mid \theta_i, \sigma_i^2) p(\theta_i) dx_i = p(\theta_i). \tag{13}$$

For an example that relates the selection biases to the two models, see Web Appendix E.

3 | SIMULATIONS

We want to answer these three questions about the *p*-hacking and publication bias models: (1) Do they work even in the absence of *p*-hacking and publication bias? Although we know these phenomena are ubiquitous and should always be corrected for, it is still important that the models do not distort the results when there is no publication bias or *p*-hacking. (2) How do they behave in extreme situations, in particular when n is small and the heterogeneity is large? (3) Does the *p*-hacking model work under the publication bias scenario and vice versa?

3.1 | Settings

We generate data under three scenarios: (i) With no publication bias nor *p*-hacking, using the normal random effect meta-analysis model. (ii) Under the presence of publication bias, using model (4). (iii) Under presence of *p*-hacking, using the random effects normal *p*-hacking model. The study-specific variances σ_i^2 are sampled uniformly from $\{20, \ldots, 80\}$. The size of the meta-analyses are $n = 5, 30, 100$, corresponding to small, medium and large meta-analyses, while the means for the effect size distribution are $0, 0.2, 0.8$. The value $\theta_0 = 0$ corresponds to no expected effect, while the positive θ_0s are the cutoffs for small and large effect sizes of Cohen (1988, pp. 24–27). The standard deviations of the random effects distributions are $\tau = 0.1$ and $\tau = 0.5$. While $\tau = 0.1$ is a reasonable amount of heterogeneity, $\tau = 0.5$ is a large amount of heterogeneity that provides a challenge for the models. The probability of acceptance of a paper is simulated to be 1 if the *p*-value is between 0 and 0.025, 0.7 if the *p*-value is between 0.025 and 0.05, and 0.1 otherwise. The *p*-hacking probabilities are 0.6, 0.3 and 0.1, for the threshold 0.025, 0.05 and 1, respectively.

In addition to the classical uncorrected model for meta-analysis, for each parameter combination we estimate the

TABLE 1 No publication bias, no *p*-hacking

True values			*p*-hacking model		Publication bias model		Classical model	
τ	θ_0	n	$\widehat{\theta_0}$	$\widehat{\tau}$	$\widehat{\theta_0}$	$\widehat{\tau}$	$\widehat{\theta_0}$	$\widehat{\tau}$
0.1	0	5	−0.02 (0.07)	0.17 (0.06)	−0.05 (0.07)	0.13 (0.05)	0.00 (0.08)	0.18 (0.07)
		30	−0.02 (0.03)	0.08 (0.03)	−0.02 (0.03)	0.07 (0.03)	0.00 (0.03)	0.10 (0.03)
		100	−0.01 (0.02)	0.08 (0.03)	−0.02 (0.02)	0.07 (0.03)	0.00 (0.02)	0.10 (0.03)
	0.2	5	0.12 (0.09)	0.21 (0.08)	0.09 (0.06)	0.17 (0.07)	0.20 (0.08)	0.19 (0.08)
		30	0.16 (0.03)	0.09 (0.04)	0.15 (0.03)	0.09 (0.04)	0.20 (0.03)	0.10 (0.04)
		100	0.18 (0.02)	0.09 (0.03)	0.17 (0.02)	0.09 (0.03)	0.20 (0.02)	0.10 (0.02)
	0.8	5	0.78 (0.09)	0.20 (0.10)	0.64 (0.14)	0.32 (0.13)	0.79 (0.08)	0.19 (0.08)
		30	0.80 (0.03)	0.10 (0.04)	0.80 (0.03)	0.11 (0.04)	0.80 (0.03)	0.10 (0.04)
		100	0.80 (0.02)	0.10 (0.03)	0.80 (0.02)	0.10 (0.03)	0.80 (0.02)	0.10 (0.02)
0.5	0	5	−0.05 (0.22)	0.57 (0.21)	−0.22 (0.19)	0.51 (0.20)	−0.01 (0.22)	0.59 (0.20)
		30	−0.03 (0.09)	0.52 (0.07)	−0.13 (0.10)	0.48 (0.07)	0.00 (0.09)	0.52 (0.07)
		100	−0.02 (0.05)	0.50 (0.04)	−0.08 (0.05)	0.48 (0.04)	0.00 (0.05)	0.50 (0.04)
	0.2	5	0.14 (0.22)	0.59 (0.20)	−0.08 (0.19)	0.56 (0.20)	0.19 (0.22)	0.59 (0.20)
		30	0.17 (0.09)	0.52 (0.07)	0.05 (0.09)	0.50 (0.08)	0.20 (0.09)	0.51 (0.07)
		100	0.18 (0.05)	0.51 (0.04)	0.11 (0.06)	0.50 (0.04)	0.20 (0.05)	0.50 (0.04)
	0.8	5	0.71 (0.24)	0.63 (0.21)	0.37 (0.25)	0.75 (0.21)	0.74 (0.23)	0.59 (0.21)
		30	0.77 (0.10)	0.53 (0.07)	0.59 (0.13)	0.60 (0.08)	0.79 (0.09)	0.51 (0.07)
		100	0.79 (0.05)	0.52 (0.04)	0.69 (0.07)	0.56 (0.05)	0.80 (0.05)	0.51 (0.04)

Note: Posterior means and, between brackets, the corresponding standard deviations for θ_0 and τ from the *p*-hacking and publication bias models when the data are simulated from the normal random effects meta-analysis model.

p-hacking model and the publication bias model using Bayesian methods. While a frequentist approach is in theory possible, it may lead to poor results if ad hoc penalizations or bias corrections are not implemented. See McShane *et al.* (2016, Appendix, 1) and Moss (2020a) for further details. All models have normal likelihoods and normal effect size distributions. We use one-sided significance cutoffs at 0.025 and 0.05 for both the publication bias and the *p*-hacking models. We use standard normal priors for θ_0, $\theta_0 \sim N(0,1)$, a standard half-normal prior for τ, that is $\tau = |\tau^\star|$, with $\tau^\star \sim N(0,1)$, and, in the *p*-hacking model, a uniform Dirichlet prior for π, $\pi \sim \text{Dir}(1)$. For the ρ in the publication bias model, we use a uniform Dirichlet that constrains $\rho_1 \geq ... \geq \rho_J$. That is, the publication probability is a decreasing function of the *p*-value.

All of these priors are reasonable. A standard normal for θ_0 is reasonable because we know that θ_0 has a small magnitude in pretty much any meta-analysis, and most are clustered around 0. A half-normal prior for τ is also reasonable, as τ is much more likely to be very small than very big. The priors for ρ and π are harder to reason about, but a uniform Dirichlet seems like a natural and neutral choice. These are the standard prior of the R package `publipha` (Moss, 2020b), which we used for all computations. `publipha` uses `STAN` (Carpenter *et al.*, 2017) to estimate the models, and each estimation uses eight chains. As suggested by a reviewer, we also tried different priors (uniform on [0,3] and the inverse gamma with shape = 2 and scale = 0.5) for τ, because this quantity is critical in Bayesian meta-analyses (see, e.g., Turner *et al.*, 2012). The results in Web Appendix F show that our models are robust to this choice.

The number of simulations is $N = 1000$ for each parameter combination. The code used to run the simulations is available in the Online Supporting Information and in an OSF repository (https://osf.io/tx8qn/).

3.2 | Results

No publication bias, no p*-hacking*

The results under this scenario are reported in Table 1. When the amount of heterogeneity is reasonable ($\tau = 0.1$) both the *p*-hacking and the publication bias perform well. The publication bias model performs slightly worse than the *p*-hacking model when the mean effect size is large ($\theta_0 = 0.8$) and the number of studies small ($n = 5$), but it catches up as n increases. With $\tau = 0.5$, the *p*-hacking model outperforms the publication bias model, with the latter tending to underestimate the mean effect. While increasing n alleviates the problem, there is still a substantial underestimation of θ_0 even in the case of $n = 100$. In contrast, both models seem to estimate τ pretty well. Obviously, without any publication bias or

TABLE 2 Publication bias

True values			*p*-hacking model		Publication bias model		Classical model	
τ	θ_0	n	$\widehat{\theta_0}$	$\widehat{\tau}$	$\widehat{\theta_0}$	$\widehat{\tau}$	$\widehat{\theta_0}$	$\widehat{\tau}$
0.1	0	5	−0.01 (0.11)	0.24 (0.08)	0.00 (0.08)	0.18 (0.07)	0.13 (0.10)	0.24 (0.09)
		30	0.02 (0.05)	0.12 (0.04)	0.01 (0.04)	0.10 (0.03)	0.14 (0.04)	0.16 (0.04)
		100	0.02 (0.03)	0.12 (0.03)	0.00 (0.03)	0.10 (0.02)	0.13 (0.02)	0.16 (0.02)
	0.2	5	0.14 (0.14)	0.28 (0.09)	0.11 (0.07)	0.21 (0.07)	0.33 (0.07)	0.15 (0.06)
		30	0.22 (0.05)	0.12 (0.05)	0.19 (0.06)	0.10 (0.04)	0.33 (0.03)	0.06 (0.03)
		100	0.24 (0.03)	0.10 (0.04)	0.20 (0.04)	0.09 (0.03)	0.33 (0.01)	0.04 (0.02)
	0.8	5	0.78 (0.09)	0.20 (0.08)	0.63 (0.15)	0.32 (0.13)	0.79 (0.08)	0.19 (0.07)
		30	0.80 (0.03)	0.10 (0.04)	0.80 (0.03)	0.10 (0.04)	0.80 (0.03)	0.10 (0.04)
		100	0.80 (0.02)	0.10 (0.02)	0.80 (0.02)	0.10 (0.02)	0.80 (0.02)	0.09 (0.02)
0.5	0	5	0.32 (0.21)	0.54 (0.21)	0.04 (0.22)	0.56 (0.19)	0.41 (0.18)	0.47 (0.23)
		30	0.36 (0.09)	0.47 (0.08)	0.01 (0.17)	0.50 (0.08)	0.43 (0.08)	0.42 (0.09)
		100	0.36 (0.05)	0.47 (0.05)	0.00 (0.11)	0.50 (0.05)	0.43 (0.04)	0.42 (0.05)
	0.2	5	0.46 (0.20)	0.52 (0.20)	0.16 (0.20)	0.58 (0.19)	0.54 (0.17)	0.44 (0.21)
		30	0.51 (0.09)	0.43 (0.08)	0.18 (0.17)	0.51 (0.09)	0.56 (0.07)	0.38 (0.08)
		100	0.51 (0.05)	0.43 (0.05)	0.18 (0.12)	0.50 (0.05)	0.56 (0.04)	0.38 (0.04)
	0.8		0.84 (0.20)	0.54 (0.20)	0.49 (0.25)	0.71 (0.21)	0.87 (0.18)	0.50 (0.19)
		30	0.91 (0.08)	0.45 (0.07)	0.67 (0.19)	0.57 (0.12)	0.93 (0.08)	0.42 (0.06)
		100	0.91 (0.05)	0.44 (0.04)	0.75 (0.11)	0.53 (0.07)	0.92 (0.04)	0.41 (0.04)

Note: Posterior means and, between brackets, the corresponding standard deviations for θ_0 and τ from the *p*-hacking and publication bias models when the data are simulated from the publication bias model with cutoffs at 0.025 and 0.05, with selection probabilities equal to 1, 0.7 and 0.1 in the intervals $[0, 0.025)$, $[0.025, 0.05)$ and $[0.5, 1]$.

p-hacking, the classical uncorrected model gives good results.

Publication bias

Overall, the publication bias model outperforms the *p*-hacking model when the data are generated from the publication bias model, but not by much (see Table 2). When $\tau = 0.5$, the *p*-hacking model tends to overestimates θ_0 while the publication bias model tends to underestimate it. The overestimation of the *p*-hacking model is most extreme when $\theta_0 = 0.2$, but not as strong as the classical uncorrected model. When $\tau = 0.1$, the publication bias and *p*-hacking models produce almost indistinguishable results, outperforming the uncorrected model (especially if the effect θ_0 is null or small). Just as in the *p*-hacking scenario, both models estimate τ reasonably well.

p-*hacking*

The simulation results for the *p*-hacking model are in Table 3. As before, the largest differences are in the most difficult case of $\tau = 0.5$, while the two models tend to agree in the more realistic case of $\tau = 0.1$. When $\tau = 0.5$, the publication bias model severely underestimates θ_0, even getting the sign wrong in some instances. This should not come as a surprise given the interpretation of θ_0 in the publication bias model, but shows that we should be cautious in interpreting the θ_0 estimates. In basically all cases, the *p*-hacking model outperforms the uncorrected model, with the latter surprisingly working better than the publication bias model when the effect size θ_0 is large (0.8).

4 | EXAMPLES

In this section, we apply the models on the two meta-analyses of Cuddy *et al.* (2018) and Anderson *et al.* (2010). As in the simulation study, we use normal models for each effect size with one-sided significance cutoff at 0.025 and 0.05 for both models. We use the same priors as we did in the simulation study. To compare the fit of the models, we use the leave-one-out cross-validation information criterion (LOOIC) (Vehtari *et al.*, 2017), calculated using the R package `loo` (Vehtari *et al.*, 2018). LOOIC equals $-2 \cdot \text{ELPD}_{\text{LOO}}$, where ELPD_{LOO} is a leave-one-out cross validation-based estimate of ELPD. In turn, ELPD is the expected log pointwise predictive density for a new data set. Just as the AIC, smaller values indicate better model fit. As for the simulation study, the analyses have been done with the `R` package `publipha` (Moss, 2020b), which in turn uses `STAN` (Carpenter *et al.*, 2017). Each model has been estimated with eight chains. The code used to run the examples can be found in the Online Supporting

TABLE 3 *p*-hacking

True values			*p*-hacking model		Publication bias model		Classical model	
τ	θ_0	n	$\widehat{\theta_0}$	$\widehat{\tau}$	$\widehat{\theta_0}$	$\widehat{\tau}$	$\widehat{\theta_0}$	$\widehat{\tau}$
0.1	0	5	−0.06 (0.15)	0.28 (0.06)	0.05 (0.06)	0.17 (0.05)	0.29 (0.06)	0.15 (0.08)
		30	−0.01 (0.07)	0.13 (0.05)	0.02 (0.06)	0.07 (0.03)	0.29 (0.02)	0.05 (0.03)
		100	0.00 (0.04)	0.10 (0.04)	−0.01 (0.04)	0.05 (0.02)	0.28 (0.01)	0.03 (0.02)
	0.2	5	0.12 (0.14)	0.28 (0.09)	0.10 (0.06)	0.20 (0.06)	0.35 (0.05)	0.13 (0.05)
		30	0.19 (0.06)	0.12 (0.05)	0.16 (0.06)	0.09 (0.04)	0.34 (0.02)	0.04 (0.02)
		100	0.20 (0.03)	0.09 (0.04)	0.16 (0.05)	0.08 (0.03)	0.34 (0.01)	0.02 (0.01)
	0.8	5	0.78 (0.09)	0.20 (0.09)	0.63 (0.15)	0.32 (0.13)	0.79 (0.08)	0.19 (0.07)
		30	0.80 (0.03)	0.10 (0.04)	0.80 (0.03)	0.10 (0.04)	0.80 (0.03)	0.09 (0.04)
		100	0.80 (0.02)	0.09 (0.03)	0.80 (0.02)	0.10 (0.03)	0.80 (0.02)	0.09 (0.02)
0.5	0	5	0.07 (0.22)	0.48 (0.19)	0.00 (0.15)	0.37 (0.20)	0.36 (0.12)	0.29 (0.20)
		30	0.07 (0.10)	0.43 (0.08)	−0.25 (0.20)	0.35 (0.10)	0.36 (0.05)	0.24 (0.10)
		100	0.06 (0.05)	0.44 (0.04)	−0.36 (0.14)	0.37 (0.06)	0.37 (0.03)	0.25 (0.05)
	0.2	5	0.20 (0.22)	0.52 (0.20)	0.04 (0.14)	0.45 (0.21)	0.43 (0.13)	0.33 (0.19)
		30	0.24 (0.10)	0.47 (0.08)	−0.19 (0.20)	0.46 (0.10)	0.45 (0.06)	0.29 (0.08)
		100	0.23 (0.05)	0.47 (0.04)	−0.29 (0.16)	0.47 (0.06)	0.45 (0.03)	0.28 (0.04)
	0.8	5	0.73 (0.22)	0.61 (0.20)	0.36 (0.24)	0.74 (0.21)	0.80 (0.18)	0.52 (0.18)
		30	0.80 (0.10)	0.50 (0.08)	0.39 (0.24)	0.66 (0.12)	0.85 (0.08)	0.43 (0.06)
		100	0.80 (0.05)	0.50 (0.04)	0.43 (0.19)	0.64 (0.09)	0.85 (0.04)	0.43 (0.03)

Note: Posterior means and standard deviations from the *p*-hacking and publication bias models when the data are simulated from the *p*-hacking model with cutoffs at 0.025 and 0.05, with *p*-hacking probabilities equal to 0.6, 0.3 and 0.1 for α equal to 0.025, 0.05 and 1

Information and in an OSF repository (https://osf.io/tx8qn/).

4.1 | Power posing

Cuddy *et al.* (2018) conducted a meta-analysis of the effect of power posing, an alleged phenomenon where adopting expansive postures has positive psychological feedback effects. Their meta-analysis is not conventional, but a *p*-curve analysis (Simonsohn *et al.*, 2014a). A *p*-curve analysis is not based on estimated effect sizes and standard errors, but directly on *p*-values. The data from Cuddy *et al.* (2018) can be accessed via the Open Science Framework (https://osf.io/pfh6r/). Here, we only consider studies with outcome 'mean difference', design '2 cell', and test statistic that is either F or t. The F-statistics are all with 1 denominator degree of freedom, and the root of these are distributed as the absolute value of a t-distributed variable. The t-values and the roots of the F-statistics are converted to standardized mean differences by using $d = t\sqrt{2/\nu}$, where ν is the degrees of freedom for the t-test. The standardized mean differences are to the left in Figure 1. Note the outlier $x_{12} = 1.72$. As it has a large effect on all the models, we analyse the data both with and without x_{12}.

The estimates of the *p*-hacking model, the publication bias model and the uncorrected meta-analysis models are in Table 4. According to the LOOIC, the corrected models account much better for the data than the uncorrected model. Both the *p*-hacking model and the publication bias models estimate larger τs and smaller θ_0s than the classical model, with the publication bias model estimating the surprising $\theta_0 \approx 0$. But recall the results of the simulation study, where the publication bias model severely underestimates θ_0 when the *p*-hacking model is true.

The publication bias selection affects not only the observed x_is, but also the θ_is. As a consequence, the posterior mean of the selected effect size distribution (this equals 0.37, is not shown in the table, and equals the average of the posterior means for the θ_is) is much closer to the uncorrected model's estimate than the *p*-hacked estimate. This effect can be most easily understood by looking at a specific θ, for example, the θ_2 reported in the right plot of Figure 1, where $x_2 = 0.62$. In this case, the publication bias posterior for is close to the uncorrected posterior even though $\theta_0 \approx 0$. On the other hand, the *p*-hacking model pushes 0.62 down to 0.17, towards the meta-analytic mean of 0.18.

Finally, the surprisingly low value for θ_0 obtained with the publication bias model can be a side effect of the presence of the outlier $x_{12} = 1.72$. Its presence on the right tail of an hypothetical true effect size distribution implies unobserved low and negative effects not reported due to publication bias. When the outlier is removed from the

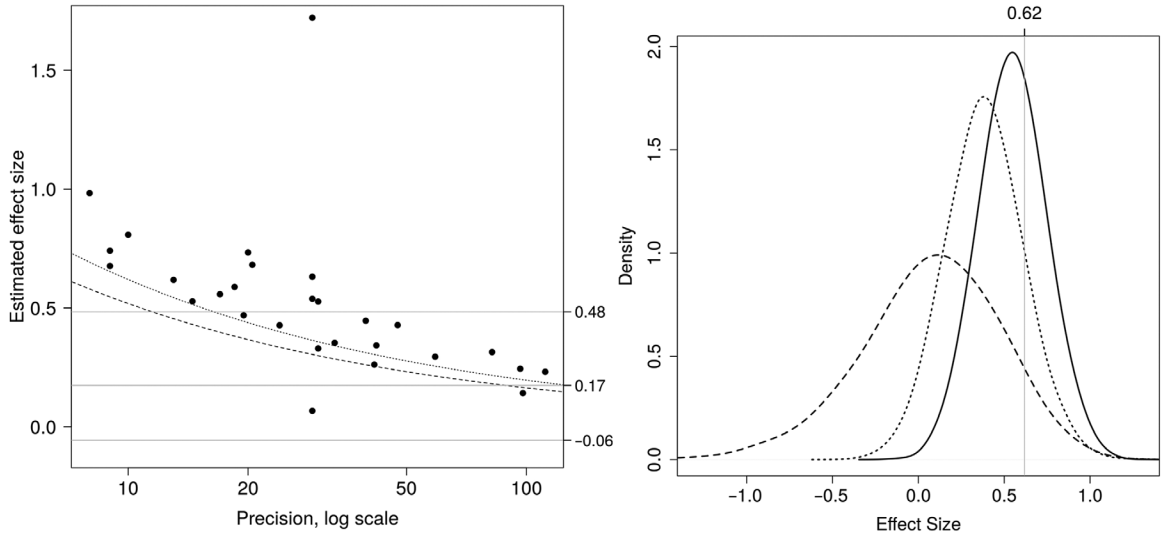

FIGURE 1 (Left) Effect sizes for the power posing example. The dotted black line is 1.96/sd and the dashed black line is 1.64/sd. The ticks on the right-hand side are the meta-analytic means: 0.48 is from the uncorrected model, 0.17 is the mean of the selected effect size distribution under the *p*-hacking model, while −0.06 is the mean under the publication bias model. (Right) Posterior densities for θ_2 in the power posing example. The dashed density belongs to the *p*-hacking model, the dotted density to the publication bias model and the solid density to the uncorrected model. The point $x_2 = 0.62$ is marked for reference

TABLE 4 Power posing example

All studies					
	LOOIC	θ_0	τ	π_1/ρ_1	π_2/ρ_2
Uncorrected	16 (18)	0.48 (0.07)	0.27 (0.06)		
p-hacking	−18 (14)	0.18 (0.12)	0.45 (0.10)	0.62 (0.15)	0.23 (0.14)
Publication bias	−5.1 (22)	−0.06 (0.23)	0.37 (0.09)	0.39 (0.22)	0.03 (0.03)
Without outlier					
	LOOIC	θ_0	τ	π_1/ρ_1	π_2/ρ_2
Uncorrected	−7.1 (5.7)	0.39 (0.04)	0.09 (0.05)		
p-hacking	−38 (10)	0.18 (0.07)	0.09 (0.07)	0.62 (0.15)	0.24 (0.15)
Publication bias	−35 (11)	0.16 (0.09)	0.08 (0.06)	0.26 (0.17)	0.03 (0.03)

Note: Posterior means for LOOICs and parameters (mean effect θ, standard deviation τ, probabilities of *p*-hacking π/probabilities of being published ρ) of the *p*-hacking, publication bias and classical meta-analysis (uncorrected) model estimated on the data by Cuddy *et al.* (2018). The results in the top table are obtained with all studies, those in the bottom without the outlier x_{12}. Posterior standard deviations are reported between brackets

analysis, the estimate of θ_0 goes up and agrees with the estimate from the *p*-hacking model, which does not change. Once the outlier is removed, the fit of the publication bias model increases tremendously, reaching a level close to that of the *p*-hacking model. Moreover, the estimates of τ are strongly affected by the removal of x_{12}. In particular, the estimate of τ decreases from 0.45 to 0.09 in the *p*-hacking model.

In conclusion, the *p*-hacking and publication bias models suggest there is selection bias in these studies. Both models have much better fit than the uncorrected one and it is reasonable to accept their parameter estimates as more realistic. Nonetheless, both models agree on a value of θ_0 that is likely to be different from 0. The results of Table 4 supports Cuddy *et al.* (2018)'s conclusion that there is evidence for some positive effect of power posing. The *p*-hacking model does not suffer the presence of an outlier, and, in contrast to the publication bias model, provides similar results with and without x_{12} in the data.

4.2 | Violent video games

Anderson *et al.* (2010) conducted a large meta-analysis of the effects of violent video games on seven negative outcomes such as aggressive behaviour and aggressive cognition. As part of their analysis, they classified some experiments as best practice experiments (for more details,

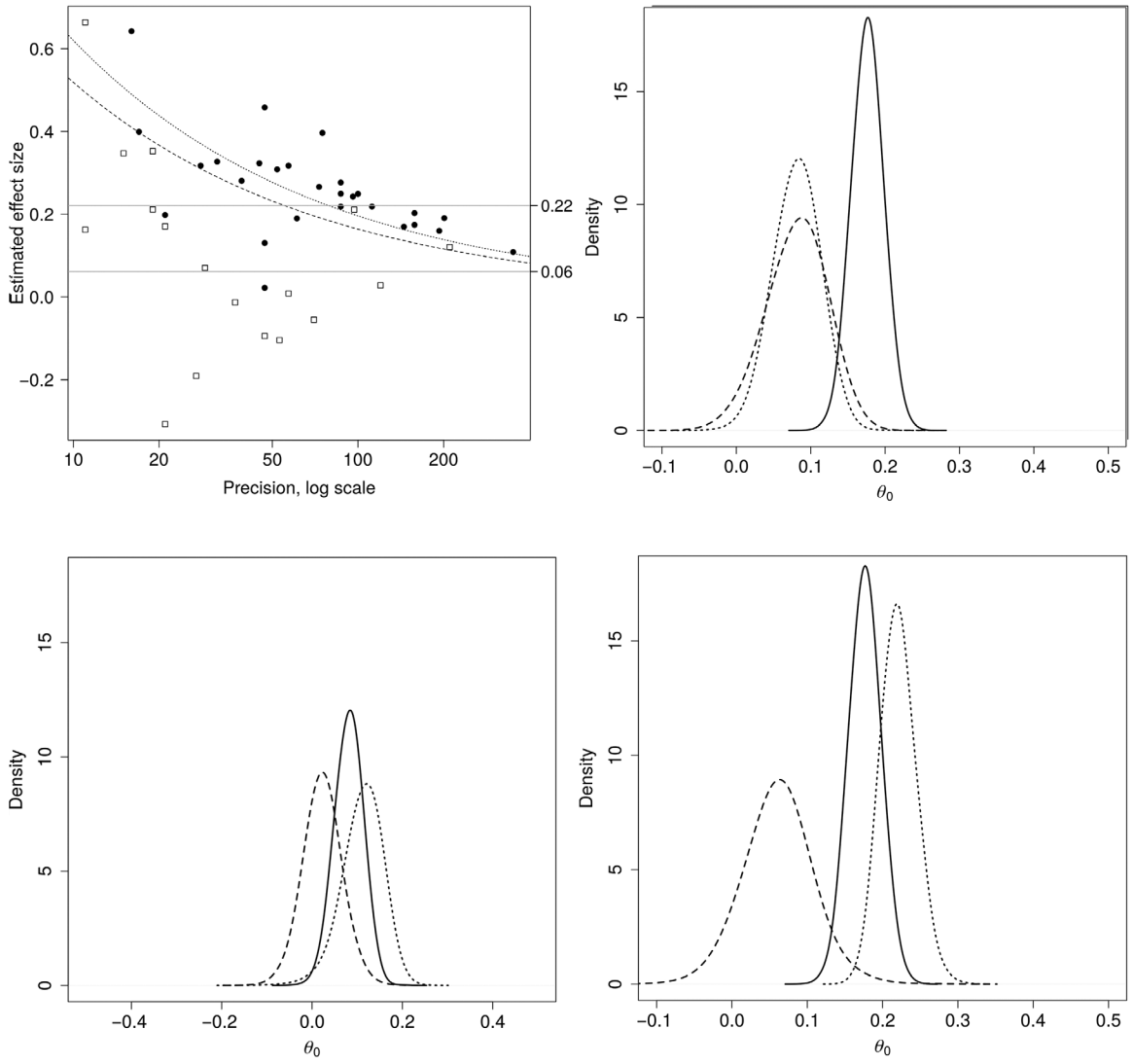

FIGURE 2 Violent video games example with outcome variable aggressive behaviour. (Top-left) Effect sizes. The dotted black line is 1.96/sd and the dashed black line is 1.64/sd. The ticks on the right-hand side are the uncorrected meta-analytical means for each group: 0.29 for the best practices group, 0.08 for the rest. The outlier $x = 1.33$ has been removed from the plot. (Top-right) Posterior densities for θ_0 with all experiments included. The dashed density belongs to the *p*-hacking model, the dotted to the publication bias model and the solid to the uncorrected model. (Bottom-left) Posterior densities for θ_0 from the publication bias model. The solid curve is the model with all experiments, the dotted curve the model with the best practice experiments and the dashed line the model without the best experiments. The posteriors for the *p*-hacking model are similar to this one. (Bottom-right) Posterior densities for θ_0 (solid line: all experiments; dotted line: best practice experiments only and dashed line without the best experiments) from the uncorrected meta-analysis model

see tab. 2 of Anderson *et al.*, 2010). Suspecting publication bias, Hilgard *et al.* (2017) reanalysed the data using an array of tools to detect and adjust for publication bias. For the outcome variable aggressive cognition, Hilgard *et al.* (2017) noted that 'Application of best-practices criteria seems to emphasize statistical significance, and a knot of experiments just reach statistical significance'. The data can be found on the web (Hilgard, 2017) and are visualized to the top left in Figure 2. In the plot, the best practice experiments are represented by solid circles, all other experiments by hollow squares. An outlier $x = 1.33$ has been removed from the data set, and excluded from our analyses. Its removal substantially improves the fit for all the models.

In this example, we fit the three models (*p*-hacking, publication bias and uncorrected models) to three data subsets (all experiments, only best practice experiments, without best practice experiments). The outcome variable is aggressive behaviour. Our aim is to answer the following: (1) What are the parameter estimates, in each subset, for each model? (2) Which model has the best fit? (3) Do we have a reason to believe the best practice experiments are drawn from a different underlying distribution than the other experiments, as Hilgard *et al.* (2017) and the top

TABLE 5 Violent video games example

All experiments					
	LOOIC	θ_0	τ	π_1/ρ_1	π_2/ρ_2
Uncorrected	−38 (11)	0.18 (0.02)	0.04 (0.03)		
p-hacking	−48 (13)	0.09 (0.04)	0.05 (0.04)	0.25 (0.11)	0.23 (0.11)
Publication bias	−54 (13)	0.08 (0.03)	0.03 (0.02)	0.44 (0.18)	0.13 (0.07)
Only best practice experiments					
	LOOIC	θ_0	τ	π_1/ρ_1	π_2/ρ_2
Uncorrected	−42 (6.2)	0.22 (0.02)	0.03 (0.02)		
p-hacking	−59 (12)	0.10 (0.05)	0.06 (0.04)	0.37 (0.17)	0.41 (0.17)
Publication bias	−61 (11)	0.11 (0.04)	0.03 (0.02)	0.46 (0.21)	0.06 (0.05)
Without best practice experiments					
	LOOIC	θ_0	τ	π_1/ρ_1	π_2/ρ_2
Uncorrected	−7.4 (5.7)	0.06 (0.04)	0.08 (0.05)		
p-hacking	−6.2 (5.1)	0.01 (0.05)	0.07 (0.05)	0.10 (0.07)	0.11 (0.08)
Publication bias	−7.7 (5)	0.02 (0.04)	0.06 (0.04)	0.61 (0.23)	0.35 (0.19)

Note: Posterior means for LOOICs and parameters (mean effect θ, standard deviation τ, probabilities of *p*-hacking π/probabilities of being published ρ) of the *p*-hacking, publication bias and classical meta-analysis (uncorrected) model estimated on the aggressive behaviour data from Anderson *et al.* (2010). Posterior standard deviations are reported between brackets.

left plot of Figure 2 suggest? (4) Is there a large difference between the posterior for θ_0 and the mean posterior for the θ_is, as we saw in the previous example?

The first three questions can be answered by looking at Table 5. The estimates of θ_0 are approximately the same for the publication bias and *p*-hacking models, and roughly half of the uncorrected estimate in all cases. In particular, when all experiments or only the best experiments are considered, there is a noticeable difference. In these two cases, the LOOICs suggest that some *p*-hacking or publication bias is present, as they are smaller than the LOOIC for the uncorrected models. Although the publication bias model seems to work slightly better than the *p*-hacking model, we can state that the two models agree and we have little reason to prefer one to the other. Basically, we can interpret this as converging evidence that the parameter estimates obtained with these two models for θ_0 and τ are in the ballpark of their true values.

Interestingly, when we exclude the experiments not considered best practice by Anderson *et al.* (2010), the differences between the estimates provided by the corrected and uncorrected models reduce and the LOOICs are almost the same. The question is if the differences between best practice and non-best practice studies reflect a different underlying distribution or not. To answer this question, let us take a look at the posterior densities for θ_0 when all experiments are included, as reported in the top right plot of Figure 2. In this case, the posterior distributions computed with the *p*-hacking and publication bias models are similar (dashed and dotted lines, respectively), which strengthens the agreement seen in Table 5. There is no large difference between the posterior for θ_0 and the mean posterior for the θ_is as in the previous example. The answer to question (4) is therefore no.

Back to question (3), we have good reasons to believe the best practice experiments have been drawn from a different underlying distribution than the other experiments if there is negligible overlap between the posteriors for the parameters θ_0. The uncorrected model supports this hypothesis (bottom right plot of Figure 2), but the *p*-hacking and publication bias models to do not. See the bottom left plot of Figure 2 for the posteriors for θ_0 in the publication bias model (those obtained with the *p*-hacking model are indistinguishable). In this case, the overlap between the posteriors for the different subsets is not negligible, and there is no evidence against hypotheses of equal θ_0s in both groups. The same conclusion can be reached from Table 5 by looking at the posterior standard deviations and posterior means.

5 | CONCLUDING REMARKS

In this paper, we studied two models to handle the effect of *p*-hacking and publication bias. Although the *p*-hacking model worked really well in the simulation study, we have to admit that the *p*-hacking scenario described in Section 2.4 is less plausible than the publication bias scenario of Section 2.3. The assumption of Bob's *p*-hacking omnipotence is strong. For while some researchers are able *p*-hackers, most give up at some point. Does truncation actually model *p*-hacking in the wild? Analysing *p*-hacking is hard without serious simplifying assumptions. The model we proposed is interpretable and

implementable, and it appears to work well in practice, as one can see in the examples of Section 4.

In this paper, we only considered normal densities, but the theory holds more generally. A remaining concern is identifiability, but we show in Web Appendix G that the publication bias and the *p*-hacking models are identifiable under weak conditions on f.

We are often interested in understanding and modelling the sources of heterogeneity in a meta-analysis (Thompson, 1994). A way to do this is to let θ_i linearly depend on covariates, in the meta-analysis context known as moderators. If we extend the one-sided discrete models publication bias and *p*-hacking models to include covariates, we will be able to estimate their effect while keeping the *p*-hacking probability or the selection probability fixed, as done by, for example, Vevea and Hedges (1995) in the publication bias model. Another option is to allow the *p*-hacking probability or the selection probability to depend on covariates themselves. For instance, the difficulty of *p*-hacking is likely to increase with n, the sample size of the study. Similarly, the selection probability is also likely to be influenced by n; for example, when n is large, null-effects are more publishable.

Although the common practice in meta-analysis studies is to treat the standard deviations as nuisance parameter, the actual tests usually contain an estimate of the standard error and this can also influence the selection mechanism. Further modifications to the models can be obtained by allowing for this.

We saw in the simulations and in Example 4.1 that the publication bias and the *p*-hacking models can give remarkably different results even with similar priors and the same cutoff vector. A way to react to this situation is to choose the best-fitting model in terms of, for example, LOOIC. To be safe, one can present the results of both models and try to understand the differences between them, as we did in the examples of Section 4. In the publication bias model, it is especially important to be aware of the interpretation of θ_0 as the mean of the underlying effect size distribution, not the effect size distribution of the observed studies. Therefore, the best response to the question 'Should one use the *p*-hacking and publication bias model?' is probably 'Use both!'

Finally, it would be interesting to model publication bias and *p*-hacking at the same time:

> Bob *p*-hacks his research to a *p*-value drawn from ω and sends it to Alice's journal. Alice accepts the paper with probability $w(u_i)$. Every rejected study is lost.

In this scenario, the original density $\phi(x_i \mid \theta_i, \sigma_i^2)$ is transformed twice: First by *p*-hacking, then by publication bias. The resulting model is

$$f(x_i \mid \theta_i, \sigma_i^2) \propto w(u_i) \int_{[0,1]} \phi_\alpha(x_i \mid \theta_i, \sigma_i^2) d\omega(\alpha). \quad (14)$$

This is a reasonable model, but its normalizing constant is hard to calculate, even when ω is discrete and w is a step function. Additional work on this problem is required.

ACKNOWLEDGEMENTS

The authors thank the Associate Editor and the Referee for their suggestions, which led to an improved version of the paper.

OPEN RESEARCH BADGES

This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at https://osf.io/tx8qn/.

DATA AVAILABILITY STATEMENT

The data that support the findings of this paper are available in Open Science Framework at https://osf.io/tx8qn/ and in the R package `publipha`. These data were derived from the following resources available in the public domain: https://osf.io/tx8qn/ (Power posing) and https://github.com/Joe-Hilgard/Anderson-meta (Violent video games).

ORCID

Jonas Moss https://orcid.org/0000-0002-6876-6964
Riccardo De Bin https://orcid.org/0000-0002-7441-6880

REFERENCES

Anderson, C.A., Shibuya, A., Ihori, N., Swing, E.L., Bushman, B.J., Sakamoto, A., et al. (2010) Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: a meta-analytic review. *Psychological Bulletin*, 136, 151–173.

Becker, B.J. (2005) Failsafe *N* or file-drawer number. In: *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Chichester: Wiley, pp. 111–125.

Boulesteix, A.-L. (2009) Over-optimism in bioinformatics research. *Bioinformatics*, 26, 437–439.

Bruns, S.B. and Ioannidis, J.P. (2016) P-curve and p-hacking in observational research. *PLoS ONE*, 11, e0149144.

Callaham, M.L., Wears, R.L., Weber, E.J., Barton, C. and Young, G. (1998) Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting. *Journal of American Medical Association*, 280, 254–257.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017) STAN: a probabilistic programming language. *Journal of Statistical Software*, 76, 1–32.

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

Citkowicz, M. and Vevea, J.L. (2017) A parsimonious weight function for modeling publication bias. *Psychological Methods*, 22, 28–41.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Mahwah: Lawrence Erlbaum Associates.

Copas, J. and Shi, J.Q. (2000) Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics*, 1, 247–262.

Cuddy, A.J., Schultz, S.J. and Fosse, N.E. (2018) P-curving a more comprehensive body of research on postural feedback reveals clear evidential value for power-posing effects: reply to Simmons and Ssimonsohn (2017). *Psychological Science*, 29, 656–666.

Dear, K.B.G. and Begg, C.B. (1992) An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*, 7, 237–245.

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

Egger, M. and Smith, G.D. (1998) Meta-analysis bias in location and selection of studies. *BMJ: British Medical Journal*, 316, 61–66.

Friese, M. and Frankenbach, J. (2020) p-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, 25, 456–471.

Hedges, L.V. (1984) Estimation of effect size under nonrandom sampling: the effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, 9, 61–85.

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

Hilgard, J. (2017) Anderson-meta. GitHub repository. Available at: https://github.com/Joe-Hilgard/Anderson-meta. Accessed September 20, 2021.

Hilgard, J., Engelhardt, C.R. and Rouder, J.N. (2017) Overstated evidence for short-term effects of violent games on affect and behavior: a reanalysis of Anderson and others (2010). *Psychological Bulletin*, 143, 757–774.

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

Marks-Anglin, A. and Chen, Y. (2020) A historical review of publication bias. *Research Synthesis Methods*, 11, 725–742.

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, 730–749.

Moss, J. (2020a) Infinite diameter confidence sets in hedges' publication bias model. *arXiv preprint arXiv:1912.09180*.

Moss, J. (2020b) *publipha: Bayesian Meta-Analysis with Publications Bias and P-Hacking*. R package version 0.1.1.

Rao, C.R. (1985) Weighted distributions arising out of methods of ascertainment: What population does a sample represent? In: *A Celebration of Statistics*. New York: Springer, pp. 543–569.

Rosenthal, R. (1979) The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638–641.

Rothstein, H.R., Sutton, A.J. and Borenstein, M. (2006) *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Chichester: Wiley.

Sijtsma, K. (2016) Playing with data—or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika*, 81, 1–15.

Simmons, J.P., Nelson, L.D. and Simonsohn, U. (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366.

Simonsohn, U., Nelson, L.D. and Simmons, J.P. (2014a) P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, 143, 534–547.

Simonsohn, U., Nelson, L.D. and Simmons, J.P. (2014b) p-Curve and effect size: correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9, 666–681.

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

Sterling, T.D. (1959) Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, 54, 30–34.

Thompson, S.G. (1994) Systematic review: why sources of heterogeneity in meta-analysis should be investigated. *BMJ: British Medical Journal*, 309, 1351–1355.

Turner, R.M., Davey, J., Clarke, M.J., Thompson, S.G. and Higgins, J.P. (2012) Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane database of systematic reviews. *International Journal of Epidemiology*, 41, 818–827.

van Assen, M.A., van Aert, R. and Wicherts, J.M. (2015) Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, 20, 293–309.

van Houwelingen, H.C., Arends, L.R. and Stijnen, T. (2002) Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21, 589–624.

Vehtari, A., Gabry, J., Yao, Y. and Gelman, A. (2018) loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R Package Version 2.0.0.

Vehtari, A., Gelman, A. and Gabry, J. (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27, 1413–1432.

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

von Neumann, J. (1951) Various techniques used in connection with random digits. *Applied Math Series*, 12, 36–38.

SUPPORTING INFORMATION

The R code (including the data) to reproduce the results and the Web Appendices referenced in Sections 2, 3 and 4 are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Moss, J., De Bin, R. Modelling publication bias and *p*-hacking. *Biometrics*. 2021;1–13. https://doi.org/10.1111/biom.13560