Correcting for Publication Bias in a Meta-Analysis with the *P*-uniform* Method

Robbie C. M. van Aert¹ and Marcel A. L. M. van Assen^{1,2}

¹Tilburg University ²Utrecht University

Author Note

The preparation of this article was supported by Grant 406-13-050 from the Netherlands Organization for Scientific Research (NWO) and European Research Council. Grant Number: 726361 (IMPROVE). A previous version of this work was presented at the International Meeting of the Psychometric Society 2018.

Correspondence concerning the article should be addressed to Robbie C. M. van Aert, Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, the Netherlands

Abstract

Publication bias is a major threat to the validity of a meta-analysis resulting in overestimated effect sizes. We propose an extension and improvement of the publication bias method *p*-uniform called *p*-uniform*. *P*-uniform* improves upon *p*-uniform in three ways, as it (i) entails a more efficient estimator, (ii) eliminates the overestimation of effect size caused by between-study variance in true effect sizes, and (iii) enables estimating and testing for the presence of the between-study variance. We compared the statistical properties of *p*-uniform* with *p*-uniform, two implementations of the three-parameter selection model (3PSM) approach, and the random-effects model. Statistical properties of *p*-uniform* and 3PSM were comparable and generally outperformed *p*-uniform and the random-effects model if publication bias was present. We demonstrate that *p*-uniform* and 3PSM estimate average effect size and between-study variance rather well with ten or more studies in the meta-analysis when publication bias is not extreme. We offer recommendations for applied researchers, provide an R package as well as an easy-to-use web application for applying *p*-uniform*.

Keywords: publication bias, meta-analysis, *p*-uniform, selection model approach

Correcting for Publication Bias in a Meta-Analysis with the *P*-uniform* Method

Effect sizes from multiple primary studies can be statistically combined by means of a meta-analysis in order to obtain a quantitative summary of the studied relationship. A major threat to the validity of a meta-analysis is publication bias (Borenstein, Hedges, Higgins, & Rothstein, 2009; Rothstein, Sutton, & Borenstein, 2005). Publication bias refers to situations where the published literature is not a representative reflection of the population of completed studies (Rothstein et al., 2005). In its most extreme case this implies that studies with statistically significant results in the expected direction get published and studies with statistically nonsignificant results do not get published. Publication bias is not only caused by reviewers and editors who are reluctant to accept studies without statistically significant results, but also by researchers who do not submit studies with nonsignificant results (Cooper, DeNeve, & Charlton, 1997; Coursol & Wagner, 1986). The consequences of publication bias are severe and may hamper scientific progress, because publication bias results in overestimated effect sizes in primary studies and combining these overestimated effect sizes in a meta-analysis yields an overestimated average effect size as well (Kraemer, Gardner, Brooks, & Yesavage, 1998; Lane & Dunlap, 1978).

Evidence for publication bias has been observed in multiple research fields. Fanelli (2010, 2012) studied how often the authors declared to have found support for the tested hypothesis in a random sample of published papers from a variety of research fields. In psychiatry and psychology, 90% of the papers concluded that the hypothesis was supported which was the largest percentage across all included research fields. However, this large percentage is not in line with the estimated average statistical power of approximately 50% (or lower) in psychological research (Bakker, van Dijk, & Wicherts, 2012; Cohen, 1990). Other more direct evidence of publication bias in psychology was found in Franco, Malhotra, and Simonovits (2014). They

studied whether experiments in the social sciences were more likely to be published if the authors of these experiments deemed the results to be strong, mixed, or null. Franco et al. (2014) observed evidence that experiments with null or mixed results were less often published compared to strong results.

Although the evidence for publication bias in multiple research fields is strong, we want to emphasize that publication bias is not omnipresent. For instance, a meta-analysis of meta-analyses (i.e., meta-meta-analysis) did not observe publication bias in meta-analyses about posttraumatic stress disorder (Niemeyer et al., 2020) and another meta-meta-analysis concluded that there was only weak evidence of mild publication bias in meta-analyses published in research fields of psychology and medicine (van Aert, Wicherts, & van Assen, 2019). Not observing strong evidence for publication bias in these meta-meta-analyses may be explained by effects being included in these meta-analyses that were not the primary outcome of a study. Publication bias is expected to work on the primary outcome of a study and less on secondary outcomes.

Numerous methods have been developed to assess and test for publication bias in a meta-analysis, including fail-safe *N* (Orwin, 1983; Rosenthal, 1979), funnel plot (Light & Pillemer, 1984), Egger's regression test (Egger, Smith, Schneider, & Minder, 1997), rank-correlation test (Begg & Mazumdar, 1994), test of excess significance (Ioannidis & Trikalinos, 2007), and *p*-uniform's publication bias test (van Aert, Wicherts, & van Assen, 2016; van Assen, van Aert, & Wicherts, 2015). Other methods were developed to provide an effect size estimate corrected for publication bias as a sensitivity analysis: trim and fill (Duval & Tweedie, 2000a, 2000b), PET-PEESE (Stanley & Doucouliagos, 2014), *p*-uniform (van Aert et al., 2016; van Assen et al., 2015), *p*-curve (Simonsohn, Nelson, & Simmons, 2014), selection model approaches (for an

overview see Hedges & Vevea, 2005; Jin, Zhou, & He, 2014; Sutton, Song, Gilbody, & Abrams, 2000), and methods based on the 10% most precise effect size estimates in a meta-analysis (Stanley, Jarrell, & Doucouliagos, 2010).

In this paper, we focus on estimating effect size corrected for publication bias, because we believe that from the perspective of an applied researcher it is most relevant to know what the effect size is corrected for publication bias. More specifically, we focus on estimating effect size corrected for publication bias using the new method *p*-uniform* in this paper. This new method is an extension and improvement of *p*-uniform (van Aert et al., 2016; van Assen et al., 2015) and *p*-curve (Simonsohn et al., 2014) to correct effect size for publication bias. *P*-uniform and *p*-curve are based on the same methodology but slightly differ in implementation (for a comparison of the two methods see van Aert et al. (2016)). These methods use the statistical principle that the *p*-values should be uniformly distributed at the true effect size.

Three major drawbacks of *p*-uniform and *p*-curve in their current implementation are that (i) the methods only use statistically significant effect sizes which makes the methods inefficient (i.e., estimates often have large variance), (ii) effect size estimates are positively biased in the presence of between-study variance in true effect sizes (Carter, Schönbrodt, Gervais, & Hilgard, 2019; McShane, Böckenholt, & Hansen, 2016; van Aert et al., 2016)¹, and (iii) they do not estimate and test for the presence of this between-study variance. In this paper, we propose a revised method called *p*-uniform* that solves all three drawbacks: statistically nonsignificant effect sizes are also included in the estimation with *p*-uniform* (i) making it a more efficient estimator than *p*-uniform, (ii) eliminating the overestimation of effect size in case of between-study variance in true effect sizes, and (iii) enabling estimation and testing for the presence of the between-study variance in true effect sizes.

P-uniform, *p*-curve, and the newly proposed *p*-uniform* can be seen as selection model

approaches, which are a wide class of approaches to correct for publication bias in a meta-analysis. Selection model approaches combine two models to correct for publication bias: an effect size model and a selection model. The *effect size model* is the distribution of primary studies' effect sizes in the absence of publication bias and the *selection model* determines how the effect size model is affected by publication bias (Hedges & Vevea, 2005). Several types of selection model approaches have been proposed, varying from approaches that estimate the selection model to those that assume a specific selection model and from frequentist to Bayesian approaches (Andrews & Kasy, 2019; Cleary & Casella, 1997; Copas & Shi, 2000; Iyengar & Greenhouse, 1988a; Kicinski, 2013; Vevea & Woods, 2005). The supplementary materials (Supplement 1, <https://osf.io/jngwk/>) provides a general overview of selection models.

The goal of this paper is twofold; we introduce *p*-uniform* and examine the statistical properties of *p*-uniform*, *p*-uniform, and the three-parameter selection model (3PSM) via Monte-Carlo simulations including both challenging and realistic conditions with a small number of studies in the meta-analysis. We compare *p*-uniform* with *p*-uniform and 3PSM, because (1) the comparison allows us to examine the conditions where *p*-uniform* is an improvement over *p*-uniform, (2) it is assumed in 3PSM that the selection model is unknown and has to be estimated which is more realistic than other methods (Vevea & Woods, 2005) that assume that the selection model is known, (3) 3PSM was among the best performing methods in recent simulation studies (Carter et al., 2019; McShane et al., 2016), and (4) 3PSM is frequently applied in practice.

Selection model approaches and the three-parameter selection model

All selection method approaches share the common characteristic that they combine an effect size and selection model to correct for publication bias. The effect size model is usually

either the equal-effect (a.k.a. the fixed-effect or common-effect model) or random-effects model. The random-effects model assumes that k independent effect size estimates, y_i with $i=1, \dots, k$, are extracted from primary studies. The random-effects model can be written as

$$y_i = \mu + \zeta_i + \epsilon_i$$

where μ is the average true effect size, ζ_i is a random effect that denotes the difference between μ and the i th primary study's true effect size, and ϵ_i is the i th primary study's sampling error. In the random-effects model, it is commonly assumed that $\zeta_i \sim N(0, \tau^2)$ where τ^2 reflects the between-study variance in true effects, and $\epsilon_i \sim N(0, \sigma_i^2)$ where σ_i^2 is the sampling variance of the i th primary study. The ζ_i and ϵ_i are assumed to be mutually independent of each other, and σ_i^2 is estimated in practice and then assumed to be known. If $\tau^2 = 0$, there is no between-study variance in the true effect sizes, and the random-effects model is equal to the equal-effect model.

The selection model is a non-negative weight function that determines the likelihood of a primary study getting included in a meta-analysis (Hedges & Vevea, 2005). The weight function, $w(y_i, \sigma_i)$, is combined with the effect size model to get a weighted density of y_i ,

$$\frac{w(y_i, \sigma_i)f(y_i, \sigma_i)}{\int w(y_i, \sigma_i)f(y_i, \sigma_i)dy_i} \quad (1)$$

where $f(y_i, \sigma_i)$ denotes the (unweighted) density distribution as in the equal-effect or random-effects model. Note that if $w(y_i, \sigma_i) = 1$ for all y_i , the weighted density is the same as the density of the effect size model (Hedges & Vevea, 2005) and estimates of the selection model approach coincide with those of the equal-effect or random-effects model. This weighted density can be used to estimate parameters (e.g., μ , τ^2 , and parameters in the weight function) in a

selection model approach by means of maximum likelihood estimation.

Different selection model approaches mainly differ in how the selection model is specified. Hedges (1984) proposed the first selection model approach that assigned a weight of 1 to all statistically significant effect sizes and a weight of 0 to nonsignificant effect sizes. This method is comparable to *p*-uniform and *p*-curve except for that maximum likelihood estimation is used in this selection model approach. The selection model approach of Hedges (1984) was extended by Iyengar and Greenhouse (1988a, 1988b) who assigned weights to statistically significant and nonsignificant effect sizes using two different parametric functions. Subsequently, the model by Iyengar and Greenhouse was extended in Hedges (1992) where a step weight function based on *p*-values was introduced for an arbitrary number of steps. The user of this selection model approach has to specify the location of the steps that determine the intervals of the *p*-values. The weight of the first interval is always fixed to one to make sure that the model is identified. The probability of publication is, relative to the first interval, estimated for each interval and these probabilities are used in the weight function. The selection model approach by Hedges (1992) was later extended to enable the inclusion of predictor variables (Vevea & Hedges, 1995) such that the effect size model is a (random-effects) meta-regression model (Borenstein et al., 2009).

We focus in this paper on the 3PSM selection model approach that is commonly used in practice and has similarities with the model proposed by Iyengar and Greenhouse (1988a, 1988b) and (Hedges, 1992). The similarity of 3PSM with the model of Iyengar and Greenhouse (1988a, 1988b) is that statistically significant and nonsignificant effect sizes are receiving different weights. 3PSM is similar to the model in Hedges (1992), because a step function is used as weight function with one step located at the α -level. However, 3PSM only estimates μ , τ^2 , and a single weight parameter whereas the model in Hedges (1992) allows for more than two intervals

and therefore allows for estimation of multiple weights.

Multiple simulation studies have been conducted to study the performance of 3PSM and other selection model approaches as well as to compare these approaches to other publication bias methods. The overarching conclusion of these studies is that selection model approaches are among the best methods to correct for publication bias in a meta-analysis. However, Hong and Reed (2021) also emphasize that performance of the methods depends on the used simulation design and metrics used for evaluating the methods. We summarize the results of previously conducted simulation studies in the supplementary materials (Supplement 1, <https://osf.io/jngwk/>).

From *p*-uniform to *p*-uniform*

***P*-uniform**

P-uniform (van Aert et al., 2016; van Assen et al., 2015) uses the statistical principle that *p*-values are uniformly distributed at the true effect size. The method discards statistically nonsignificant effect sizes and only uses the significant effect sizes to correct for publication bias. Assumptions of the method are that a fixed true effect underlies the primary studies included in the meta-analysis and that all primary studies' effect sizes that are statistically significant in the same direction have an equal probability of getting included in a meta-analysis. Statistical significance is taken into account - and hence there is corrected for publication bias - by computing probabilities of observing an effect size or larger conditional on the effect size being statistically significant. The conditional probability of the *i*th study (q_i) can be written as

$$q_i = \frac{1 - \Phi\left(\frac{y_i - \mu}{\sigma_i}\right)}{1 - \Phi\left(\frac{y_{cv} - \mu}{\sigma_i}\right)} \quad (2)$$

where the numerator is the probability of observing an effect size at the true effect size larger than the effect size in the i th primary study and the denominator is the probability of observing a (statistically significant) effect size (i.e., larger than the critical value y_{cv} , which is the smallest statistically significant effect size given an α -level and σ_i). *P*-uniform can also be seen as a selection model approach with the equal-effect model as effect size model, and a selection model assuming equal weights for statistically significant effect sizes to get published. Discarding nonsignificant effect sizes in *p*-uniform is tantamount to assuming a study has a constant probability of being included in a meta-analysis.

P-uniform's effect size estimate is equal to the value of μ where a statistic that is computed based on the q_i equals its expected value assuming a uniform distribution. Van Assen et al. (2015) proposed to use Fisher's test (Fisher, 1925), $-\sum_{i=1}^k \ln(q_i)$, to estimate the effect size in *p*-uniform and draw inferences. Van Assen et al. (2015) compared *p*-uniform with trim-and-fill (Duval & Tweedie, 2000a, 2000b) to correct for publication bias and concluded that *p*-uniform outperformed trim-and-fill if publication bias exists and between-study variance in true effect size is absent or small. However, they also showed that overestimation of *p*-uniform increased as a function of the between-study variance in true effect sizes.

Another proposed estimator² for estimating the effect size and compute the confidence interval with *p*-uniform is based on the distribution of the sum of independently uniformly distributed random variables, which is called the Irwin-Hall distribution (van Aert et al., 2016).

Van Aert et al. (2016) recommended to use the estimator based on the Irwin-Hall distribution as the default estimator, because (i) summing the conditional probabilities q_i is easy to understand, and (ii) it has the nice property that $\hat{\mu}$ is equal to, larger than, smaller than zero if the average of the statistically significant p -values is equal to, smaller than, larger than α if one-tailed tests or $\alpha/2$ if two-tailed tests were used in the primary studies. Moreover, the estimator based on the Irwin-Hall distribution is less susceptible to outlying effect sizes than the estimator using the Fisher's test.

Critique on p -uniform

McShane et al. (2016) criticizes p -uniform for three reasons; p -uniform (i) assumes a fixed true effect underlying all y_i , (ii) discards statistically nonsignificant effect sizes, and (iii) uses method of moments estimators instead of maximum likelihood estimation. The first critique is indeed a limitation of p -uniform, because assuming that the true effect size is fixed results in an overestimated effect size if this assumption is violated (Carter et al., 2019; McShane et al., 2016; van Aert et al., 2016; van Assen et al., 2015). Moreover, methods that do not assume that the true effect size is fixed are favorable, because heterogeneity is often present in meta-analyses (Higgins, 2008; Higgins, Thompson, & Spiegelhalter, 2009; van Erp, Verhagen, Grasman, & Wagenmakers, 2017).

The second critique by McShane et al. (2016) relates to the loss of efficiency of p -uniform's estimators because of discarding nonsignificant effects. Efficiency loss may be limited in many applications, as the vast majority of published results in psychology are statistically significant (Fanelli, 2010, 2012). Nevertheless, as meta-analyses including many nonsignificant effects do occur (van Aert et al., 2019) and the potential efficiency loss is both particularly prevalent and detrimental in the important cases where the average true effect size is (close to)

zero, we generalized our methodology and also include statistically nonsignificant effect sizes in *p*-uniform*.

The final critique by McShane et al. (2016) relates to the optimal large-sample properties of the maximum likelihood estimator compared to method of moments estimators (Casella & Berger, 2002). Although van Aert et al. (2016) and van Assen et al. (2015) were aware of these large-sample optimal properties, they intentionally selected method of moments estimators, because these yield exact confidence intervals even if only one statistically significant effect size is included in a meta-analysis. This is in contrast with the conventional Wald-based confidence interval and hypothesis test that is accompanied by maximum likelihood estimation, because these assume that the log-likelihood around the maximum likelihood estimate is regular (Pawitan, 2013). As *p*-uniform uses conditional probabilities (given statistical significance) as likelihoods that are truncated, the log-likelihood is not well approximated by the normal distribution for a large and relevant part of the parameter space (parameter values close to 0) especially in case of a small number of effect sizes in a meta-analysis. Hence, Wald-based confidence intervals and hypothesis tests are generally inappropriate. To bypass problems of non-normally distributed log-likelihoods, we implemented *p*-uniform* using maximum likelihood estimation but computed confidence intervals of μ and τ^2 by inverting the likelihood-ratio test and using the likelihood-ratio test for testing the null hypothesis $\mu = 0$. These procedures do not assume asymptotic normality distributions as the Wald-based confidence intervals do (Agresti, 2013; Pawitan, 2013) and are therefore expected to have better statistical properties.

P*-uniform

P-uniform* is a selection model approach with the random-effects model as effect size model. The selection model assumes that the probability of publishing a statistically significant

effect size as well as a nonsignificant effect size are constant, but these two probabilities may be different from each other. Hence, p -uniform* works by only treating the primary studies' effect sizes differently depending on whether they are statistically significant or not. P -uniform* can be seen as a selection model approach where the selection model has one cut-off at the critical value determining whether an effect size is statistically significant or not.

Maximum likelihood estimation is used in p -uniform* where truncated densities are being used instead of the conditional probabilities in Equation (2)³. Truncated densities (q_i^{ML*}) are computed for both the statistically significant and nonsignificant effect sizes and are a function of both μ and τ^2 ,

$$q_i^{ML*} = \begin{cases} \frac{\frac{1}{\sqrt{\sigma_i^2 + \tau^2}} \phi\left(\frac{y_i - \mu}{\sqrt{\sigma_i^2 + \tau^2}}\right)}{1 - \Phi\left(\frac{y_{cv} - \mu}{\sqrt{\sigma_i^2 + \tau^2}}\right)} & \text{if } p_i \leq \alpha, \\ \frac{\frac{1}{\sqrt{\sigma_i^2 + \tau^2}} \phi\left(\frac{y_i - \mu}{\sqrt{\sigma_i^2 + \tau^2}}\right)}{\Phi\left(\frac{y_{cv} - \mu}{\sqrt{\sigma_i^2 + \tau^2}}\right)} & \text{if } p_i > \alpha \end{cases} \quad (3)$$

where ϕ denotes the standard normal probability density function. The likelihood function is the product of the q_i^{ML*} :

$$L(\mu, \tau^2) = \prod_{i=1}^k q_i^{ML^*}. \quad (4)$$

The ‘weights’ of the selection model of *p*-uniform* are included and implied in the computation of the truncated densities. The numerators of the truncated densities in Equation (3) (i.e., the usual likelihoods) are weighed by the reciprocal of the probability of observing a (non)significant effect size given a value for μ , τ^2 , and σ_i^2 .

Confidence intervals for μ and τ^2 are obtained using *p*-uniform* by inverting the likelihood-ratio test statistic. The likelihood-ratio test (Agresti, 2013; Pawitan, 2013) is used to test the null hypotheses $\mu = 0$ and $\tau^2 = 0$ that compares the null model where either μ or τ^2 is fixed to 0 with the alternative model where both parameters are estimated. We also implemented *p*-uniform* with method of moments estimation using the Irwin-Hall distribution and the Fisher’s test as estimator. We describe the procedure for estimation with the method of moments estimators and the results of the simulation study using these estimators in the supplementary materials (Supplement 2, <https://osf.io/jngwk/>).

***P*-uniform* compared to 3PSM**

3PSM requires that at least one study is observed in the intervals of the step function to estimate the weights of the selection model. One study per interval technically allows for estimation of the weights. However, it has been recommended for selection model approaches that at least 10 to 15 studies have to be available in each interval to allow accurate estimation of the weights (Hedges & Vevea, 1996, 2005; Vevea & Woods, 2005), which implies that at least 20 studies are recommended when modeling publication bias (i.e., when distinguishing statistically significant from nonsignificant effect sizes). The main difference between *p*-

uniform* and 3PSM is that *p*-uniform* does not (need to) estimate the weight parameter, and is therefore a more parsimonious model. More parsimonious models are generally preferred over more complicated models if the performance of the models is comparable.

P-uniform* does not require that at least one study is observed in the intervals of the step function, preliminary analyses and the known statistical properties of *p*-uniform (van Aert et al., 2016) indicate that *p*-uniform* may also not perform well in these situations. For example, the average effect size estimate of *p*-uniform* is expected to be very negative if there are only statistically significant studies in a meta-analysis and some of these studies have *p*-values slightly smaller than the α -level. Estimation of the between-study variance in true effect sizes will also be challenging for both methods if the average true effect size is close to zero and there are only statistically significant effect sizes in the meta-analysis. The methods then need to estimate the variance of a truncated normal distribution from which only a very small portion is observed.

Monte-Carlo simulation study

Method

Standardized mean differences were the effect size measure of interest using a two-independent groups design with a sample size of $n_i = 50$ per group. We used Hedges' *g* standardized mean difference rather than Cohen's *d*, because a small positive bias in Cohen's *d* is corrected for by Hedges' *g* (Hedges, 1981). The Hedges' *g* effect size is obtained by multiplying

Cohen's *d* with the correction factor $J = \frac{\Gamma\left(\frac{df}{2}\right)}{\sqrt{\frac{df}{2}} \Gamma\left(\frac{df-1}{2}\right)}$ where Γ refers to the gamma function

and *df* to the degrees of freedom.

We started the Monte-Carlo simulation study by first sampling a true effect size θ_i for the i th primary study from $N(\mu, \tau^2)$. Let $\tilde{n}_i = \frac{n_i}{2}$ and g_i being the observed Hedges' g effect size in the i th study. The transformation $J^{-1}\sqrt{\tilde{n}_i}g_i$ approximates a non-central t -distribution with df degrees of freedom and non-centrality parameter $\theta_i\sqrt{\tilde{n}_i}$ (Hedges, 1981, 1983; Viechtbauer, 2005). Hedges' g effect size g_i was then obtained by sampling a t -value from this non-central t -distribution and dividing the sampled t -value by $J^{-1}\sqrt{\tilde{n}_i}$. The unbiased estimate of the sampling variance of Hedges' g (see Equation 26 in Viechtbauer (2007a)) was computed with

$$\frac{1}{\tilde{n}_i} + \left(1 - \frac{df - 2}{df \times J}\right)^2 g_i^2.$$

The effect size of the i th primary study was always included in the meta-analysis if it was statistically significant. A one-tailed hypothesis test with $\alpha=.025$ was used to resemble common practice of research in psychology where a two-tailed test with $\alpha=.05$ is conducted and only effect sizes in the predicted direction are reported. Statistically nonsignificant effect sizes were included in the meta-analysis if a randomly drawn number from a uniform distribution ranging from zero to one was smaller than $1 - pub$, where $1 - pub$ represents the probability of a statistically nonsignificant effect size to be included in a meta-analysis with $pub = 1$ referring to extreme publication bias (only statistically significant studies get published). This procedure for generating data of primary studies was repeated until k primary studies' effect sizes were included in a meta-analysis.

Simulated data of meta-analyses may be generated with only statistically significant or statistically nonsignificant effect sizes. This implies that there are no studies in one of the intervals of the selection model. Since these are challenging conditions for p -uniform* and 3PSM cannot be applied in this situation (also see section “ P -uniform* compared to 3PSM”), we also

ran the simulation study where we replaced a study from the meta-analysis with a statistically (non)significant study if there were only (non)significant studies in a simulated meta-analysis. That is, if the aforementioned outlined procedure resulted in a meta-analysis with no studies in one of the intervals, then we continued sampling studies until one study was obtained in the previously empty interval, and replaced one random originally simulated study with this new study.

The following variables were varied in the Monte-Carlo simulations: μ , τ , k , and pub . Three different levels were selected for μ (0; 0.2; 0.5) reflecting no, a small, and a medium effect (Cohen, 1988). The square root of the between-study variance in true effect sizes (τ) was 0, 0.163, or 0.346 representing I^2 -statistics equal to 0%, 40%, and 75% (zero, small-medium, large (Higgins, Thompson, Deeks, & Altman, 2003)). The number of effect sizes in a meta-analysis (k) was equal to 10, 30, 60, and 120; 10 and 30 are close to the median (12) and mean (38.7) number of effect sizes in meta-analyses in psychology (van Erp et al., 2017), respectively. We also included 60 and 120 because previous research (Field & Gillett, 2010; Hedges & Vevea, 2005; Vevea & Woods, 2005) suggests that a large number of effect sizes in a meta-analysis are required in order for selection model approaches to perform well. Four different levels for pub were selected: 0, 0.5, 0.9, and 1. Combining the different levels of these variables resulted in $3 \times 3 \times 4 \times 4 = 144$ conditions. For each condition 10,000 runs were conducted.⁴

P-uniform* and 3PSM were applied to each simulated meta-analysis. *P*-uniform* was implemented using maximum likelihood estimation and estimation based on the Irwin-Hall distribution and Fisher's test (Fisher, 1925). The random-effects model was included to be able to compare methods that correct for publication bias with the method that is usually applied and does not correct for publication bias. We used the Paule-Mandel estimator (Paule & Mandel,

1982) to estimate the between-study variance in true effect sizes in the random-effects model, because two papers reviewing existing estimators of the between-study variance recommend this estimator (Langan, Higgins, & Simmonds, 2016; Veroniki et al., 2016). The outcome variables were the average, median, and standard deviation of the estimates, RMSE, and coverage probability and average width of the 95% confidence intervals for μ and τ .⁵

The Monte-Carlo simulation study was programmed in R (R Core Team, 2022) and the packages “metafor” (Viechtbauer, 2010), “weightr” (Coburn & Vevea, 2016), and “puniform” (van Aert, 2022) were used for applying the random-effects model, 3PSM, and *p*-uniform respectively. 3PSM was applied by using the implementation of “weightr” and “metafor”, because both implementations slightly differ which may impact the estimation and possible non-convergence. Furthermore, profile likelihood confidence intervals are implemented for τ^2 in “metafor” whereas confidence intervals for τ^2 are not implemented in “weightr”. Other R packages that were used to decrease the computing time of the simulations were the “parallel” package (R Core Team, 2022) for parallelizing the simulations and the “Rcpp” package (Eddelbuettel, 2013) for executing C++ functions. R code of the Monte-Carlo simulation study where a study is replaced in case of only statistically (non)significant effect sizes is available at <https://osf.io/dyt4b>, and R code of the simulation study without replacement is available at <https://osf.io/f7ygu>.⁶

Results

Results of simulations that allow only (non)significant effect sizes in a meta-analysis

We first present the results of the simulations where no extra steps were taken to have at least one statistically significant and nonsignificant effect size in every simulated meta-analysis.

We focus on the conditions where $\mu=0$, $\tau=0.346$, $k=10$, and $pub=0$ and 1, because the condition with $pub=1$ is most challenging since only statistically significant effect sizes are in the meta-analysis and there is large heterogeneity. The results of all other conditions are available in the supplemental materials (<https://osf.io/gycf2>).

Table 1 shows the non-convergence rate as well as the mean and median of the estimated parameters μ and τ of the methods across the 10,000 simulation runs. The random-effects model and *p*-uniform* always converged. *P*-uniform and 3PSM as implemented in “metafor” did not converge when $pub=0$ in 17.2% of the simulation runs, because there were no statistically significant effect sizes in these runs. This is in line with the assumptions of *p*-uniform and 3PSM, because *p*-uniform only uses statistically significant effect sizes and 3PSM needs effect sizes in both intervals in order to estimate the weight parameter. In case of $pub=1$, there are only statistically significant studies causing non-convergence in all simulation runs of 3PSM as implemented in “metafor”. 3PSM as implemented in “weightr” suffered less from convergence problems than its implementation in “metafor”. The reason for this is that the developers of “weightr” made the arbitrary choice to fix the weight parameter to 0.01 in case it could not be estimated.

If $pub=0$, the mean of the estimates of μ across all simulation runs was close to the true effect size for all methods except for *p*-uniform*. *P*-uniform*’s mean of the estimates was unrealistically large (4.451). This was caused by estimates for a small number of simulated meta-analyses that were very large due to effect sizes in these meta-analyses that were slightly larger than the α -level. The median of the estimates across all simulation runs of *p*-uniform* was close to the true effect size. If $pub=1$, the random-effects model, *p*-uniform, and 3PSM as implemented in “weightr” all yielded positively biased estimates of μ . *P*-uniform* severely underestimates the true effect size in this condition (-575.299). This was caused by very negative estimates of a

small number of simulated meta-analyses with effect sizes that had p -values that were slightly smaller than the α -level. There was a small negative bias in the estimates of τ for all methods if $pub=0$. If $pub=1$, τ was severely underestimated by the random-effects model and 3PSM as implemented in “weightr” and overestimated by p -uniform*.

We also studied whether parameter estimation of p -uniform* in case of only statistically significant effect sizes was still acceptable for meta-analyses with particular characteristics. If this would be the case, meta-analysts can be recommended to apply p -uniform* in these cases. We examined the performance of p -uniform* by focusing on the simulation runs where (1) the sign of the estimate of p -uniform* was the opposite of the random-effects model, (2) the null-hypothesis of no effect was rejected with p -uniform*, and (3) the absolute value of the estimate of p -uniform* was larger than of the random-effects model. The mean of the effect size estimates of only these simulation runs were no longer negatively biased but had a positive bias (0.290). This shows that p -uniform*'s negative bias was not present if the criterion was applied, but the estimates after applying the criterion were positively biased just as the other methods.

In the remainder of this results section, we focus on the simulations where at least one statistically significant and nonsignificant effect size is in the meta-analysis. This is the condition 3PSM was developed for and the vast majority of meta-analyses in practice contains at least one statistically significant and nonsignificant effect size (van Aert et al., 2019).

Results of simulations with at least one significant and nonsignificant effect sizes in a meta-analysis

We only present the results of p -uniform* with maximum likelihood estimation in the paper, because this estimator outperformed the estimators using Fisher's test (Fisher, 1925) and the Irwin-Hall distribution (see supplements at <https://osf.io/gycf2/>). The width of the confidence

intervals is not presented, because coverage probabilities often substantially deviated from the nominal coverage rate, thereby decreasing the usefulness of assessing the width of confidence intervals. Finally, we only present the results for $k=10$ and 60 in this section, because these conditions already illustrate how the methods' performances increase in k ; the condition $k=120$ was omitted because the methods' performance in that condition was not remarkably different from that in $k=60$. All results that are not presented in the paper are available online (<https://osf.io/gycf2/>).

Average estimates of μ Figures 1 and 2 show the average of the estimates of μ when μ (columns of the figures), τ (rows of the figures), and pub (x-axis of the figures) were varied for $k=10$ and $k=60$, respectively. All the figures are centered at the true effect size μ (dashed gray line) to facilitate comparability of the different subfigures as we varied μ . We first describe the results of $k=10$ and then illustrate how the results change if $k=60$.

Highlighting common issues with a lack of correction for publication bias, the random-effects model overestimated μ under publication bias and this overestimation decreased in μ and increased in τ and pub . *P*-uniform yielded a large negative bias if $pub < 1$ especially in conditions where only a small number of effect sizes per meta-analysis was statistically significant. This negative bias was caused by effect sizes with p -values close to the α -level. The properties of *p*-uniform* and both implementations of 3PSM were similar resulting in overlapping lines and symbols in Figure 1. The bias was negligible for all levels of pub . For the conditions where $\mu=0$ and $pub=1$, the bias of 3PSM as implemented in “weightr” was slightly larger than *p*-uniform* and “metafor” its implementation of 3PSM.

Bias of the random-effects model was unaffected by increasing the number of studies to 60 (see Figure 2). Bias of *p*-uniform* and “metafor” its implementation of 3PSM was also

negligible for $k=60$. The positive bias of the “weightr” implementation of 3PSM when $pub=1$ increased for $\mu=0$ and was also present for $\mu=0.2$. The positive bias of the “weightr” implementation was a function of the number of studies in the meta-analysis, because the bias was larger for $k=120$ compared to 60 (see results in the supplements).

RMSE for estimating μ Figures 3 and 4 show the RMSE for estimating μ for $k=10$ and 60. The RMSE for the random-effects model followed the patterns observed for its bias; RMSE increased in publication bias and τ , and decreased in μ . Not only for $pub = 0$, but also for $pub = 0.5$, the random-effects model had a lower RMSE than the other methods, because its bias was small while at the same time the standard deviation of its estimates was lower than for the other methods (see <https://osf.io/gycf2/>).

P-uniform’s RMSE was in the vast majority of conditions larger than those of *p*-uniform* and the two implementations of 3PSM. *P*-uniform was expected to outperform all other methods in the conditions with $\tau=0$ in combination with $pub=1$, because there is no heterogeneity as assumed in *p*-uniform and there is extreme publication bias. However, its RMSE is larger than that of *p*-uniform* and the two implementations of 3PSM. This was caused by the single statistically nonsignificant effect size that replaced a significant effect size in these simulations. *P*-uniform was then based on only the nine statistically significant effect sizes and this resulted in a higher RMSE than *p*-uniform* and the two implementations of 3PSM.

RMSE of *p*-uniform* and both implementations of 3PSM were comparable in all conditions except for $pub=1$. If $pub=1$, the implementation of 3PSM in “weightr” yielded a smaller RMSE if $\mu=0$ or 0.2. The reason for this is that the weight parameter in the 3PSM implementation of “weightr” is constrained to be at least 0.01. Constraining this weight made sure that estimates did not become extremely negative and yielded a larger RMSE than *p*-

uniform* and the implementation of 3PSM in “metafor”.

P -uniform benefitted most from increasing k to 60. It was among the methods with the smallest RMSE in the conditions with $\tau=0$ in combination with $pub=1$. RMSE of the other methods that correct for publication bias were comparable. The RMSE of the 3PSM implementation in “weightr” was again sometimes lower than that of p -uniform* and the implementation of 3PSM in “metafor” due to constraining the weight parameter to 0.01 or larger.

Coverage probability of confidence interval for μ A confidence interval for μ could always be computed for the random-effects model and p -uniform* but not with the two implementations of 3PSM. Confidence intervals of 3PSM could not be computed in at most 47.6% for the implementation in “weightr” and 26.8% for the implementation in “metafor” (condition $\mu = 0, \tau = 0, k=10, pub = 1$). P -uniform’s confidence interval could, similarly to its effect size estimation, only be computed if a meta-analysis contained statistically significant effect sizes. The coverage probabilities shown in this section were determined based on all confidence intervals that could be computed for a particular method. This makes directly comparing the methods difficult since the number of confidence intervals that could be computed differed across methods. Furthermore, it is important to emphasize that differences in coverage probabilities between 3PSM and p -uniform* can be caused by the differences between the methods or by the fact a Wald-based confidence interval is computed by 3PSM and a profile likelihood confidence interval by p -uniform*.

Table 1 presents the coverage probability of the 95% confidence interval for μ if $k=10$ and $k=60$. Coverage probabilities between 0.9 and 0.975 are marked in bold and are deemed to be acceptable. Coverage probabilities of the random-effects model decreased as a function of pub . Coverage probabilities of all methods that correct for publication bias were close to the nominal

coverage rate if $k=10$ in combination with $\tau=0$. These results confirm that *p*-uniform yields exact confidence intervals if its assumptions are met. Nonzero values of τ in combination with the presence of publication bias resulted in undercoverage of all methods. This undercoverage was most severe if $pub=1$ and performance of none of the methods was acceptable in this condition. The undercoverage of all methods increased for $k=60$.

Average estimates of τ Figures 5 and 6 show the average estimates of τ for $k=10$ and 60, respectively. Note that *p*-uniform does not estimate τ . All methods were positively biased if $\tau = 0$ in combination with $k=10$, and the random-effects model yielded the least biased estimator for these conditions if $pub = 1$. The random-effects model yielded sometimes a larger bias than the other methods if $\tau = 0$ in combination with $pub=0$. This was caused by replacing a statistically nonsignificant effect size with a significant effect size in the simulations if there were only nonsignificant effect sizes in a meta-analysis. The bias was comparable to those of the other methods for the simulations without replacement (see supplements at <https://osf.io/gycf2>). *P*-uniform* and the two implementations of 3PSM yielded comparable results. The methods were negatively biased if $\tau > 0$. Bias of the random-effects model was generally less than of the methods that correct for publication bias if $pub < 0.9$. Increasing k to 60 did especially decreased the bias of *p*-uniform* and 3PSM as implemented in “metafor”. The implementation of 3PSM in “weightr” was more negatively biased when than the other methods to correct for publication bias if $\tau > 0$ in combination with $pub = 1$. This was again caused by the restricting the weight parameter to 0.01 or larger.

RMSE for estimating τ Figures 7 and 8 present the RMSE for estimating τ . For $k=10$ (Figure 7), the RMSE of the random-effects model was higher than of the other methods for the condition $\tau = 0$ in combination with $pub=0$. This was caused by the replacement procedure that

was used in these simulations. The random-effects model had a smaller RMSE than the other methods if $\tau = 0$ in combination with $\mu > 0$ and $pub = 1$. RMSE of all methods were comparable if $\tau > 0$. For $k=60$, RMSE of especially the methods that corrected for publication bias decreased compared to $k=10$.

Coverage probability of confidence interval for τ Note that the implementation of 3PSM in “weightr” does not compute confidence intervals for τ . Table 2 presents the coverage probabilities of the random-effects model, *p*-uniform*, and 3PSM as implemented in “metafor”. Coverage probabilities between 0.9 and 0.975 are marked in bold and are evaluated as acceptable. For $k=10$, the random-effects model yielded accurate coverage if $\mu=0.5$ and this was also the case if publication bias was present. Coverage probabilities of *p*-uniform* were especially too low if $pub \geq 0.9$ and $\tau=0.346$. The implementation of 3PSM in “metafor” yielded overcoverage for $\tau < 0.346$ and acceptable coverage if $\tau=0.346$. Increasing the number of studies to $k=60$ resulted in more conditions with acceptable coverage probabilities for *p*-uniform* and the implementation of 3PSM in “metafor” and less conditions with acceptable coverage of the random-effects model.

Application

We apply random-effects meta-analysis, *p*-uniform, *p*-uniform*, and 3PSM using the implementations in the R packages “weightr” and “metafor” to two published meta-analyses in this section that differed in the amount of estimated between-study variance. The meta-analysis by Rabelo, Keller, Pilati, and Wicherts (2015) consists of 25 Hedges’ *g* standardized mean differences on whether participant’s judgement of importance of, for instance, morality-related outcomes changes if they are holding a heavy- or lightweight object. The total sample sizes of the primary studies ranged from 30 to 100 (mean = 61.12, median = 60). All effect sizes were

positive and 21 (84%) effect sizes were statistically significant based on a two-tailed test with $\alpha=.05$.

The methods were also applied to the meta-analysis by Bangert-Drowns, Hurley, and Wilkinson (2004). This meta-analysis consisted of 48 Hedges' *g* standardized mean differences on the effect of writing-to-learn interventions on academic achievement. That is, students' academic achievement in these primary studies was compared between an experimental group where there was an explicit focus on learning by writing and a control group where traditional teaching methods were used. Sampling variances of the Hedges' *g* effect sizes were computed assuming equal sample sizes in both groups. The total sample sizes ranged from 16 to 542 (mean = 116.2, median = 67.5). Based on a two-tailed test with $\alpha=.05$, 14 effect sizes were significantly larger than zero and 1 effect size was significantly smaller than zero. R code of applying the methods to these to meta-analysis is available at <https://osf.io/9k3bp/>.

Table 4 shows the results of applying the methods to the meta-analysis of Rabelo et al. (2015) (first five rows) and Bangert-Drowns et al. (2004) (last five rows). The average effect size estimate of Rabelo et al. (2015) was substantially smaller when correcting for publication bias with *p*-uniform (-0.179), *p*-uniform* (0.244), and both implementations of 3PSM (0.254) when compared with the estimate not corrected for bias (0.571). The between-study variance was estimated as zero by all methods that estimated this parameter. These results suggest that the effect of holding a heavy- or lightweight object on judgement of importance is most likely a small rather than a large effect.

The correction for publication bias was smaller in the meta-analysis of Bangert-Drowns et al. (2004), because the average effect size estimate of random-effects meta-analysis (0.228) was only slightly larger than estimates obtained with *p*-uniform* (0.179) and both implementations of

3PSM (0.148). *P*-uniform's estimate (0.245) was larger than the estimate of random-effects meta-analysis, which was likely caused by *p*-uniform overestimating the average effect size due to heterogeneity in true effect sizes. The estimated between-study variance by *p*-uniform* (0.027) and both implementations of 3PSM (0.028) was smaller than estimated in the random-effects meta-analysis (0.069). To conclude, correcting for potential publication bias did not affect the conclusions that writing-to-learn interventions affects academic achievement.

Discussion

Publication bias distorts the results of meta-analyses yielding overestimated effect sizes. Multiple methods were developed to correct for publication bias in a meta-analysis, and selection model approaches are seen as the state-of-the-art methods (McShane et al., 2016). The *p*-uniform method (van Aert et al., 2016; van Assen et al., 2015) can also be seen as a selection model approach, and we extended and improved this method in this paper. The new method, *p*-uniform*, does not only use statistically significant primary studies' effect sizes for estimation as is done in *p*-uniform, but also uses the nonsignificant effect sizes. Including the statistically nonsignificant primary studies' effect sizes results in three major improvements of *p*-uniform* over *p*-uniform: (i) it makes *p*-uniform* a more efficient estimator than *p*-uniform, (ii) overestimation of effect size by *p*-uniform in case of between-study variance in true effect is eliminated, and (iii) it enables estimation and testing for the presence of the between-study variance in true effect sizes.

The aim of this paper was to introduce *p*-uniform* and compare the statistical properties of the method with those of *p*-uniform, two implementations of 3PSM, and the random-effects model that is commonly used but does not correct for publication bias. The Monte-Carlo simulation study confirmed previous research that the random-effects model yields overestimated

average effect size and unpredictable bias in the between-study variances in true effect sizes (Augusteijn, van Aert, & van Assen, 2019; Jackson, 2006, 2007) if publication bias is present. The simulation study also showed that p -uniform* is an improvement over p -uniform if between-study variance is present. Statistical properties of p -uniform* and 3PSM were generally comparable, but these were not acceptable in case of extreme publication bias with only statistically significant primary studies' effect sizes in a meta-analysis. However, the simulation study and previous research (van Aert et al., 2016; van Assen et al., 2015) showed that p -uniform can be used in such a situation if heterogeneity in true effect sizes is zero or small.

Given that p -uniform* showed to be a major improvement over p -uniform if between-study variance is present and p -uniform* showed comparable statistical properties to 3PSM, p -uniform* is a viable method to correct for publication bias. P -uniform* has the advantage over 3PSM that the weight in the selection model does not need to be estimated. P -uniform* only assumes that these weights are the same for the statistically significant and the same for the nonsignificant primary studies' effect sizes, without requiring estimation of these weights. This makes p -uniform* a more parsimonious model than 3PSM.

We provide recommendations for meta-analysts in practice based on the results of our Monte-Carlo simulation study. These recommendations are built upon guidelines for conducting meta-analyses as the Meta-Analytic Reporting Standards (MARS; Appelbaum et al., 2018), and the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA; Moher, Liberati, Tetzlaff, Altman, & The Prisma Group, 2009). First of all, we recommend to not solely rely on the traditional meta-analysis models if publication bias may have affected the meta-analysis, but to supplement these models by both p -uniform* and 3PSM. Examining publication bias in a meta-analysis is in agreement with the MARS and PRISMA that both recommend to assess the risk and consequences of bias in any meta-analysis. We advise researchers to use so-

called triangulation where researchers do not rely on one particular publication bias method but use multiple publication bias methods that are known to have good statistical properties for the characteristics of the meta-analysis (Carter et al., 2019; Coburn & Vevea, 2015; Kepes, Banks, McDaniel, & Whetzel, 2012). This is necessary because research, including ours, has shown that there is no single publication bias methods that outperforms all the other available publication bias methods. If both *p*-uniform* and 3PSM reach similar conclusions, the meta-analysts can be more confident in their results and conclusions based upon these results. In case of diverging conclusions by the two methods, the results of both methods are recommended to be reported as well and interpretation of the results must be done with caution.

Importantly, we do not recommend applying 3PSM using the implementation in “weightr”. Fixing the weight parameter to 0.01 if there are no statistically (non)significant effect sizes in a meta-analysis is unwarranted just as forcing it to be 0.01 or larger in case of a mix of significant and nonsignificant effect sizes. Setting the weight parameter to 0.01 is not evidence-based and 0.01 seems unrealistically small, implying that 99% of the statistically nonsignificant effect sizes end up in the file drawer. We do also not recommend to use *p*-uniform* if the meta-analysis only contains statistically significant effect sizes caused by publication bias. Although, *p*-uniform* can technically, in contrast to 3PSM, be applied, our Monte-Carlo simulation study showed bad performance in this condition. If the between study variance is expected to be zero or small, *p*-uniform can then better be applied since it provides estimates close to the true effect size and exact confidence intervals in these situations. However, we also suspect good performance of *p*-uniform* in this condition when all statistically significant effect sizes are accompanied by very small *p*-values (say < .001), suggesting that these significant effects are not caused by publication bias but by high power of the original studies.

P-uniform* can be easily applied by meta-analysts using the function “puni_star()” in the

R package “puniform”. Users can currently analyze primary studies based on a two-samples or one-sample *t*-test and correlation coefficient or can supply the function directly with standardized effect sizes. Meta-analysts who are not familiar with R can also apply *p*-uniform* to their data via an easy-to-use web application (<https://rvanaert.shinyapps.io/p-uniformstar/>). This web application can be applied if the primary study’s effect size measure is a one-sample mean, two-independent means or correlation coefficient. 3PSM can be applied using the “weightr” package (Coburn & Vevea, 2016), the corresponding web application (<https://vevealab.shinyapps.io/WeightFunctionModel/>) or by using the “metafor” package (Viechtbauer, 2010).

A limitation of *p*-uniform* and 3PSM is that the probability of publishing a statistically nonsignificant primary study’s effect size is assumed to be the same for all nonsignificant effect sizes in a meta-analysis. This assumption may be violated in practice causing biased estimates of the methods (see Simonsohn, Simmons, and Nelson (2017, December 20) for a discussion about this assumption). To counteract a violation of this assumption, 3PSM can be extended to allow for more than two intervals of the selection model. *P*-uniform*’s selection model can, in principle, also be extended to allow for more than two intervals. For example, researchers may expect that the probability of publishing a primary study’s effect size with a marginally significant *p*-value (i.e., *p*-value between 0.05 and 0.1) is different from publishing a statistically nonsignificant effect size with a *p*-value larger than 0.1. This can be incorporated in *p*-uniform* where the truncated densities are computed differently for these three intervals. That is, the denominators of the truncated density in Equation (3) for an effect size with a *p*-value smaller than 0.05, between 0.05 and 0.1, and larger than 0.1 equal the probability of observing a *p*-value in this interval given μ and τ . However, information needs to be available to select appropriate thresholds for these intervals. We simulated data in this study that exactly matched the selection

models of *p*-uniform* and 3PSM, so future research may study to what extent statistical properties of the methods deteriorate if the selection model for simulating data differs from the methods' selection model. Future research may also study the effect of publication bias on the meta-analytic results if covariates are included in a meta-analysis model.

Another limitation of the methods is that their results will be distorted if *p*-hacking (a.k.a. questionable research practices or researcher degrees of freedom, see Simmons, Nelson, and Simonsohn (2011); Wicherts et al. (2016)) are used in the primary studies that are included in a meta-analysis, but this limitation applies to any meta-analysis method. As our study did not address the performance of publication bias tests, we also recommend further research on developing and examining publication bias tests. A publication bias test was also proposed for *p*-uniform (van Assen et al., 2015) and in the context of 3PSM. Future research may study the statistical properties of these publication bias tests. Future research may also consider to implement *p*-uniform* in a Bayesian framework. This Bayesian version of *p*-uniform* can then together with *p*-uniform* and 3PSM be compared with other selection model approaches that were developed in a Bayesian framework. Two recently developed Bayesian methods that deserve attention in future research are the BALM method (Du, Liu, & Wang, 2017) and the Bayesian model averaging method proposed by Guan and Vandekerckhove (2015).

To conclude, scientific progress can best be achieved by using meta-analysis (Cumming, 2008), but this progress is hampered by publication bias causing false positive and overestimated effect sizes. Hence, there is a need for methods that can accurately estimate the effect size and between-study variance in a meta-analysis in the presence of publication bias. *P*-uniform* is an extension and substantial improvement over *p*-uniform and showed promising results in Monte-Carlo simulations.

Data availability statement

R codes of the Monte-Carlo simulation studies are available on the OSF page of the paper:

<https://osf.io/ebq6m/>.

References

- Agresti, A. (2013). *Categorical data analysis*. Hoboken, N.J.: Wiley-Interscience.
- Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8), 2766-2794. doi:10.1257/aer.20180310
- Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *The American psychologist*, 73(1), 3-25. doi:10.1037/amp0000191
- Augusteijn, H. E. M., van Aert, R. C. M., & van Assen, M. A. L. M. (2019). The effect of publication bias on the Q test and assessment of heterogeneity. *Psychological Methods*, 24(1), 116-134. doi:10.1037/met0000197
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7(6), 543-554. doi:10.1177/1745691612459060
- Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, 74(1), 29-58. doi:10.3102/00346543074001029
- Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50(4), 1088-1101.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons, Ltd.
- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2), 115-144. doi:10.1177/2515245919847196
- Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Belmont, CA: Duxbury.
- Cleary, R. J., & Casella, G. (1997). An application of Gibbs sampling to estimation in meta-analysis: Accounting for publication bias. *Journal of Educational and Behavioral Statistics*, 22(2), 141-154.
- Coburn, K. M., & Vevea, J. L. (2015). Publication bias as a function of study characteristics. *Psychological Methods*, 20(3), 310-330. doi:10.1037/met0000046
- Coburn, K. M., & Vevea, J. L. (2016). *weightr*: Estimating weight-function models for publication bias. Retrieved from <https://CRAN.R-project.org/package=weightr>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304-1312.
- Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*, 2(4), 447-452. doi:10.1037/1082-989X.2.4.447
- Copas, J. B., & Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics*, 1(3), 247-262. doi:10.1093/biostatistics/1.3.247
- Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice*, 17(2), 136-137. doi:10.1037/0735-7028.17.2.136
- Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, 3(4), 286-300. doi:10.1111/j.1745-6924.2008.00079.x

- Du, H., Liu, F., & Wang, L. (2017). A Bayesian "fill-In" method for correcting for publication bias in meta-analysis. *Psychological Methods*, 22(4), 799-817. doi:10.1037/met0000164
- Duval, S., & Tweedie, R. L. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95(449), 89-98. doi:10.1080/01621459.2000.10473905
- Duval, S., & Tweedie, R. L. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2), 455-463.
- Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. New York, NY: Springer.
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629-634.
- Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLOS ONE*, 5(4), e10068. doi:10.1371/journal.pone.0010068
- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3), 891-904. doi:10.1007/s11192-011-0494-7
- Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 63(3), 665-694. doi:10.1348/000711010X502733
- Fisher, R. A. (1925). *Statistical methods for research workers* (1st ed.). London: Oliver & Boyd.
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502-1505. doi:10.1126/science.1255484
- Guan, M., & Vandekerckhove, J. (2015). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin and Review*, Advance online publication. doi:10.3758/s13423-015-0868-6
- Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal*, 41(8), 901-916.
- Hartung, J., & Knapp, G. (2001a). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*, 20(12), 1771-1782. doi:10.1002/sim.791
- Hartung, J., & Knapp, G. (2001b). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, 20(24), 3875-3889. doi:10.1002/sim.1009
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107-128.
- Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, 93(2), 388-395. doi:10.1037//0033-2909.93.2.388
- 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(1), 61-85.
- Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, 7(2), 246-255.
- Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, 21(4), 299-332.
- Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester: UK: Wiley.

- Higgins, J. P. T. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, 37(5), 1158-1160. doi:10.1093/ije/dyn204
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327(7414), 557-560. doi:10.1136/bmj.327.7414.557
- Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society*, 172(1), 137-159.
- Hong, S., & Reed, W. R. (2021). Using Monte Carlo experiments to select meta-analytic estimators. *Research Synthesis Methods*, 12(2), 192-215. doi:10.1002/jrsm.1467
- IntHout, J., Ioannidis, J. P., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, 14. doi:10.1186/1471-2288-14-25
- Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4(3), 245-253. doi:10.1177/1740774507079441
- Iyengar, S., & Greenhouse, J. B. (1988a). Selection models and the file drawer problem. *Statistical Science*, 3(1), 109-117. doi:10.1214/ss/1177013012
- Iyengar, S., & Greenhouse, J. B. (1988b). Selection models and the file drawer problem: Rejoinder. *Statistical Science*, 3(1), 133-135. Retrieved from <http://www.jstor.org/stable/2245932>
- Jackson, D. (2006). The implications of publication bias for meta-analysis' other parameter. *Statistics in Medicine*, 25(17), 2911-2921. doi:10.1002/sim.2293
- Jackson, D. (2007). Assessing the implications of publication bias for two popular estimates of between-study variance in meta-analysis. *Biometrics*, 63(1), 187-193. doi:10.1111/j.1541-0420.2006.00663.x
- Jackson, D. (2013). Confidence intervals for the between-study variance in random effects meta-analysis using generalised Cochran heterogeneity statistics. *Research Synthesis Methods*, 4(3), 220-229. doi:10.1002/jrsm.1081
- Jin, Z. C., Zhou, X. H., & He, J. (2014). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine*, 34(2), 343-360. doi:10.1002/sim.6342
- Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods*, 15(4), 624-662. doi:10.1177/1094428112452760
- Kicinski, M. (2013). Publication bias in recent meta-analyses. *PLOS ONE*, 8(11). doi:10.1371/journal.pone.0081823
- Kraemer, H. C., Gardner, C., Brooks, J., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods*, 3(1), 23-31. doi:10.1037/1082-989X.3.1.23
- Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical & Statistical Psychology*, 31, 107-112.
- Langan, D., Higgins, J. P. T., & Simmonds, M. (2016). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. *Research Synthesis Methods*, 8(2), 181-198. doi:10.1002/jrsm.1198
- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

- McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, 11(5), 730-749. doi:10.1177/1745691616662243
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The Prisma Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine*, 6(7), e1000097. doi:10.1371/journal.pmed.1000097
- Niemeyer, H., van Aert, R. C. M., Schmid, S., Uelsmann, D., Knaevelsrud, C., & Schulte-Herbrüggen, O. (2020). Publication bias in meta-analyses of posttraumatic stress disorder interventions. *Meta-Psychology*, 4. doi:10.15626/MP.2018.884
- Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics*, 8(2), 157-159.
- Paule, R. C., & Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, 87(5), 377-385.
- Pawitan, Y. (2013). *In all likelihood : Statistical modelling and inference using likelihood*. Oxford: OUP Oxford.
- R Core Team. (2022). R: A language and environment for statistical computing. Retrieved from <http://www.R-project.org/>
- Rabelo, A. L. A., Keller, V. N., Pilati, R., & Wicherts, J. M. (2015). No effect of weight on judgments of importance in the moral domain and evidence of publication bias from a meta-analysis. *PLOS ONE*, 10(8), e0134808. doi:10.1371/journal.pone.0134808
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638-641.
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, UK: Wiley.
- Röver, C., Knapp, G., & Friede, T. (2015). Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology*, 15. doi:10.1186/s12874-015-0091-1
- Sidik, K., & Jonkman, J. N. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine*, 21(21), 3153-3159. doi:10.1002/sim.1262
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359-1366. doi:10.1177/0956797611417632
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9(6), 666-681. doi:10.1177/1745691614553988
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2017, December 20). Why p-curve excludes $p > .05$ [Web log message]. Retrieved from <http://datacolada.org/61>
- Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1), 60-78.
- Stanley, T. D., Jarrell, S. B., & Doucouliagos, H. (2010). Could it be better to discard 90% of the data? A statistical paradox. *The American Statistician*, 64(1), 70-77. doi:10.1198/tast.2009.08205
- Sutton, A. J., Song, F., Gilbody, S. M., & Abrams, K. R. (2000). Modelling publication bias in meta-analysis: A review. *Statistical Methods in Medical Research*, 9(5), 421-445. doi:10.1177/096228020000900503

- van Aert, R. C. M. (2022). *puniform*: Meta-analysis methods correcting for publication bias. (Version 0.2.5). Retrieved from <https://CRAN.R-project.org/package=puniform>
- van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses on p-values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*, 11(5), 713-729. doi:10.1177/1745691616650874
- van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLOS ONE*, 14(4). doi:10.1371/journal.pone.0215052
- van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, 20(3), 293-309. doi:10.1037/met0000025
- van Erp, S. J., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990-2013. *Journal of Open Psychology Data*, 5(1). doi:10.5334/jopd.33
- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., . . . Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1), 55-79. doi:10.1002/jrsm.1164
- Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60(3), 419-435. doi:10.1007/bf02294384
- Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, 10(4), 428-443. doi:10.1037/1082-989X.10.4.428
- Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, 30(3), 261-293. doi:10.3102/10769986030003261
- Viechtbauer, W. (2007a). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. *Journal of Educational and Behavioral Statistics*, 32(1), 39-60. doi:10.3102/1076998606298034
- Viechtbauer, W. (2007b). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, 26(1), 37-52. doi:10.1002/sim.2514
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48. doi:10.18637/jss.v036.i03
- Wicherts, J. M., Veldkamp, C. L. S., Augusteyn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7(1832). doi:10.3389/fpsyg.2016.01832
- Wiksten, A., Rücker, G., & Schwarzer, G. (2016). Hartung-Knapp method is not always conservative compared with fixed-effect meta-analysis. *Statistics in Medicine*, 35(15), 2503-2515. doi:10.1002/sim.6879

Footnotes

¹ Simonsohn et al. (2014) argue that *p*-curve accurately estimates effect size in the presence of heterogeneity, because it estimates the average true effect size of exactly the studies included in a meta-analysis. This is in contrast to conventional random-effects meta-analysis where the included studies are assumed to be a random sample of a population of studies and the estimated parameter is the mean of this population (Borenstein et al., 2009). See van Aert et al. (2016) for an elaborate discussion.

² Effect sizes in *p*-uniform in its current implementation can be estimated using six different methods: maximum likelihood, and using the Irwin-Hall distribution, Fisher's test, the Kolmogorov-Smirnov test, Anderson-Darling test, and a variant of Fisher's test, $-\sum_{i=1}^k \ln(1 - q_i)$. All methods are implemented in the R package "puniform".

³ *P*-uniform* reduces to *p*-uniform where μ is estimated with maximum likelihood estimation if τ^2 is constrained to zero.

⁴ We conducted an additional Monte-Carlo simulation study to examine whether varying the primary study's sample size influenced the results (R code <https://osf.io/d4tq9>). The same variables were varied in this simulation study as in the one described above where was made sure that all simulated meta-analyses contained at least one statistically (non)significant effect size. The only difference between the two simulation studies was that *k* was fixed to 30 rather than systematically varied. Sample sizes were varied such that the median sample size per group of the primary studies included in a meta-analysis was equal to 50; ten studies had 25 observations per group, eight had 50 per group, six had 100 per group, four had 150 per group, and two had 300 per group. We only report the results of this Monte-Carlo simulation study in the supplemental

materials (<https://osf.io/gycf2/>), because the results of this simulation study were close to those with fixed sample size.

⁵ Confidence intervals of the random-effects model for μ were computed using the adjustment proposed by Hartung and Knapp (Hartung, 1999; Hartung & Knapp, 2001a, 2001b) and Sidik and Jonkman (Sidik & Jonkman, 2002), because coverage probability after applying this adjustment is closer to the nominal coverage rate (IntHout, Ioannidis, & Borm, 2014; Röver, Knapp, & Friede, 2015; Wiksten, Rucker, & Schwarzer, 2016). Veroniki et al. (2016) (Veroniki et al., 2016) reviewed existing methods to compute a confidence interval for τ^2 and concluded that the *Q*-profile method (Viechtbauer, 2007b) and generalized *Q*-statistic method (Jackson, 2013) are the two methods with the best statistical properties. We decided to include the *Q*-profile method in our Monte-Carlo simulations, because this method does not require arbitrary choices with respect to the primary study's weights as compared to the generalized *Q*-statistic method.

⁶ We also examined the statistical properties of *p*-uniform* and 3PSM as implemented in the R package “weightr” (Coburn & Vevea, 2016) by means of a numerical study. In this numerical study, we examined the most extreme case of a meta-analysis with just one statistically significant and one nonsignificant observed effect size. This numerical study is described in Supplement 3 at <https://osf.io/jngwk>, and all its results are reported in <https://osf.io/gycf2>.

Table 1 *Results of the simulations without the requirement of having at least one statistically significant and nonsignificant effect size in the meta-analysis for the conditions $\mu = 0$, $\tau = 0.346$, $k = 10$, and $pub = 0$ and 1 . The included methods are the random-effects model (RE), *p*-uniform, *p*-uniform* using maximum likelihood estimation (ML), and the three-parameter selection model (3PSM) as implemented in the R packages “weightr” and “metafor”. The non-convergence rate are shown in the first two columns. The mean and median (between brackets) of the estimates of μ and τ are shown in the last four columns, and these are based on all simulation runs where a particular method reached convergence.*

<i>pub</i>	Non-convergence rate		$\hat{\mu}$		$\hat{\tau}$	
	0	1	0	1	0	1
RE model	0%	0%	-0.001 (0.000)	0.605 (0.600)	0.324 (0.325)	0.030 (0.000)
<i>p</i> -uniform	17.2%	0%	0.060 (0.465)	0.457 (0.470)	-	-
<i>p</i> -uniform* (ML)	0%	0%	4.451 (-0.017)	-575.299 (0.320)	0.328 (0.290)	3.592 (0.178)
3PSM (weightr)	4.3%	0%	0.020 (-0.017)	0.378 (0.386)	0.297 (0.290)	0.130 (0.110)
3PSM (metafor)	17.2%	100%	0.005 (-0.015)	-	0.294 (0.292)	-

Table 2 Coverage probability of the confidence interval for μ computed with the random-effects model (RE), *p*-uniform, *p*-uniform* using maximum likelihood estimation (ML), and the three-parameter selection model (3PSM) as implemented in the R packages “weightr” and “metafor”. The coverage probabilities are shown as a function of μ , τ , the severity of publication bias (*pub*), and the number of primary studies’ observed effect sizes (*k*). Coverage probabilities between 0.9 and 0.975 are marked in bold.

		<i>k</i> =10											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	<i>pub</i>	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
$\tau=0$	RE model	0.960	0.957	0.867	0.000	0.972	0.920	0.354	0.003	0.960	0.965	0.926	0.924
	<i>p</i> -uniform	0.941	0.946	0.949	0.948	0.952	0.950	0.951	0.948	0.949	0.952	0.945	0.947
	<i>p</i> -uniform* (ML)	0.952	0.955	0.957	0.947	0.950	0.955	0.957	0.946	0.954	0.952	0.940	0.941
	3PSM (weightr)	0.961	0.966	0.965	0.982	0.956	0.955	0.962	0.983	0.957	0.968	0.973	0.973
	3PSM (metafor)	0.969	0.974	0.968	0.986	0.969	0.968	0.964	0.978	0.965	0.975	0.973	0.977
$\tau=0.163$	RE model	0.967	0.952	0.657	0.000	0.966	0.897	0.290	0.009	0.959	0.945	0.871	0.867
	<i>p</i> -uniform	0.896	0.888	0.870	0.802	0.893	0.886	0.848	0.824	0.907	0.897	0.893	0.893
	<i>p</i> -uniform* (ML)	0.923	0.925	0.924	0.858	0.917	0.922	0.908	0.871	0.928	0.921	0.914	0.911
	3PSM (weightr)	0.946	0.945	0.946	0.957	0.927	0.928	0.949	0.967	0.940	0.944	0.948	0.948
	3PSM (metafor)	0.951	0.949	0.938	0.942	0.932	0.931	0.936	0.935	0.944	0.947	0.952	0.956
$\tau=0.346$	RE model	0.972	0.917	0.353	0.003	0.958	0.875	0.251	0.030	0.956	0.915	0.740	0.704
	<i>p</i> -uniform	0.657	0.607	0.371	0.231	0.634	0.555	0.369	0.316	0.663	0.601	0.567	0.564
	<i>p</i> -uniform* (ML)	0.901	0.904	0.867	0.712	0.882	0.895	0.806	0.734	0.902	0.872	0.834	0.827
	3PSM (weightr)	0.919	0.920	0.904	0.832	0.892	0.902	0.870	0.844	0.916	0.904	0.888	0.882
	3PSM (metafor)	0.920	0.920	0.896	0.835	0.894	0.902	0.863	0.834	0.917	0.905	0.890	0.886

Table 2 Continued

		<i>k</i> =60											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	<i>pub</i>	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
$\tau=0$	RE model	0.958	0.932	0.299	0.000	0.950	0.707	0.000	0.000	0.947	0.665	0.012	0.000
	<i>p</i> -uniform	0.951	0.952	0.951	0.951	0.951	0.949	0.946	0.949	0.947	0.948	0.946	0.947
	<i>p</i> -uniform* (ML)	0.952	0.951	0.953	0.837	0.943	0.951	0.952	0.827	0.944	0.925	0.862	0.832
	3PSM (weightr)	0.968	0.966	0.958	0.160	0.980	0.968	0.960	0.955	0.959	0.977	0.998	0.999
	3PSM (metafor)	0.962	0.960	0.956	0.998	0.971	0.961	0.951	0.996	0.955	0.962	0.989	0.994
$\tau=0.163$	RE model	0.953	0.876	0.012	0.000	0.951	0.613	0.000	0.000	0.950	0.618	0.004	0.000
	<i>p</i> -uniform	0.871	0.827	0.580	0.230	0.788	0.687	0.457	0.329	0.760	0.714	0.673	0.670
	<i>p</i> -uniform* (ML)	0.922	0.932	0.946	0.503	0.907	0.939	0.912	0.547	0.938	0.915	0.810	0.753
	3PSM (weightr)	0.949	0.947	0.948	0.840	0.932	0.942	0.944	0.972	0.952	0.962	0.966	0.968
	3PSM (metafor)	0.950	0.947	0.948	0.943	0.934	0.942	0.945	0.947	0.952	0.962	0.963	0.965
$\tau=0.346$	RE model	0.951	0.684	0.000	0.000	0.951	0.504	0.000	0.000	0.950	0.544	0.000	0.000
	<i>p</i> -uniform	0.238	0.076	0.001	0.000	0.083	0.018	0.000	0.000	0.086	0.043	0.016	0.011
	<i>p</i> -uniform* (ML)	0.911	0.932	0.931	0.500	0.923	0.940	0.865	0.542	0.936	0.915	0.775	0.680
	3PSM (weightr)	0.940	0.945	0.942	0.776	0.937	0.940	0.938	0.863	0.945	0.951	0.944	0.928
	3PSM (metafor)	0.940	0.945	0.942	0.856	0.937	0.940	0.938	0.877	0.945	0.951	0.944	0.928

Table 3 Coverage probability of the confidence interval for τ^2 computed with the random-effects model (RE), p -uniform* using maximum likelihood estimation (ML), and the three-parameter selection model (3PSM) as implemented in the R package “metafor”. The coverage probabilities are shown as a function of μ , τ , the severity of publication bias (pub), and the number of primary studies’ observed effect sizes (k). Coverage probabilities between 0.9 and 0.975 are marked in bold.

		$k=10$											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	pub	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
$\tau=0$	RE model	0.931	0.922	0.775	0.824	0.966	0.949	0.910	0.758	0.956	0.951	0.941	0.935
	p -uniform* (ML)	0.983	0.982	0.985	0.969	0.971	0.982	0.985	0.964	0.984	0.972	0.964	0.963
	3PSM (metafor)	0.991	0.991	0.990	0.991	0.990	0.993	0.991	0.992	0.991	0.988	0.991	0.991
$\tau=0.163$	RE model	0.964	0.939	0.818	0.778	0.959	0.947	0.894	0.750	0.957	0.942	0.923	0.920
	p -uniform* (ML)	0.971	0.971	0.976	0.962	0.972	0.980	0.979	0.961	0.984	0.972	0.964	0.962
	3PSM (metafor)	0.988	0.986	0.978	0.991	0.988	0.987	0.989	0.991	0.990	0.988	0.992	0.991
$\tau=0.346$	RE model	0.968	0.946	0.903	0.752	0.956	0.950	0.863	0.771	0.952	0.945	0.904	0.900
	p -uniform* (ML)	0.902	0.915	0.893	0.769	0.903	0.914	0.842	0.781	0.911	0.892	0.845	0.847
	3PSM (metafor)	0.930	0.931	0.925	0.922	0.927	0.927	0.927	0.931	0.933	0.943	0.941	0.945

Table 3 Continued

		$k=10$											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	pub	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
$\tau=0$	RE model	0.965	0.917	0.159	0.000	0.953	0.887	0.890	0.000	0.953	0.845	0.246	0.099
	p -uniform* (ML)	0.980	0.981	0.988	0.867	0.968	0.979	0.986	0.862	0.986	0.973	0.908	0.874
	3PSM (metafor)	0.988	0.986	0.991	0.987	0.987	0.988	0.989	0.990	0.984	0.987	0.988	0.986
$\tau=0.163$	RE model	0.952	0.876	0.212	0.000	0.950	0.915	0.836	0.000	0.953	0.860	0.217	0.061
	p -uniform* (ML)	0.926	0.934	0.943	0.484	0.912	0.934	0.904	0.498	0.931	0.899	0.781	0.710
	3PSM (metafor)	0.944	0.944	0.937	0.984	0.941	0.942	0.939	0.985	0.952	0.955	0.964	0.968
$\tau=0.346$	RE model	0.955	0.888	0.895	0.000	0.952	0.942	0.556	0.001	0.949	0.875	0.200	0.030
	p -uniform* (ML)	0.915	0.937	0.927	0.502	0.927	0.937	0.865	0.547	0.934	0.911	0.768	0.668
	3PSM (metafor)	0.949	0.949	0.939	0.942	0.946	0.940	0.939	0.954	0.947	0.951	0.954	0.956

Table 4 *Results of applying the random-effects meta-analysis model (RE), p-uniform, p-uniform* using maximum likelihood (ML) estimation, and 3PSM as implemented in the R package “weightr” and “metafor” to the data of the meta-analysis Rabelo et al. (2015) (first five rows) and Bangert-Drowns et al. (2004) (last five rows). Between-study variance in the random-effects meta-analysis model was estimated using the Paule-Mandel estimator.*

	$\hat{\mu}$	(95% CI μ)	$\hat{\tau}^2$	(95% CI τ^2)
Rabelo et al. (2015)				
RE model	0.571	(0.524;0.618)	0.000	(0.000;0.000)
<i>p</i> -uniform	-0.179	(-0.676;0.159)	-	-
<i>p</i> -uniform* (ML)	0.244	(0.038;0.436)	0.000	(0.000;0.018)
3PSM (weightr)	0.254	(0.220;0.289)	0.000	-
3PSM (metafor)	0.254	-	0.000	(0.000;0.019)
Bangert-Drowns et al. (2004)				
RE model	0.228	(0.126;0.330)	0.069	(0.027;0.153)
<i>p</i> -uniform	0.245	(-0.236;0.531)	-	-
<i>p</i> -uniform* (ML)	0.179	(0.068;0.297)	0.027	(0.006;0.073)
3PSM (weightr)	0.148	(0.004;0.291)	0.028	-
3PSM (metafor)	0.148	(0.004;0.291)	0.028	(0.000;0.095)

Note. CI refers to confidence interval. *P*-uniform does not estimate and computes a CI for τ^2 . “weightr” does not compute a CI for τ^2 . “metafor” could not compute a CI for μ for Rabelo et al. (2015).

Figure 1. Average of the estimates of μ for the random-effects model (RE), p -uniform, p -uniform* using maximum likelihood estimation (ML), and the three-parameter selection model (3PSM) as implemented in the R packages “weightr” and “metafor”. The average of the estimates of μ is shown as a function of μ , τ , and the severity of publication bias (pub) with the number of primary studies’ observed effect sizes (k) equal to 10.

Figure 2. Average of the estimates of μ for the random-effects model (RE), p -uniform, p -uniform* using maximum likelihood estimation (ML), and the three-parameter selection model (3PSM) as implemented in the R packages “weightr” and “metafor”. The average of the estimates of μ is shown as a function of μ , τ , and the severity of publication bias (pub) with the number of primary studies’ observed effect sizes (k) equal to 60.

Figure 3. Root mean square error (RMSE) of estimating μ for the random-effects model (RE), p -uniform, p -uniform* using maximum likelihood estimation (ML), and the three-parameter selection model (3PSM) as implemented in the R packages “weightr” and “metafor”. The RMSE of estimating μ is shown as a function of μ , τ , and the severity of publication bias (pub) with the number of primary studies’ observed effect sizes (k) equal to 10. The RMSE was multiplied by 1,000 to facilitate the interpretation.

Figure 4. Root mean square error (RMSE) of estimating μ for the random-effects model (RE), p -uniform, p -uniform* using maximum likelihood estimation (ML), and the three-parameter selection model (3PSM) as implemented in the R packages “weightr” and “metafor”. The RMSE of estimating μ is shown as a function of μ , τ , and the severity of publication bias (pub) with the number of primary studies’ observed effect sizes (k) equal to 60. The RMSE was multiplied by 1,000 to facilitate the interpretation.

Figure 5. Average of the estimates of τ for the random-effects model (RE), p -uniform* using maximum likelihood estimation (ML), and the three-parameter selection model (3PSM) as implemented in the R packages “weightr” and “metafor”. The average of the estimates of τ is shown as a function of μ , τ , and the severity of publication bias (pub) with the number of primary studies’ observed effect sizes (k) equal to 10.

Figure 6. Average of the estimates of τ for the random-effects model (RE), p -uniform* using maximum likelihood estimation (ML), and the three-parameter selection model (3PSM) as implemented in the R packages “weightr” and “metafor”. The average of the estimates of τ is shown as a function of μ , τ , and the severity of publication bias (pub) with the number of primary studies’ observed effect sizes (k) equal to 60.

Figure 7. Root mean square error (RMSE) of estimating τ for the random-effects model (RE), p -uniform* using maximum likelihood estimation (ML), and the three-parameter selection model (3PSM) as implemented in the R packages “weightr” and “metafor”. The RMSE of estimating τ is shown as a function of μ , τ , and the severity of publication bias (pub) with the number of primary studies’ observed effect sizes (k) equal to 10. The RMSE was multiplied by 1,000 to facilitate the interpretation.

Figure 8. Root mean square error (RMSE) of estimating τ for the random-effects model (RE), p -uniform* using maximum likelihood estimation (ML), and the three-parameter selection model (3PSM) as implemented in the R packages “weightr” and “metafor”. The RMSE of estimating τ is shown as a function of μ , τ , and the severity of publication bias (pub) with the number of primary studies’ observed effect sizes (k) equal to 60. The RMSE was multiplied by 1,000 to facilitate the interpretation.