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

Robbie C. M. van Aert[1] and Marcel A. L. M. van Assen[1,2]

[1]Tilburg University [2]Utrecht 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. *P*-uniform is a meta-analysis method that corrects estimates for publication bias but overestimates average effect size if heterogeneity in true effect sizes (i.e., between-study variance) is present. We propose an extension and improvement of *p*-uniform called *p*-uniform*. *P*-uniform* improves upon *p*-uniform in three important ways, as it (i) entails a more efficient estimator, (ii) eliminates the overestimation of effect size in case of 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, the selection model approach of Hedges (1992), and the random-effects model. Statistical properties of *p*-uniform* and the selection model approach were comparable and generally outperformed *p*-uniform and the random-effects model if publication bias was present. We demonstrate that *p*-uniform* and the selection model approach 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. *P*-uniform* generally provides more accurate estimates of the between-study variance in meta-analyses containing many studies (e.g., 60 or more) and if publication bias is present. However, both methods do not perform well if the meta-analysis only includes statistically significant studies. *P*-uniform performed best in this case but only when between-study variance was zero or small. We offer recommendations for applied researchers, and provide an R package and 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. Meta-analysis is now seen as the "gold standard" for synthesizing evidence from multiple studies studying the same relationship (Aguinis, Gottfredson, & Wright, 2011; Head, Holman, Lanfear, Kahn, & Jennions, 2015). However, a major threat to the validity of a meta-analysis is publication bias (e.g., 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 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. For example, publication bias causes an overrepresentation of false positives in the published literature (van Assen, van Aert, & Wicherts, 2015), because false positives are more easily published than statistically nonsignficant results. Publication bias also 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 (e.g., 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., 2019) 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).

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 et al., 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 publication bias tests have low statistical power especially if the number of effect sizes in a meta-analysis is small (Begg & Mazumdar, 1994; Sterne, Gavaghan, & Egger, 2000; van Assen et al., 2015). Furthermore, we believe that from the perspective of an applied researcher it is more relevant to know what the effect size is corrected for publication bias than to know that publication bias distorted the results of a meta-analysis without knowing the consequences on effect size estimation. More specifically, we focus in this paper on selection model approaches to correct for publication bias, because recently published work suggests that these methods have better statistical properties than other methods (Carter, Schönbrodt, Gervais, & Hilgard, 2019; Du, Liu, & Wang, 2017; McShane, Böckenholt, & Hansen, 2016), and are even nowadays called the state-of-the-art methods to correct for publication bias (McShane et al., 2016).

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, via those that assume a specific selection model and from frequentist to Bayesian approaches (e.g., 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. Here we focus on the selection model approach of Hedges (1992), which performance is examined below.

Selection model approaches have hardly been used in meta-analyses (Hunter & Schmidt, 2015), because these methods require the user to make sophisticated assumptions and choices (Borenstein et al., 2009) and are often not implemented in user-friendly software for applying these methods. However, easy-to-use software has recently been developed that can be used for applying several types of selection model approaches (R packages "weightr" [Coburn & Vevea, 2016] , "selectMeta" [Rufibach, 2015] , and "metasens" [Schwarzer, Carpenter, & Rücker, 2016]).

Two recently developed methods to correct effect sizes for publication bias are *p*-uniform (van Aert et al., 2016; van Assen et al., 2015) and *p*-curve (Simonsohn et al., 2014). *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. Estimation is only based on the statistically significant effect sizes and nonsignificant effect sizes are discarded. For that reason, conditional probabilities (i.e., *p*-values conditional on being statistically significant) are evaluated for being uniformly distributed instead of the traditional *p*-values.

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 size (Carter et al., 2019; McShane et al., 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 size, and (iii) enabling estimation and testing for the presence of the between-study variance in true effect size.

The goal of this paper is twofold; we introduce the new method *p*-uniform* and examine the statistical properties of *p*-uniform*, *p*-uniform, and the selection model approach of Hedges (1992) 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 the selection model approach of Hedges (1992) for five reasons. First, it allows us to examine the conditions where *p*-uniform* is an improvement over *p*-uniform. Second, Hedges' method enables estimation of both the effect size as well as the between-study variance in true effect. Third, this selection model approach assumes that the selection model is unknown and has to be estimated which is more realistic than other methods (e.g., Vevea & Woods, 2005) that assume that the selection model is known. Fourth, easy-to-use software is available for applying this method in the R package "weightr" (Coburn & Vevea, 2016) and this method suffers less from convergence problems than for instance the selection model approach proposed by Copas and colleagues (Copas, 1999; Copas & Shi, 2000, 2001). Finally, statistically significant and nonsignificant effect sizes can be included in this selection model approach whereas other selection model approaches only use the significant effect sizes (e.g., Hedges, 1984).

Selection model approach of Hedges (1992)

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 fixed-effect (a.k.a. the equal-effects or common-effects model) or random-effects model. The random-effects model assumes that *k* independent effect size estimates, y_i with *i*=1, …, *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 $\zeta_i \sim \mathrm{N}(0, \tau^2)$ where τ^2 reflects the between-study variance in true effects, and $\varepsilon_i \sim N(0, \sigma_i^2)$ 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 size, and the random-effects model is equal to the fixed-effect model.

The selection model is a non-negative weight function that determines the likelihood of a primary study getting published (Hedges & Vevea, 2005). The selection model of Hedges (1992) is a step function that creates intervals of *p*-values where *p*-values in the same interval get the same weight in the weight function. The probability of publication for each interval of *p*-values is estimated and these probabilities are used in the weight function. 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 in order to make sure that the model is identified.

The weight function, $w(y_i, \sigma_i)$, is combined with the effect size model to get a weighted density of y_i,

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

where $f(y_i, \sigma_i)$ denotes the (unweighted) density distribution as in the fixed-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 fixed-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.

The selection model approach by Hedges (1992) was later extended to enable the inclusion of predictor variables (see Vevea & Hedges, 1995) such that the effect size model is a mixed-effects (a.k.a. meta-regression) model instead of a random-effects model (Borenstein et al., 2009). This selection model approach of Hedges (1992) is implemented in the R package “weightr”. In order to avoid non-convergence of this method, the weight of an interval is set equal to 0.01 in the R package if there are no observations in an interval. This is, of course, a choice made in the implementation of this selection model approach and does not pertain to the approach itself.

Hedges and Vevea (1996) conducted a simulation study to examine the statistical properties of the selection model approach of Hedges (1992) and concluded that the weights of the selection model are often poorly estimated, but that the estimates of effect size and between-study variance were then still quite accurate. Furthermore, they assessed whether non-normally distributed random effects bias the estimates of the selection model approach and concluded that the approach is relatively robust to violations of the normality assumption.

A recent simulation study by Carter et al. (2019) compared the selection model approach of Hedges (1992), *p*-uniform, *p*-curve, trim-and-fill, weighted average of the adequately powered studies, and PET-PEESE and concluded that none of the methods consistently outperformed all the others. McShane et al. (2016) also conducted a Monte-Carlo simulation study by comparing the selection model approach of Iyengar and Greenhouse (1988b) with *p*-uniform and *p*-curve. They concluded that the selection model approach of Iyengar and Greenhouse (1988b) yields the best performance and should be preferred over the other methods, but also noted that the assumptions underlying this selection model approach are idealistic and unlikely to be met in practice. Du et al. (2017) proposed a Bayesian implementation of the selection model approach of Hedges (1992) using non-informative or weakly informative prior distributions (BALM). They concluded that BALM had better statistical properties than trim-and-fill, PET-PEESE, *p*-uniform, and the selection model approach of Hedges (1992) except for slightly more bias in the estimate of μ than the selection model approach.

Another simulation study (Terrin, Schmid, Lau, & Olkin, 2003) studied the statistical properties of Hedges' (1992) selection model approach and compared it to trim-and-fill. Although this selection model approach outperformed trim-and-fill, it often failed to converge. Non-convergence of Hedges' (1992) approach occurs when there are no *p*-values observed in one of the specified intervals of *p*-values. Although recent studies (e.g., Carter et al., 2019; Du et al., 2017; McShane et al., 2016) suggest less convergence problems and better performance of selection model approaches, it is currently unknown how these methods perform in meta-analyses with a small number of primary studies' effect sizes in combination with extreme publication bias. Especially such meta-analyses are quite common in the psychological literature that shows signs of strong publication bias as evidenced by over 95% articles showing positive outcomes (Fanelli, 2010). Moreover, the median number of studies in meta-analyses in psychology and medicine equals 12 and 3, respectively (Rhodes, Turner, & Higgins, 2015; Turner, Jackson, Wei, Thompson, & Higgins, 2015; van Erp, Verhagen, Grasman, & Wagenmakers, 2017). Hence, we also study in this paper the statistical properties of selection model approaches for conditions that are realistic for meta-analyses in practice, even though these conditions may at the same time be challenging for these methods.

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 published. Statistical significance is taken into account - and hence publication bias is corrected - 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

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

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 fixed-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 constant probability to get published, but without the need to estimate this probability.

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 size.

Another proposed estimator[2] 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). Hence, we recommended to only interpret *p*-uniform's effect size estimate as the estimate of the average population effect size when heterogeneity is zero or small (van Aert et al., 2016). Moreover, methods that do not assume that the true effect size is fixed are favorable, because heterogeneity is often present in meta-analyses (e.g., Higgins, 2008; Higgins, Thompson, & Spiegelhalter, 2009; van Erp et al., 2017).

The second critique by McShane et al. (2016) relates to the loss of efficiency of *p*-uniform's estimators (van Aert et al., 2016; van Assen et al., 2015) 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 (e.g., Fanelli, 2010; Fanelli, 2012). Nevertheless, as meta-analyses including many nonsignificant effects do occur (e.g.,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*. Simonsohn, Simmons, and Nelson (2017, December 20), however, argue that nonsignificant effects should not be included in *p*-curve (and therefore also not in *p*-uniform). They argue that all nonsignificant effects do not have the same probability of getting published, and it is hard to make assumptions about this probability. However, we contend that the benefits of including statistically nonsignificant effect sizes (more efficient estimator, less biased estimator if the true effect size is heterogeneous, and enabling estimation of the between-study variance in true effect size) outweigh the potential costs (possible bias in the estimator if the assumption of equal probability for publishing nonsignificant effect sizes is violated).

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 (e.g., 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). 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). 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,

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

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

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

The ‘weights’ of *p*-uniform* are included or 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. An advantage of *p*-uniform* over the selection model approach is that it is more parsimonious, because estimation of the weights is not required in *p*-uniform*.

The profile (log-)likelihood functions of Equation (4) can be iteratively optimized until $\hat{\mu}$ and $\hat{\tau}^2$ do not change anymore in consecutive steps. Confidence intervals for μ and τ^2 are obtained 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/).

Analytical approximation

We examined the statistical properties of *p*-uniform* and the selection model of Hedges (1992) implemented in the R package "weightr" (Coburn & Vevea, 2016; henceforth called Hedges1992) both by means of an analytical approximation and a simulation study. In the analytical approximation we examined the most extreme case of a meta-analysis with just one statistically significant and one nonsignificant observed effect size. This seems to be a rather extreme situation for a meta-analysis, but many meta-analyses only contain a small number of primary studies' effect sizes (median number of primary studies' effect sizes in meta-analyses in the Cochrane Database of Systematic Reviews is 3 (Rhodes et al., 2015; Turner et al., 2015).

The analytical study (see supplement 3 in the supplementary materials, https://osf.io/jngwk/, for method and results) demonstrated that convergence for estimating μ and τ and their confidence interval was not a problem for *p*-uniform* and hardly a problem for Hedges1992 in these very challenging conditions. This invalidates the critique on the selection model approach that at least 100 primary studies' effect sizes are required for estimates to converge (Field & Gillett, 2010; Hedges & Vevea, 2005; Vevea & Woods, 2005). As τ was underestimated and overcoverage of *p*-uniform*'s confidence interval for τ was severe, we concluded that two studies are (unsurprisingly) not sufficient for estimating τ and its confidence interval.

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

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

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 α=.025 was used to resemble common practice of research in psychology were a two-tailed test with α=.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.

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 size (τ) was 0, 0.163, or 0.346 representing I^2-statistics equal to 0%, 40%, and 75% (zero, small-medium, large [Higgins & Thompson, 2002]). 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, whereas 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 x 3 x 4 x 4 = 144 conditions. For each condition 10,000 replications were conducted.[3]

P-uniform* and Hedges1992 with two intervals and the threshold at $\alpha=.025$ 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). *P*-uniform with the recommended Irwin-Hall estimator (van Aert et al., 2016) was also included to study to what extent *p*-uniform* is an improvement over *p*-uniform. 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 size, because two recent 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.[4] Moreover, we also studied the Type I error rate and statistical power for the test of no effect with $\alpha = .05$.

The Monte-Carlo simulation study was programmed in R (R Core Team, 2020) and the packages "metafor" (Viechtbauer, 2010), "weightr" (Coburn & Vevea, 2016), and "puniform" (van Aert, 2020) were used for applying the random-effects model, selection model approach, and *p*-uniform respectively. Other R packages that were used to decrease the computing time of the simulations were the "parallel" package (R Core Team, 2020) for parallelizing the simulations and the "Rcpp" package (Eddelbuettel, 2013) for executing C++ functions. R code of this Monte-Carlo simulation study is available via https://osf.io/79k3p/.

Results

A requirement for applying *p*-uniform is that at least one statistically significant effect size is included in the meta-analysis. Consequently, *p*-uniform could be applied to only 22.5% of the runs in the condition with the least statistically significant effect sizes per meta-analysis (condition $\mu = 0$, $\tau = 0$, *k*=10, and $pub = 0$). However, the number of statistically significant effect sizes per meta-analysis increased as a function of μ, *τ*, *k*, and *pub*. Hence, *p*-uniform could be applied to more than 90% of the runs in 133 out of 144 conditions.

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 supplement 2 at https://osf.io/jngwk/ for a description of these results). 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 are available online (https://osf.io/gycf2/).

Average estimates of μ *P*-uniform* and Hedges1992 did not always converge whereas the random-effects model always obtained an estimate of μ. The reason for the non-convergence of *p*-uniform* was that the estimate of *p*-uniform* was equal to one of the boundaries of the parameter space. This non-convergence was most severe for the condition $\mu = 0$, $\tau = 0.346$, *k*=10, and $pub = 1$ (12.8%). Hedges1992 failed to converge in at most 19.6% of the replications for the condition $\mu = 0$, $\tau = 0$, *k*=10, and $pub = 0$. Both methods' non-convergence rate was close to zero if both statistically significant and nonsignificant primary studies' effect sizes were included in a meta-analysis.

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. Bias of *p*-uniform was small if $\tau = 0$ and $pub = 1$ (maximum bias -0.052). However, *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 and was at most -0.021 if the median rather than the average of the replications was used as outcome.

Hedges1992 and *p*-uniform* were less biased than the random-effects model with negligible bias if pub < 0.9. For $pub = 0.9$, Hedges1992 provided accurate average estimates (maximum bias 0.054). For $pub = 0.9$, *p*-uniform* also provided accurate average estimates for $\mu = 0$ and $\mu = 0.2$, but slightly underestimated μ when $\mu = 0.5$ in combination with $\tau = 0.163$ (bias = -0.069). Hedges1992 was severely positively biased in case of extreme publication bias (i.e., $pub = 1$; maximum bias 0.387), and this bias decreased in μ. In case of extreme publication bias, *p*-uniform* generally showed less severe bias than Hedges1992 (maximum bias 0.169), and tended to underestimate μ.

While bias in the random-effects model was unaffected by increasing the number of studies to 60 (see Figure 2), bias of the three other methods decreased slightly. For $pub = 1$,*P*-uniform was the least biased of all methods if $\tau = 0$; Hedges1992 still provided strongly overestimated estimates of μ if $\mu = 0$ (bias at most 0.358), whereas overestimation by *p*-

uniform* was substantially smaller (maximum bias 0.141). For *p*-uniform* and Hedges1992, bias was negligible for $pub < 1$ with maximum bias equal to -0.031 and -0.032, respectively.

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. For $pub = 0$ or $pub = 0.5$, the random-effects model had a lower RMSE than the other methods, because its bias was zero ($pub = 0$) or small ($pub = 0.5$) while at the same time the standard deviation of its estimates was lower than for the other methods (see https://osf.io/gycf2/). For severe publication bias ($pub \geq 0.9$), RMSE of the other methods was often smaller because the contribution of the higher bias of the random-effects model outweighed its higher precision.

For $pub < 1$, RMSE of *p*-uniform was larger than of Hedges1992 and *p*-uniform* if $\mu = 0$ or $\mu = 0.2$, probably because of being based on less studies (i.e., lower precision). For other conditions, RMSE of *p*-uniform was comparable to Hedges1992. Comparing Hedges1992 with *p*-uniform* showed a generally similar RMSE for $pub = 0$ and $pub = 0.5$. For $pub = 1$, *p*-uniform* had a much higher RMSE than Hedges1992, caused by a considerably larger standard deviation of the estimates of *p*-uniform* (see https://osf.io/gycf2/). This was a consequence of primary studies with *p*-values close to the α -level resulting in highly negative effect size estimates of *p*-uniform*.

As the standard deviation of estimates decreased as *k* increased, the RMSE decreased in *k* for all methods in all conditions (see Figure 4). Performance of all methods became more similar for $pub = 0$ and $pub = 0.5$. RMSE of *p*-uniform was comparable to or larger than RMSE of the other methods if $pub < 0.9$. For $pub = 1$, RMSE of *p*-uniform was the smallest or comparable to Hedges1992 except for conditions where $\tau = 0.346$. As bias mainly determined the RMSE for larger values of *k*, the RMSE of Hedges1992 (less bias) was generally smaller than of the random-effects model (most bias), and the RMSE of *p*-uniform* became smaller than the random-effects model and more comparable to Hedges1992 when *k* was increased.

Coverage probability of confidence interval for μ A confidence interval for μ could always be computed with the random-effects model but not with *p*-uniform* and Hedges1992. Non-convergence was at most 13.8% for *p*-uniform*'s confidence interval (condition $\mu = 0$, $\tau = 0.163$, *k*=10, $pub = 1$), and 29.3% for Hedges1992 (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.

Table 1 presents the coverage probability of the 95% confidence interval for μ if *k*=10 and *k*=60. Coverage probabilities of the random-effects model were equal to 0.95 in the absence of publication bias and decreased as a function of pub with coverage probabilities approaching 0. *P*-uniform's coverage probabilities were close to 0.95 if $\tau = 0$ even under extreme publication bias, but its coverage probability decreased as τ increased**.** *P*-uniform*'s coverage probabilities were close to 0.95 for $pub \leq 0.5$ and $\tau = 0$, and decreased as a function of pub and τ. Coverage probabilities of Hedges1992 were close to 0.95 in the absence of publication bias, but also decreased as a function of pub and τ. Undercoverage was, in general, more extreme for the random-effects model, Hedges1992, *p*-uniform than for *p*-uniform*.

If *k* was increased, undercoverage of the random-effects model became more severe, as detrimental effects of its bias were more pronounced for a larger number of studies. These results confirm that confidence intervals *p*-uniform yields exact confidence intervals if its assumptions are met and that performance of *p*-uniform* and Hedges1992 is not acceptable if only statistically significant results are present in the meta-analysis.

Testing null hypothesis of no effect Table 2 presents the Type I error rate and statistical power for testing the null hypothesis of no effect. The first four columns ($\mu = 0$) refer to the Type I error rate whereas the other columns illustrate the statistical power of the methods. The Type I error rate of the random-effects model was close to 0.05 in the absence of publication bias, but it increased as a function of pub with Type I error equal to 1 for $pub = 1$. Type I error also increased in *k*, whenever there was publication bias. The large Type I error rates were caused by the overestimation of effect size due to publication bias. This overestimation of effect size also caused deceivingly high statistical power of the random-effects model. *P*-uniform's Type I error rate was close to 0.05 if $\tau = 0$. It was too high if $\tau > 0$, and increased in *k*. Statistical power of *p*-uniform was typically large in case of heterogeneous true effect size, because of it was positively biased in these conditions.

P-uniform* better controlled the Type I error rate than the random-effects model and *p*-uniform with the Type I error rate being close to 0.05 except for conditions with extreme publication bias (*pub* = 1) and $\tau = 0.346$. The statistical power of *p*-uniform* decreased as a function of publication bias; even for *k*=60 power could be as low as .305 to detect a medium effect size of μ = 0.5 under extreme publication bias (*pub* = 1). Consequently, because of low statistical power we do not recommend using *p*-uniform* for testing the hypothesis of no effect under extreme publication bias (*pub* ≥ 0.9).

The Type I error rate of Hedges1992 was much too large under extreme publication bias (often > 0.8 for $pub = 1$, and converging to 1 if *k*=120, see https://osf.io/gycf2/). Hedges1992's overestimation of effect size under extreme publication bias also resulted in misleading high statistical power in that condition. Statistical power of *p*-uniform* and Hedges1992 were comparable under no or moderate publication bias (*pub* = 0 or 0.5), except when μ = 0.5; here, Hedges1992 outperforms *p*-uniform* with respect to power. All in all, we do not recommend using Hedges1992 for testing the hypothesis of no effect under extreme publication bias (*pub* ≥ 0.9) because of overestimation of effect size and too high Type I error.

Average estimates of τ Note that *p*-uniform does not estimate τ. Estimates of τ could always be obtained with the random-effects model, but *p*-uniform* and the Hedges1992 did not always converge with respect to estimating τ (non-convergence at most 12.8% and 19.6%, respectively). Figures 5 and 6 show the average estimates of τ for *k*=10 and 60, respectively. For *k*=10 and $pub = 0$, the random-effects model overestimated τ if $\tau = 0$ (maximum bias 0.051) and underestimated it if $\tau > 0$ (maximum bias -0.025). If $pub = 1$ and $\tau > 0$, the random-effects model underestimated τ for all conditions, because meta-analyses in this condition only consisted of statistically significant observed effect sizes. If $\tau = 0$, there was a small positive bias in *p*-uniform* and Hedges1992 for all levels of pub, and this bias was the largest for $pub = 1$ in combination with $\mu = 0.5$ (bias = 0.067 for *p*-uniform* and 0.042 for Hedges1992). *P*-uniform* and Hedges1992 underestimated τ if $\tau > 0$ for all levels of pub with *p*-uniform* being less negatively biased than Hedges1992.

Increasing *k* resulted in less bias of *p*-uniform* and Hedges1992 but not of the random-effects model. *P*-uniform* benefitted the most from a larger number of effect sizes in a meta-analysis and its bias was comparable or less than of Hedges1992. Hence, these results imply that *p*-uniform* should be used for estimating τ in the potential presence of publication bias, especially when the number of studies is large (i.e., $k \geq 60$).

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 increased in pub if $\tau > 0$. If $\tau = 0$, the

RMSE was very small for $pub = 1$, which was caused by τ being estimated as zero in the vast majority of meta-analyses (see Figures 5 and 6). *P*-uniform* and Hedges1992 had similar RMSEs, except that *p*-uniform*'s RMSE exceeded that of Hedges1992 if $pub = 1$, due to larger variability of *p*-uniform*'s estimates for τ (see https://osf.io/gycf2/).

While the RMSE did not substantially decrease for the random-effects model when increasing *k* to 60, it did decrease for *p*-uniform* and Hedges1992. RMSEs of *p*-uniform* and Hedges1992 were quite similar, although lower for Hedges1992 if $pub = 1$. RMSEs of *p*-uniform* and Hedges1992 were both considerably lower than that of the random-effects model if $pub \geq 0.9$ and $\tau > 0$.

Coverage probability of confidence interval for τ Note that Hedges1992 does not compute confidence intervals for τ. A confidence interval for τ could always be computed with the random-effects model but not always with *p*-uniform*. Non-convergence of *p*-uniform*'s confidence interval was the same as for estimating μ and τ. Table 3 presents the coverage probabilities of the random-effects model and *p*-uniform*. Coverage probabilities of the random-effects model were close to 0.95 for $pub = 0$ but decreased as a function of pub. Undercoverage of the random-effects model was most severe (0.047) for $pub = 1$ in combination with $\mu = 0$ and $\tau = 0$. Coverage probabilities of *p*-uniform* were close to 0.95 if $pub = 0$ and $\tau < 0.346$, but generally decreased as pub and τ were increased. Undercoverage was most severe for $pub = 1$ (minimum coverage 0.42).

Coverage probabilities of the random-effects model decreased if *k* was increased. For *k*=60, the coverage probability of the random-effects model was even equal to zero for $pub = 1$ in combination with $\mu = 0$ and $\mu = 0.2$. Coverage probabilities of *p*-uniform* deviated more from 0.95 if $pub = 1$ than when *k*=10. Hence, researchers are advised against interpreting confidence intervals of *p*-uniform* if meta-analyses only contain statistically significant effect sizes.

General conclusion and recommendations None of the methods outperformed the other methods for all studied conditions. However, some general recommendations can be made. Although the random-effects model had the best statistical properties in the absence of publication bias, we do usually not know the severity of publication bias. Hence, we recommend to always accompany traditional fixed-effect or random-effects meta-analysis with *p*-uniform, *p*-uniform*, or the selection model approach depending on characteristics of the meta-analysis. Generally, we strongly recommend not to use random-effects meta-analysis for testing the null-hypothesis of no effect because of overestimation due to publication bias; Type I error is too high and can even approach 1.

P-uniform* and Hedges1992 outperformed the random-effects model if publication bias was present. However, statistical properties of *p*-uniform* and Hedges1992 were not good in case of extreme publication bias with only statistically significant primary studies' effect sizes in a meta-analysis. As increasing the number of studies to even 120 did not always improve the statistical properties, we recommend not to put much trust in the estimates of any of the methods when a meta-analysis only consists of statistically significant studies. However, *p*-uniform is a viable alternative in case of only statistically significant studies in a meta-analysis due to publication bias and zero or small between-study variance.

Performance of *p*-uniform* and Hedges1992 was comparable which makes it impossible to recommend one method over the other. However, some recommendations can still be made based on the results of our Monte-Carlo simulations. First, we recommend to use *p*-uniform* if a researcher's main emphasis is on estimating the between-study variance, as *p*-uniform* was less biased than Hedges1992 for estimating τ and hardly suffers from convergence problems. However, average effect size estimates of *p*-uniform* can be highly negative, which resulted in a larger RMSE than that of Hedges1992 and sometimes even larger than of the random-effects model. Hence, we recommend to set *p*-uniform*'s estimate of the average effect size to zero if this occurs, which is in line with our recommendation for *p*-uniform and *p*-curve (van Aert et al., 2016). This adjustment is defensible, because it is unlikely that the average effect size estimate is (strongly) negative if statistically significant positive primary studies' effect sizes are observed.

Application

We apply random-effects meta-analysis, *p*-uniform, *p*-uniform*, and Hedges1992 to two published meta-analyses in this section. The meta-analysis by Rabelo, Keller, Pilati, and Wicherts (2015) consisted 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 α=.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 α=.05, 14 effect sizes were significantly larger than zero and 1 effect size was significantly smaller than zero.

The Paule-Mandel estimator (Paule & Mandel, 1982) was used as estimator for the between-study variance in the random-effects meta-analysis model, the Irwin-Hall estimator was used for *p*-uniform, and the maximum likelihood estimator for *p*-uniform*. The selection models of *p*-uniform* and Hedges1992 placed a cut-off at .025 that assumes that two-tailed hypothesis tests were conducted in the primary studies with α=.05 and only effect sizes in the predicted direction were reported. R code of this application 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 four rows) and Bangert-Drowns et al. (2004) (last four 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* (0.075), and Hedges1992 (0.254) when compared with the estimate not corrected for bias (0.571). The random-effects model and Hedges1992 concluded that the effect was significantly different from zero whereas *p*-uniform and *p*-uniform* did not reject the null-hypothesis of no effect. The between-study variance was estimated as zero in the random-effects model, *p*-uniform*, and Hedges1992. As the results of our simulation study show that *p*-uniform generally outperforms the other methods in case the vast majority of primary studies are statistically significant in combination with homogeneous true effect size, we conclude that there was no convincing evidence that weight affects judgments of importance.

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 Hedges1992

(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 size. The null-hypothesis of no effect was rejected by all methods except *p*-uniform. The estimated between-study variance by *p*-uniform* (0.027) and Hedges1992 (0.028) was smaller than estimated in the random-effects meta-analysis (0.069) but still statistically significant. To conclude, correcting for potential publication bias did not affect the conclusions that writing-to-learn interventions affected academic achievement, and that this effect was heterogeneous.

Discussion

Publication bias distorts the results of meta-analyses yielding overestimated effect sizes and false positives. 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 size.

The aim of this paper was to introduce *p*-uniform* and compare the statistical properties of the method with those of *p*-uniform, the selection model approach of Hedges (1992), and the random-effects model that is commonly used but does not correct for publication bias. We studied their relative performance under different levels of publication bias using a Monte-Carlo simulation study where we selected challenging and realistic conditions that are representative for meta-analyses in practice. Statistical properties of the random-effects model were better than of *p*-uniform*, *p*-uniform, and the selection model approach of Hedges (1992) if publication bias did not affect the probability of publishing a primary study. If publication bias was present, the random-effects model performed worse than methods that corrected for publication bias, confirming previous research showing that it overestimates effect size and yields unpredictable bias in estimating between-study variance when publication bias operates (Augusteijn, van Aert, & van Assen, 2019; Jackson, 2006, 2007).

Statistical properties of *p*-uniform* and the selection model approach of Hedges (1992) were generally comparable. Statistical properties of *p*-uniform* and the selection model approach of Hedges (1992) 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 is a viable alternative in such a situation if heterogeneity in true effect size is zero or small.

The comparable statistical properties of *p*-uniform* and the selection model approach were caused by the similarities between the two methods. Both methods use maximum likelihood estimation, the random effects model as effect size model, and a selection model with one threshold at the α -level. However, there are also differences between the two methods explaining why the statistical properties were not exactly the same. First, weights in the selection model have to be estimated or assumed to be known in the selection model approach by Hedges (1992) which is in contrast with *p*-uniform*. *P*-uniform* only assumes that these probabilities are the same for the statistically significant and the same for the nonsignificant primary studies' effect sizes and there is no need for estimating these probabilities. This is an advantage as the statistical model of *p*-uniform* is more parsimonious than the one of the selection model approach by Hedges (1992).

Another difference is that the selection model approach by Hedges (1992) as implemented in the R package "weightr" relies on asymptotic normality distributions for creating Wald-type confidence intervals for the effect size, and that this selection model approach, in contrast to *p*-uniform*, cannot be used for computing a confidence interval for the between-study variance. Confidence intervals computed with *p*-uniform* do not rely on asymptotic normality distributions, because these are computed by inverting the likelihood-ratio tests which may explain why the coverage probabilities of *p*-uniform's confidence interval for the effect size were closer to the nominal coverage rate than of the selection model approach.

We provide recommendations for meta-analysts in practice based on the results of our analytical approximations and 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). To summarize, we recommend to not solely rely on the fixed-effect and random-effects model if publication bias may have affected the meta-analysis, but to supplement these models by either *p*-uniform*, the selection model approach of Hedges (1992), or by both methods. 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.

Importantly, we do not recommend using either *p*-uniform* or the selection model approach of Hedges (1992) if the meta-analysis only contains statistically significant effect sizes caused by publication bias, as our Monte-Carlo simulation study showed bad performance for both methods in that condition. If heterogeneity is 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* and the selection model approach 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. The selection model approach of Hedges (1992) can be applied using the "weightr" package (Coburn & Vevea, 2016) or the web application (https://vevealab.shinyapps.io/WeightFunctionModel/).

A limitation of *p*-uniform* and the selection model approach of Hedges (1992) 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 et al., 2017, December 20 for a discussion about this assumption). To counteract a violation of this assumption, the flexibility of the selection model approach of Hedges (1992) can be used by creating more than two intervals of the method's selection model. *P*-uniform*'s selection model can 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 *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 (Du et al., 2017). We simulated data in this study that exactly matched the selection models of *p*-uniform* and the selection model approach of Hedges (1992), 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, & 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. A limitation that only applies to selection model approaches is that they do not converge if there are no primary study's *p*-values observed in each interval of the selection model, as weights of these intervals then cannot be computed. This non-convergence is circumvented in the R package "weightr" by fixing weights to 0.01 if these cannot be estimated. Although previous research has shown that the weights are often poorly estimated and that the selection model approach is quite robust to misestimated weights (Hedges & Vevea, 1996), future research may scrutinize the effects of fixing weights and whether these weights are better fixed to another value than 0.01. The value 0.01 seems unrealistically small, implying that 99% of the statistically nonsignificant effect sizes end up in the file drawer. This weight may also be estimated by other evidence in the field of the meta-analysis.

As our study did not address the performance of tests of publication bias, 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 Hedges (1992) based on the selection model approach, and a publication bias test in the framework of *p*-uniform* can also be developed. 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 the selection model approach of Hedges (1992) 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 et al., 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. The *p*-uniform* method is an extension and substantial improvement over *p*-uniform and showed promising results in an analytical study and Monte-Carlo simulations.

Open Practices Statement

All R code used in this paper is available via the Open Science Framework page: https://osf.io/ebq6m/.

References

Agresti, A. (2013). *Categorical data analysis*. Hoboken, N.J.: Wiley-Interscience.

Aguinis, H., Gottfredson, R. K., & Wright, T. A. (2011). Best-practice recommendations for estimating interaction effects using meta-analysis. *Journal of Organizational Behavior, 32*(8), 1033-1043. doi:10.1002/job.719

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. (1999). What works?: Selectivity models and meta-analysis. *Journal of the Royal Statistical Society. Series A 162*(1), 95-109.

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

Copas, J. B., & Shi, J. Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research, 10*(4), 251-265.

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

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology, 13*(3), e1002106. doi:10.1371/journal.pbio.1002106

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. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539-1558. doi:10.1002/sim.1186

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.

Hunter, J. E., & Schmidt, F. L. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, California: Sage.

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.

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-Herbrueggen, O. (2019). Publication bias in meta-analyses of posttraumatic stress disorder interventions. Manuscript submitted for publication.

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. (2020). R: A language and environment for statistical computing.

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

Rhodes, K. M., Turner, R. M., & Higgins, J. P. (2015). Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of Clinical Epidemiology*, *68*(1), 52-60. doi:10.1016/j.jclinepi.2014.08.012

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

Rufibach, K. (2015). selectMeta: Estimation of weight functions in meta analysis. Retrieved from https://CRAN.R-project.org/package=selectMeta

Schwarzer, G., Carpenter, J., & Rücker, G. (2016). metasens: Advanced statistical methods to model and adjust for bias in meta-analysis. R package version 0.3-1. https://CRAN.R-project.org/package=metasens.

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 ps>.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

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology, 53*(11), 1119-1129. doi:10.1016/S0895-4356(00)00242-0

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

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22*(13), 2113-2126. doi:10.1002/sim.1461

Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G., & Higgins, J. P. T. (2015). Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine, 34*(6), 984-998. doi:10.1002/sim.6381

van Aert, R. C. M. (2020). puniform: Meta-analysis methods correcting for publication bias. (Version 0.2.2). 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:https://doi.org/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., Augusteijn, 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] 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/ms5kn/). The same variables were varied in this simulation study as in the one described above except that *k* was now fixed to 30. 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.

[4] 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, Rücker, & Schwarzer, 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.

Table 1

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 Hedges1992 as a function of μ, τ, the severity of publication bias (pub), and the number of primary studies' observed effect sizes (k).*

		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
τ=0	RE model	0.949	0.95	0.871	0	0.951	0.912	0.357	0	0.949	0.915	0.689	0.55
	p-uniform	0.947	0.951	0.953	0.951	0.947	0.948	0.952	0.953	0.955	0.948	0.952	0.948
	p-uniform*	0.952	0.95	0.958	0.881	0.943	0.948	0.955	0.899	0.955	0.94	0.905	0.883
	Hedges1992	0.955	0.96	0.966	0.589	0.955	0.952	0.958	0.795	0.951	0.903	0.684	0.57
τ=0.163	RE model	0.952	0.943	0.656	0	0.952	0.898	0.29	0	0.954	0.9	0.58	0.371
	p-uniform	0.894	0.896	0.861	0.786	0.897	0.892	0.849	0.812	0.901	0.902	0.889	0.884
	p-uniform*	0.918	0.922	0.928	0.684	0.913	0.927	0.896	0.756	0.931	0.917	0.867	0.836
	Hedges1992	0.945	0.943	0.945	0.234	0.927	0.931	0.912	0.454	0.935	0.902	0.691	0.527
τ=0.346	RE model	0.951	0.913	0.347	0	0.954	0.872	0.253	0	0.951	0.878	0.454	0.154
	p-uniform	0.653	0.607	0.364	0.199	0.637	0.55	0.361	0.284	0.669	0.616	0.55	0.537
	p-uniform*	0.886	0.906	0.86	0.341	0.885	0.895	0.773	0.45	0.894	0.864	0.748	0.677
	Hedges1992	0.912	0.922	0.889	0.108	0.895	0.901	0.797	0.297	0.907	0.876	0.619	0.401

Table 1 Continued

		k=60											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	pub	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
τ=0	RE model	0.951	0.934	0.281	0	0.951	0.697	0	0	0.95	0.676	0.016	0
	p-uniform	0.949	0.945	0.955	0.953	0.952	0.954	0.948	0.95	0.948	0.947	0.942	0.945
	p-uniform*	0.952	0.955	0.948	0.803	0.944	0.951	0.951	0.887	0.945	0.932	0.875	0.829
	Hedges1992	0.967	0.97	0.954	0.284	0.982	0.971	0.963	0.702	0.958	0.978	0.967	0.604
τ=0.163	RE model	0.951	0.867	0.01	0	0.95	0.599	0	0	0.95	0.612	0.004	0
	p-uniform	0.871	0.826	0.576	0.231	0.779	0.67	0.456	0.339	0.754	0.714	0.682	0.666
	p-uniform*	0.923	0.932	0.941	0.147	0.91	0.938	0.902	0.347	0.933	0.913	0.811	0.718
	Hedges1992	0.951	0.952	0.943	0.219	0.936	0.941	0.936	0.388	0.948	0.958	0.944	0.51
τ=0.346	RE model	0.952	0.683	0	0	0.948	0.495	0	0	0.95	0.538	0.001	0
	p-uniform	0.237	0.074	0	0	0.092	0.019	0	0	0.09	0.043	0.016	0.014
	p-uniform*	0.91	0.932	0.922	0.245	0.922	0.94	0.86	0.378	0.934	0.916	0.774	0.613
	Hedges1992	0.94	0.944	0.935	0.137	0.934	0.942	0.929	0.259	0.944	0.951	0.922	0.42

Table 2

Type I error rate and statistical power for testing the null hypothesis of no effect with the random-effects model (RE), p-uniform, p-uniform using maximum likelihood estimation (ML), and Hedges1992 as a function of μ, τ, the severity of publication bias (pub), and the number of primary studies' observed effect sizes (k).*

		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
τ=0	RE model	0.049	0.062	0.222	1	0.89	0.945	0.998	1	1	1	1	1
	p-uniform	0.029	0.031	0.042	0.051	0.125	0.151	0.217	0.283	0.848	0.889	0.936	0.94
	p-uniform*	0.039	0.041	0.038	0.144	0.694	0.632	0.312	0.122	0.841	0.626	0.236	0.093
	Hedges1992	0.028	0.024	0.032	0.492	0.525	0.621	0.474	0.793	0.952	0.912	0.935	0.988
τ=0.163	RE model	0.049	0.082	0.487	1	0.732	0.866	0.997	1	1	1	1	1
	p-uniform	0.103	0.12	0.203	0.311	0.292	0.358	0.534	0.65	0.891	0.932	0.967	0.973
	p-uniform*	0.056	0.054	0.057	0.143	0.461	0.399	0.212	0.141	0.766	0.583	0.246	0.132
	Hedges1992	0.017	0.019	0.056	0.826	0.399	0.458	0.424	0.941	0.929	0.888	0.903	0.978
τ=0.346	RE model	0.05	0.145	0.764	1	0.413	0.689	0.987	1	0.976	0.997	1	1
	p-uniform	0.399	0.477	0.732	0.872	0.615	0.739	0.907	0.948	0.934	0.971	0.99	0.995
	p-uniform*	0.064	0.062	0.073	0.178	0.19	0.184	0.136	0.17	0.535	0.399	0.221	0.167
	Hedges1992	0.019	0.036	0.13	0.913	0.203	0.28	0.407	0.906	0.739	0.715	0.78	0.921

Table 2 Continued

		k=60											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	pub	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
τ=0	RE model	0.049	0.106	0.821	1	1	1	1	1	1	1	1	1
	p-uniform	0.037	0.049	0.045	0.039	0.284	0.393	0.662	0.805	1	1	1	1
	p-uniform*	0.044	0.042	0.049	0.112	1	1	0.984	0.252	1	1	0.927	0.657
	Hedges1992	0.033	0.035	0.049	0.801	0.996	0.999	0.993	0.903	1	0.999	0.988	0.991
τ=0.163	RE model	0.051	0.21	0.996	1	1	1	1	1	1	1	1	1
	p-uniform	0.188	0.256	0.542	0.85	0.734	0.89	0.992	0.999	1	1	1	1
	p-uniform*	0.056	0.055	0.056	0.274	0.994	0.988	0.826	0.304	1	0.999	0.871	0.513
	Hedges1992	0.028	0.036	0.06	0.824	0.997	0.994	0.916	0.899	1	1	0.981	0.996
τ=0.346	RE model	0.048	0.436	1	1	0.985	1	1	1	1	1	1	1
	p-uniform	0.84	0.955	1	1	0.994	1	1	1	1	1	1	1
	p-uniform*	0.05	0.051	0.062	0.179	0.72	0.675	0.372	0.168	0.999	0.987	0.673	0.305
	Hedges1992	0.02	0.034	0.073	0.891	0.778	0.773	0.566	0.929	1	0.997	0.894	0.981

Table 3

Coverage probability of the confidence interval for τ computed with the random-effects model (RE) and p-uniform using maximum likelihood estimation (ML) as a function of μ, τ, the severity of publication bias (pub), and the number of primary studies' observed effect sizes (k).*

		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
τ=0	RE model	0.948	0.936	0.777	0.047	0.949	0.95	0.898	0.204	0.952	0.934	0.824	0.767
	p-uniform*	0.972	0.97	0.98	0.966	0.961	0.978	0.986	0.928	0.983	0.966	0.908	0.883
τ=0.163	RE model	0.949	0.939	0.82	0.092	0.952	0.951	0.861	0.262	0.952	0.926	0.805	0.726
	p-uniform*	0.954	0.968	0.978	0.94	0.967	0.978	0.973	0.912	0.981	0.972	0.915	0.877
τ=0.346	RE model	0.946	0.943	0.898	0.186	0.949	0.949	0.812	0.351	0.951	0.929	0.782	0.666
	p-uniform*	0.891	0.911	0.884	0.42	0.902	0.911	0.814	0.532	0.907	0.881	0.765	0.699

Table 3 Continued

		k=60											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	pub	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
τ=0	RE model	0.954	0.909	0.156	0	0.949	0.886	0.895	0	0.953	0.85	0.247	0.05
	p-uniform*	0.98	0.977	0.981	0.974	0.968	0.983	0.989	0.931	0.988	0.972	0.923	0.871
τ=0.163	RE model	0.953	0.876	0.226	0	0.948	0.912	0.832	0	0.952	0.855	0.218	0.026
	p-uniform*	0.921	0.928	0.944	0.126	0.91	0.939	0.9	0.288	0.927	0.904	0.776	0.655
τ=0.346	RE model	0.948	0.882	0.892	0	0.951	0.943	0.552	0	0.95	0.878	0.202	0.008
	p-uniform*	0.91	0.928	0.922	0.267	0.922	0.934	0.855	0.389	0.934	0.912	0.769	0.607

Table 4

Results of applying the random-effects meta-analysis model, p-uniform, p-uniform, and the selection model approach of Hedges (1992) to the meta-analysis of Rabelo et al. (2015) (first four rows) and Bangert-Drowns et al (2004) (last four rows). Between-study variance in the random-effects meta-analysis model was estimated using the Paule-Mandel estimator. P-uniform neither does estimate τ^2 nor provides inferences for this parameter. Standard errors cannot be estimated with p-uniform and p-uniform*. The selection model approach of Hedges (1992) does not estimate a standard error and confidence interval for τ^2.*

	μ (SE)	(95% CI μ)	H$_0$: μ = 0	τ^2 (SE)	(95% CI τ^2)	H$_0$: τ^2 = 0
Rabelo et al. (2015)						
RE model	0.571 (0.023)	(0.524; 0.618)	t=25.036, p<.0001	0 (0.02)	(0; 0)	Q=4.553, p=1
p-uniform	-0.179 (-)	(-0.676; 0.159)	L_0=0.959, p=.831	-	-	-
p-uniform* (ML)	0.075 (-)	(-0.188; 0.307)	L_0=0.339, p=.56	0 (-)	(0; 0.022)	L_{het}=0, p=1
Hedges1992	0.254 (0.018)	(0.22; 0.289)	z=14.402, p<.0001	0 (-)	(-)	Q=4.553, p=1
Bangert-Drowns et al. (2004)						
RE model	0.228 (0.051)	(0.127; 0.330)	t=4.511, p<.0001	0.069 (0.025)	(0.027; 0.153)	Q=107.106, p<.0001
p-uniform	0.245 (-)	(-0.236; 0.531)	L_0=-1.14, p=.127	-	-	-
p-uniform* (ML)	0.179 (-)	(0.068; 0.297)	L_0=10.224, p=.001	0.027 (-)	(0.006; 0.073)	L_{het}=6.311, p=.012
Hedges1992	0.148 (0.073)	(0.004; 0.291)	z=2.02, p=.043	0.028 (0.024)	(-)	Q=107.106, p<.0001

Note. RE refers to random-effects, ML to maximum likelihood, SE to standard error, and CI to confidence interval.

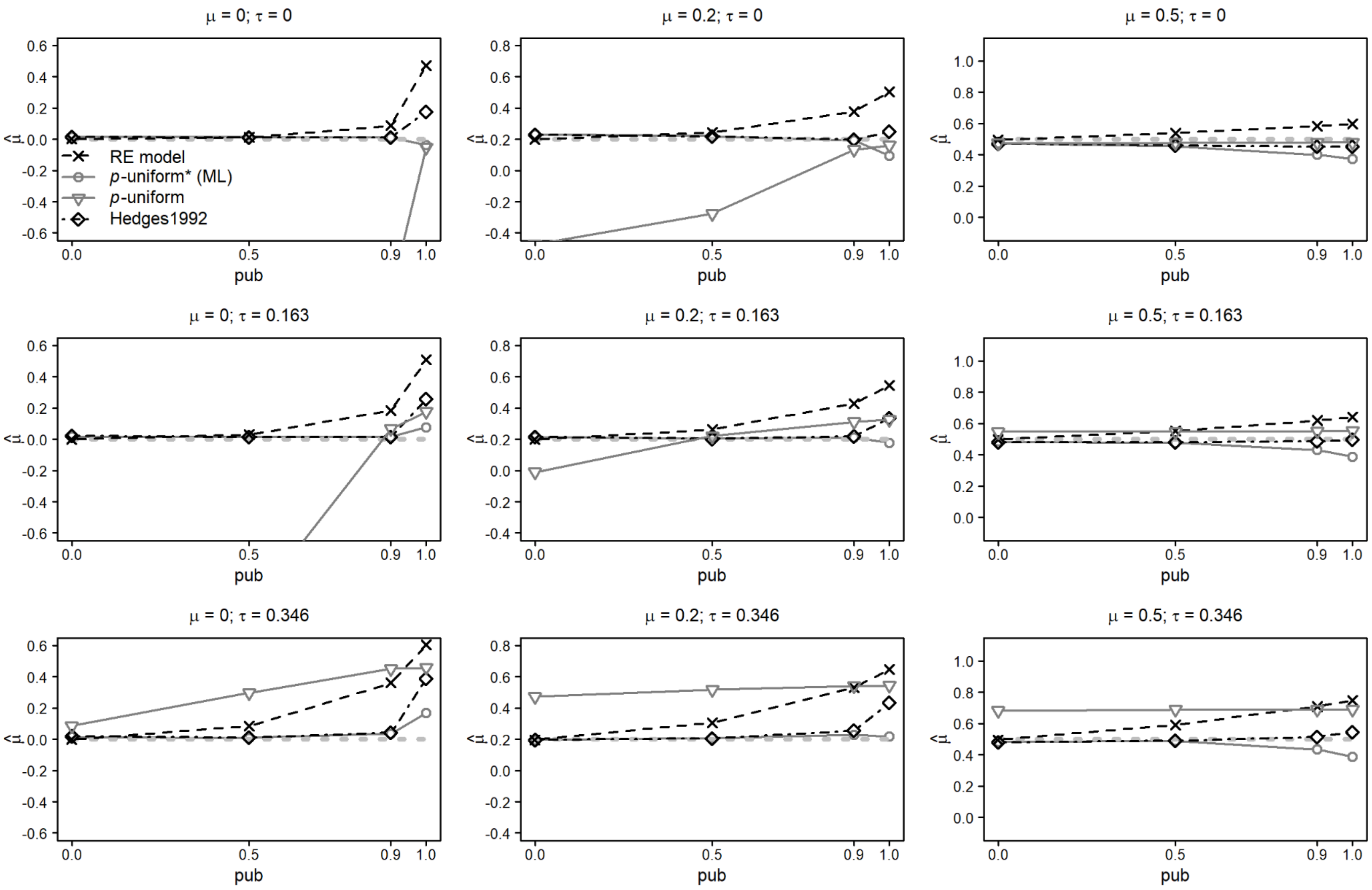
μ = 0; τ = 0
μ = 0.2; τ = 0
μ = 0.5; τ = 0
μ = 0; τ = 0.163
μ = 0.2; τ = 0.163
μ = 0.5; τ = 0.163
μ = 0; τ = 0.346
μ = 0.2; τ = 0.346
μ = 0.5; τ = 0.346
RE model
p-uniform* (ML)
p-uniform
Hedges1992
μ̂
pub
0.0
0.5
0.9
1.0
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0

Figure 1. Average of the estimates of μ for the random-effects model (RE), *p*-uniform, *p*-uniform* using maximum likelihood estimation (ML), and Hedges1992 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.

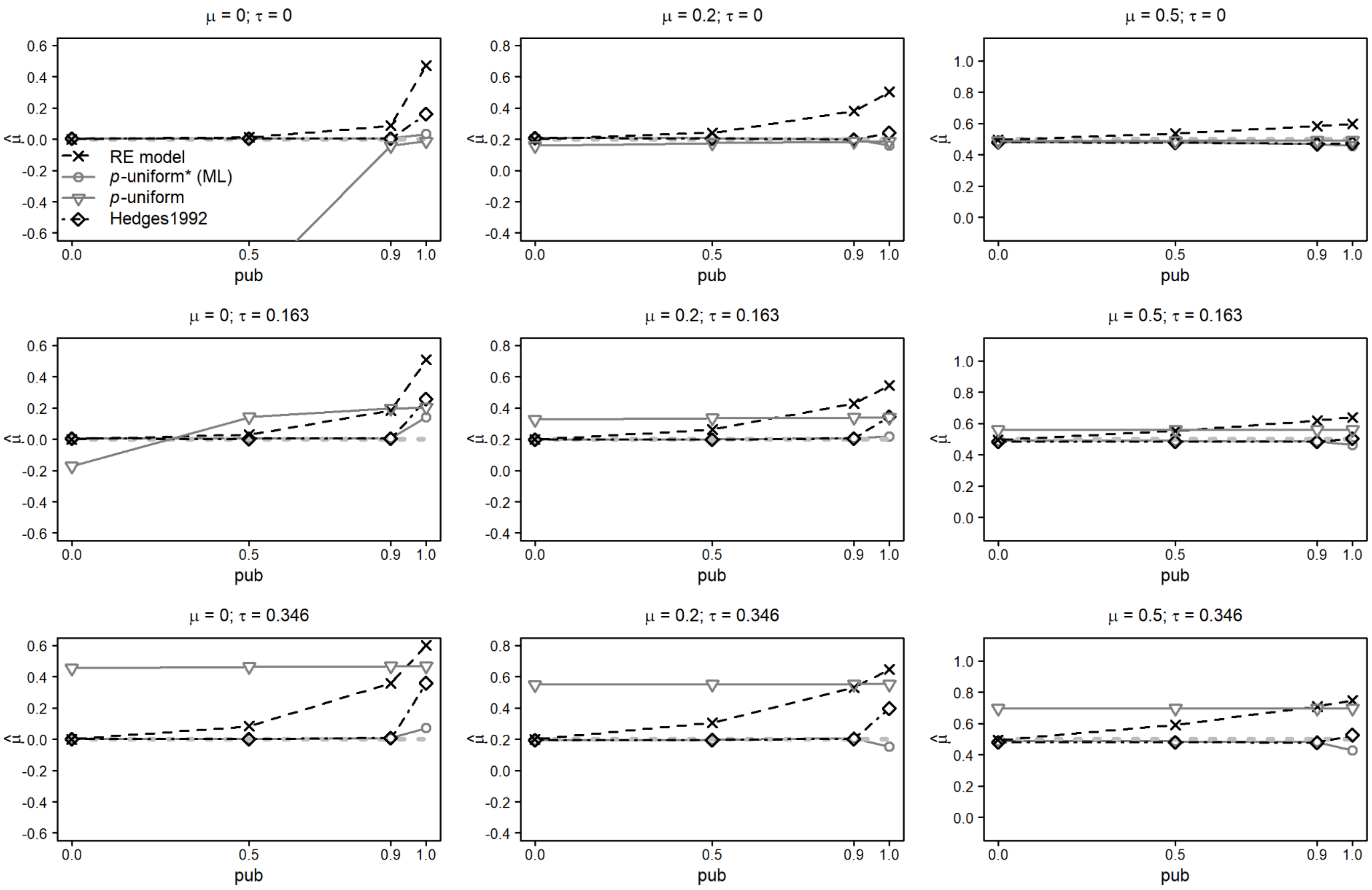
μ = 0; τ = 0
μ = 0.2; τ = 0
μ = 0.5; τ = 0
μ = 0; τ = 0.163
μ = 0.2; τ = 0.163
μ = 0.5; τ = 0.163
μ = 0; τ = 0.346
μ = 0.2; τ = 0.346
μ = 0.5; τ = 0.346
RE model
p-uniform* (ML)
p-uniform
Hedges1992
μ̂
pub
0.0
0.5
0.9
1.0
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8

Figure 2. Average of the estimates of μ for the random-effects model (RE), *p*-uniform, *p*-uniform* using maximum likelihood estimation (ML), and Hedges1992 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.

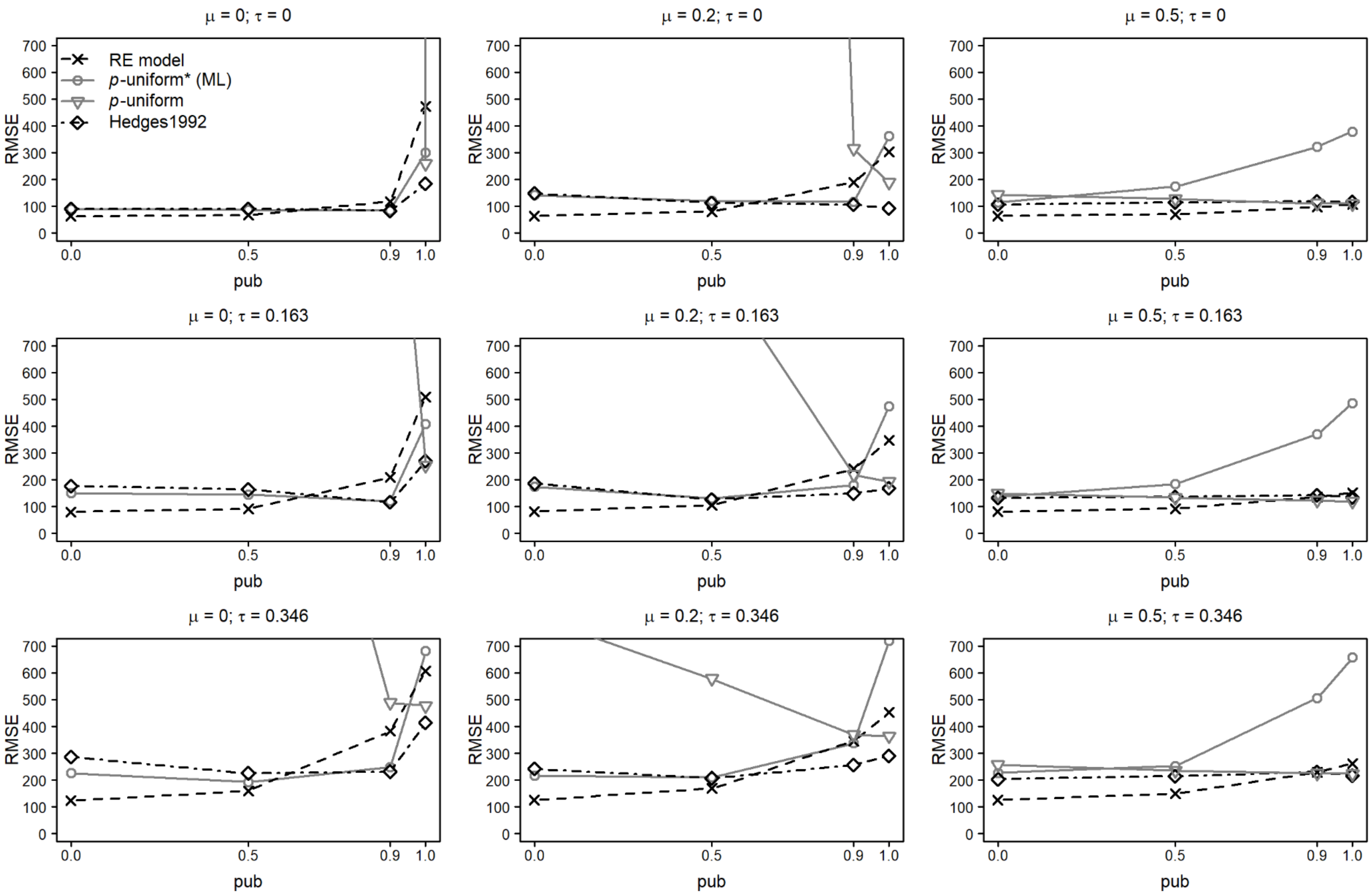
μ = 0; τ = 0
μ = 0.2; τ = 0
μ = 0.5; τ = 0
μ = 0; τ = 0.163
μ = 0.2; τ = 0.163
μ = 0.5; τ = 0.163
μ = 0; τ = 0.346
μ = 0.2; τ = 0.346
μ = 0.5; τ = 0.346
RE model
p-uniform* (ML)
p-uniform
Hedges1992
RMSE
0
100
200
300
400
500
600
700
pub
0.0
0.5
0.9
1.0

Figure 3. Root mean square error (RMSE) of estimating μ with the random-effects model (RE), *p*-uniform, *p*-uniform* using maximum likelihood estimation (ML), and Hedges1992 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.

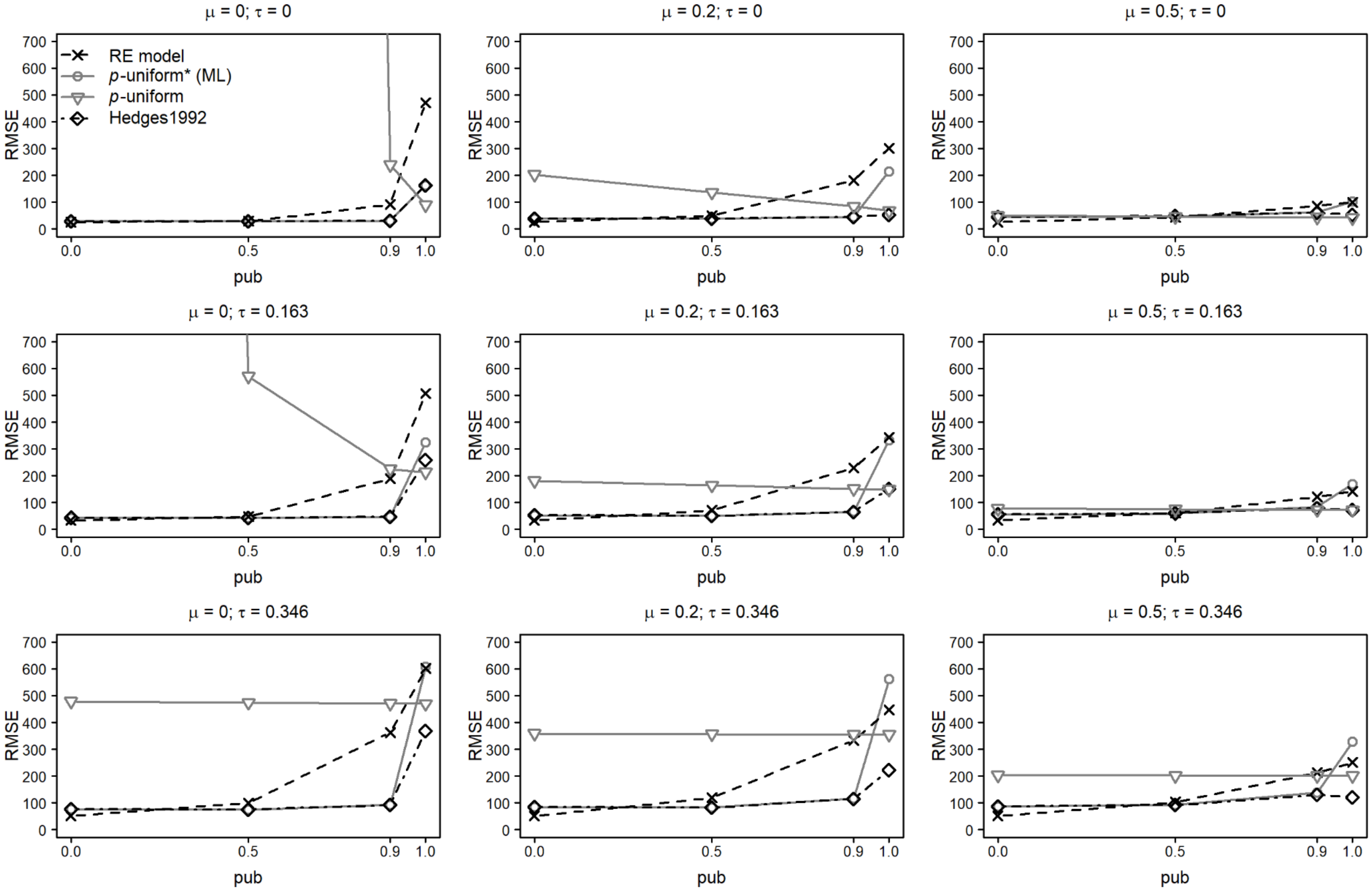
μ = 0; τ = 0
μ = 0.2; τ = 0
μ = 0.5; τ = 0
μ = 0; τ = 0.163
μ = 0.2; τ = 0.163
μ = 0.5; τ = 0.163
μ = 0; τ = 0.346
μ = 0.2; τ = 0.346
μ = 0.5; τ = 0.346
RE model
p-uniform* (ML)
p-uniform
Hedges1992
RMSE
pub
0
100
200
300
400
500
600
700
0.0
0.5
0.9
1.0

Figure 4. Root mean square error (RMSE) of estimating μ with the random-effects model (RE), *p*-uniform, *p*-uniform* using maximum likelihood estimation (ML), and Hedges1992 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.

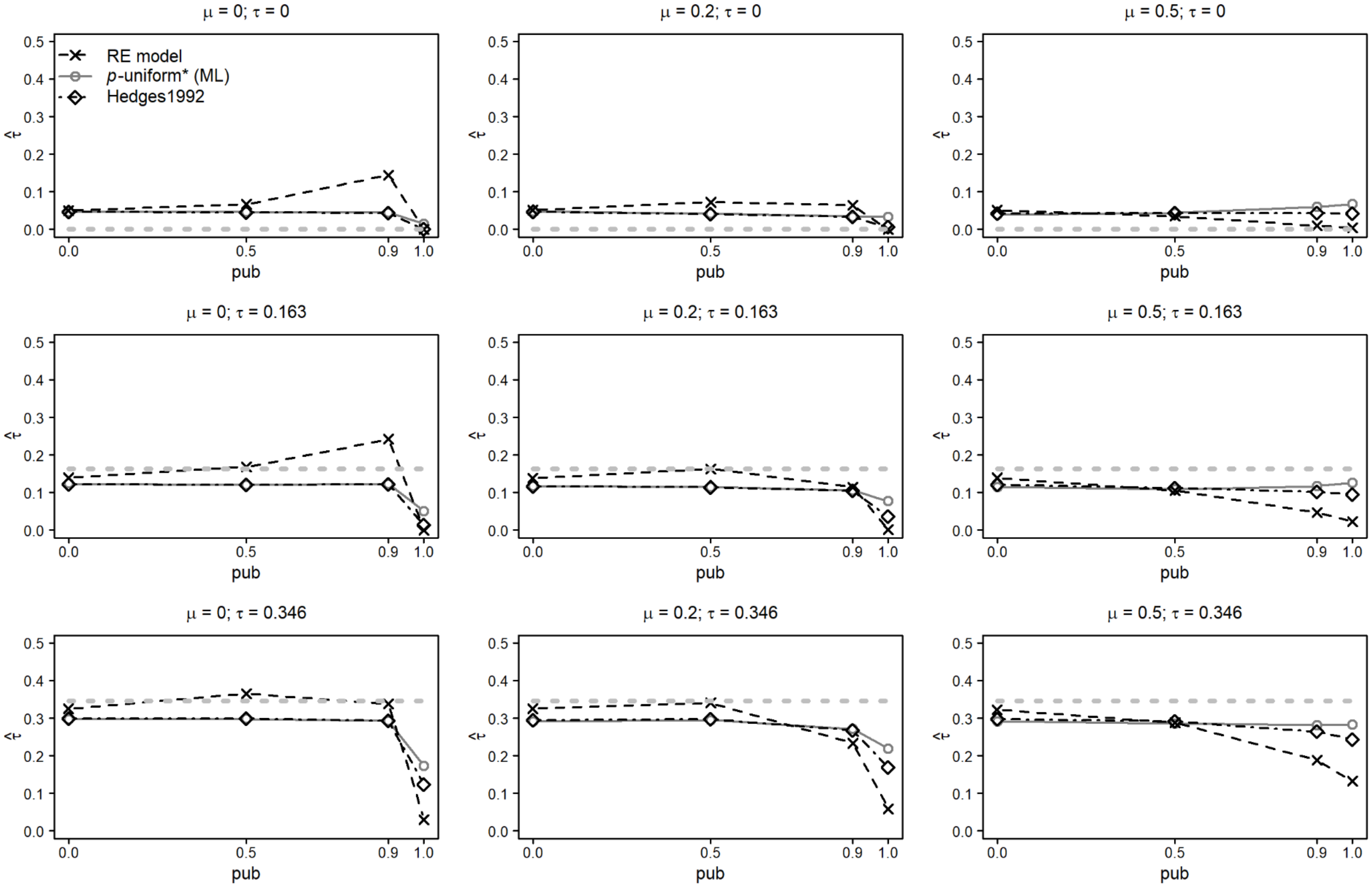
μ = 0; τ = 0
μ = 0.2; τ = 0
μ = 0.5; τ = 0
μ = 0; τ = 0.163
μ = 0.2; τ = 0.163
μ = 0.5; τ = 0.163
μ = 0; τ = 0.346
μ = 0.2; τ = 0.346
μ = 0.5; τ = 0.346
RE model
p-uniform* (ML)
Hedges1992
$\hat{\tau}$
0.0
0.1
0.2
0.3
0.4
0.5
pub
0.0
0.5
0.9
1.0

Figure 5. Average of the estimates of τ for the random-effects model (RE), *p*-uniform* using maximum likelihood estimation (ML), and Hedges1992 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.

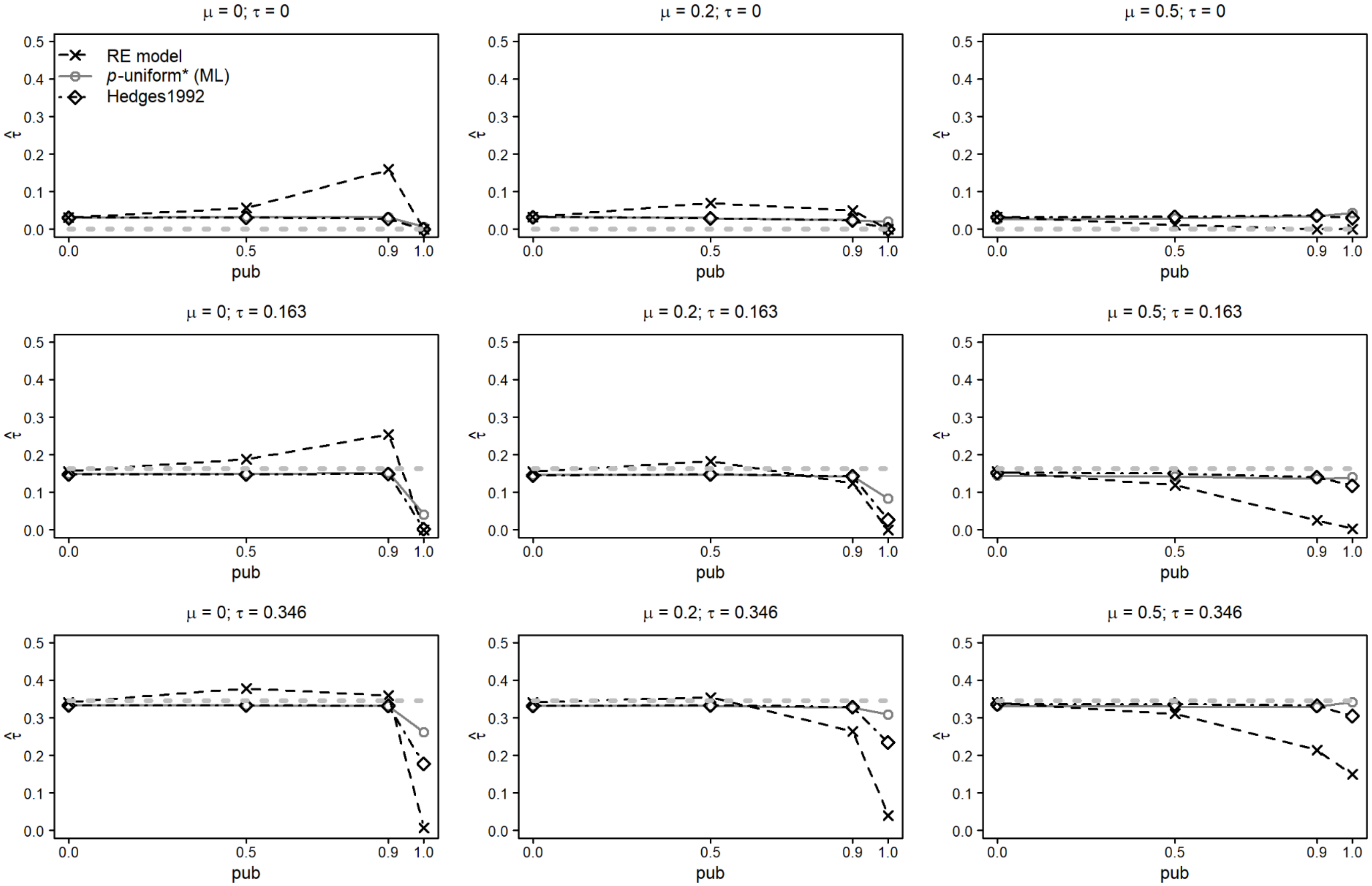
μ = 0; τ = 0
μ = 0.2; τ = 0
μ = 0.5; τ = 0
μ = 0; τ = 0.163
μ = 0.2; τ = 0.163
μ = 0.5; τ = 0.163
μ = 0; τ = 0.346
μ = 0.2; τ = 0.346
μ = 0.5; τ = 0.346
RE model
p-uniform* (ML)
Hedges1992
τ̂
0.0
0.1
0.2
0.3
0.4
0.5
pub
0.0
0.5
0.9
1.0

Figure 6. Average of the estimates of τ for the random-effects model (RE), *p*-uniform* using maximum likelihood estimation (ML), and Hedges1992 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.

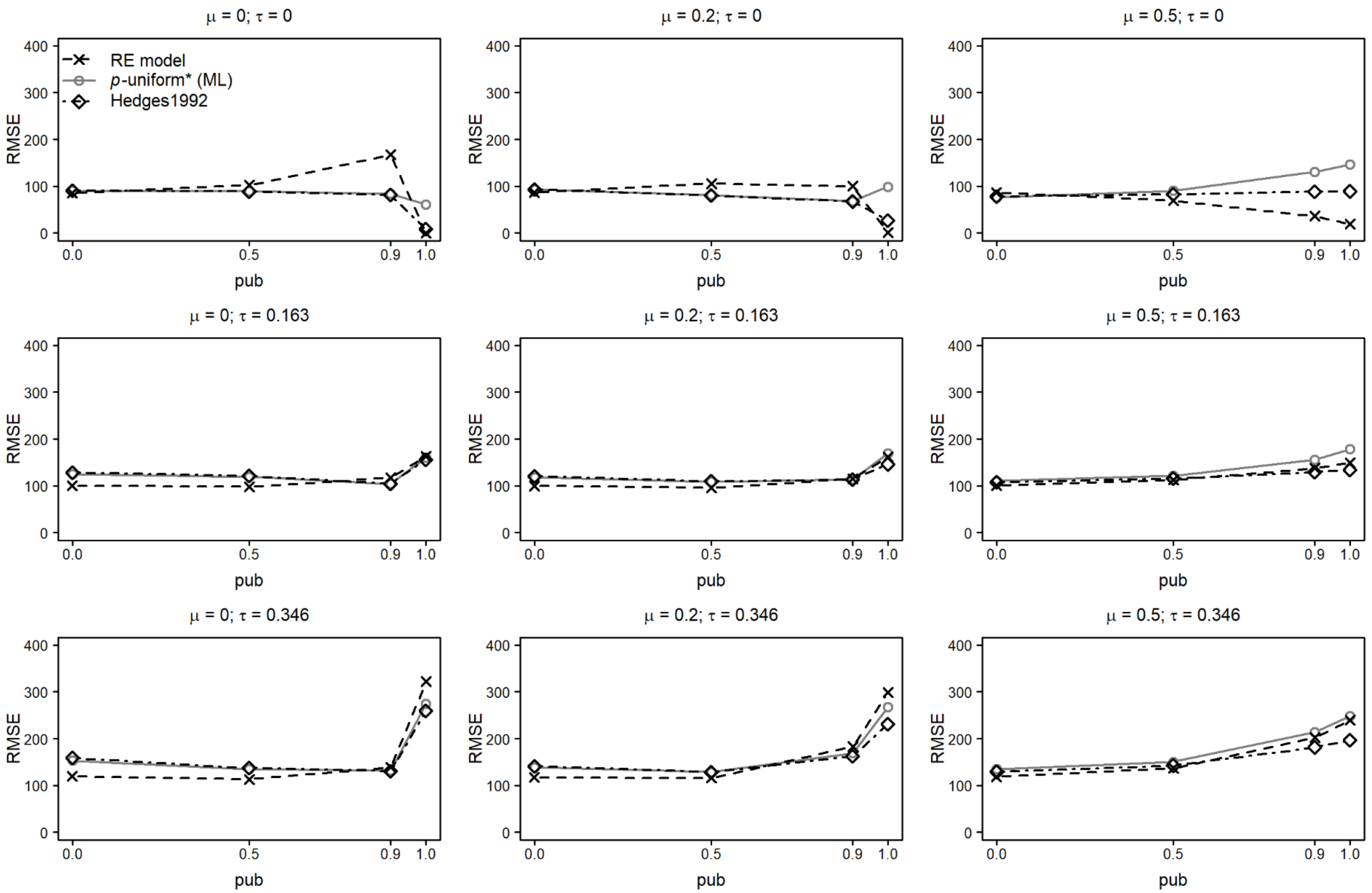
μ = 0; τ = 0
μ = 0.2; τ = 0
μ = 0.5; τ = 0
μ = 0; τ = 0.163
μ = 0.2; τ = 0.163
μ = 0.5; τ = 0.163
μ = 0; τ = 0.346
μ = 0.2; τ = 0.346
μ = 0.5; τ = 0.346
RE model
p-uniform* (ML)
Hedges1992
RMSE
0
100
200
300
400
pub
0.0
0.5
0.9
1.0

Figure 7. Root mean square error (RMSE) of estimating τ with the random-effects model (RE), *p*-uniform* using maximum likelihood estimation (ML), and Hedges1992 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.

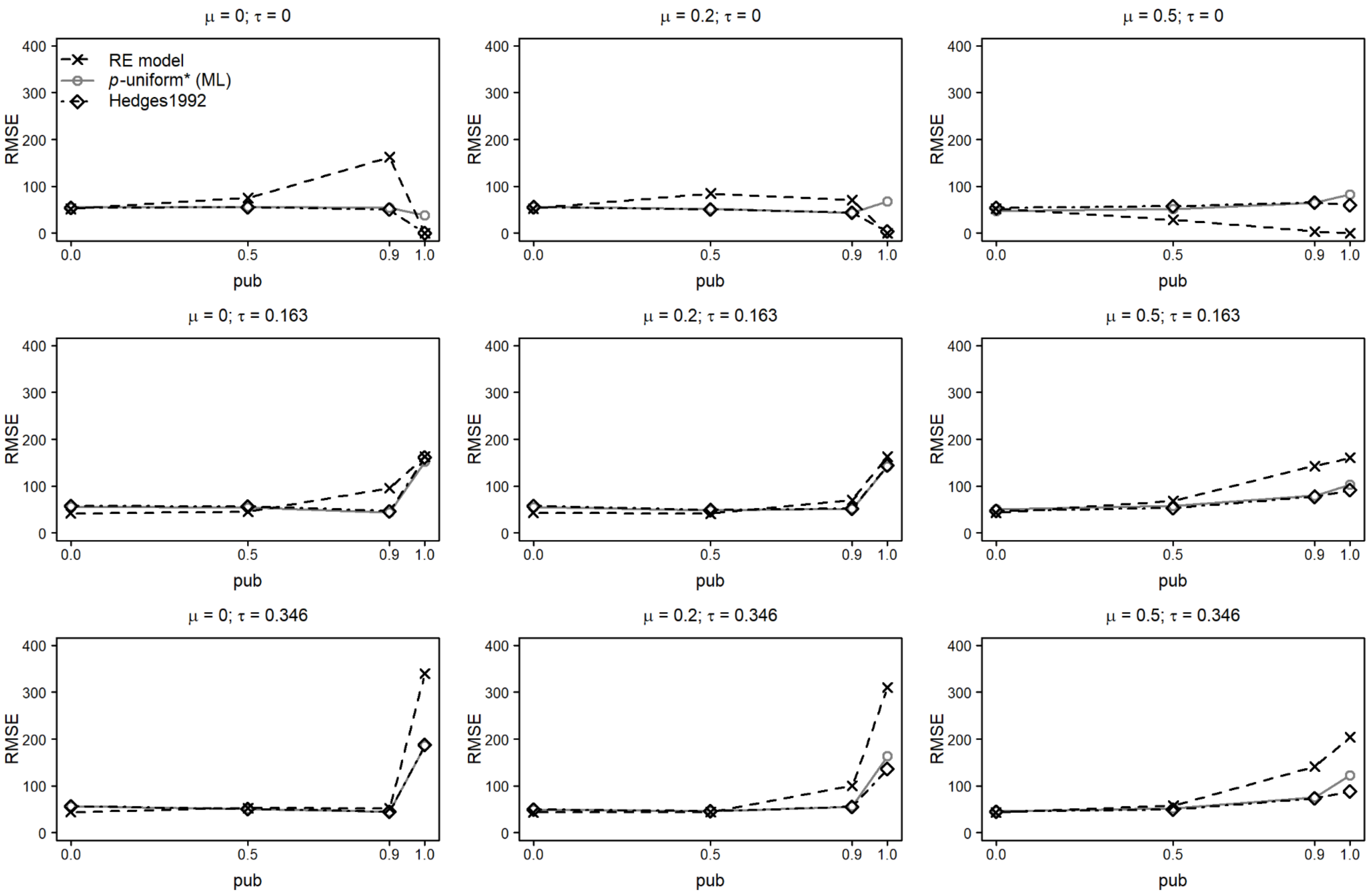
μ = 0; τ = 0
μ = 0.2; τ = 0
μ = 0.5; τ = 0
μ = 0; τ = 0.163
μ = 0.2; τ = 0.163
μ = 0.5; τ = 0.163
μ = 0; τ = 0.346
μ = 0.2; τ = 0.346
μ = 0.5; τ = 0.346
RE model
p-uniform* (ML)
Hedges1992
RMSE
0
100
200
300
400
pub
0.0
0.5
0.9
1.0

Figure 8. Root mean square error (RMSE) of estimating τ with the random-effects model (RE), *p*-uniform* using maximum likelihood estimation (ML), and Hedges1992 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.