Correcting for bias in psychology: A comparison of meta-analytic methods

Evan C. Carter*
University of Minnesota, Minneapolis, MN, USA

Felix D. Schönbrodt*
Ludwig-Maximilians-University, Munich, Germany

Joseph Hilgard
University of Pennsylvania, Philadelphia, PA, USA

Will M. Gervais
University of Kentucky, Lexington, KY, USA

Publication bias and questionable research practices in primary research can lead to badly overestimated effects in meta-analysis. Methodologists have proposed a variety of statistical approaches to correcting for such overestimation. However, much of this work has not been tailored specifically to psychology, so it is not clear which methods work best for data typically seen in our field. Here, we present a comprehensive simulation study to examine how some of the most promising meta-analytic methods perform on data typical of psychological research. We tried to mimic realistic scenarios by simulating several levels of questionable research practices, publication bias, and heterogeneity, using study sample sizes empirically derived from the literature. Our results indicate that one method–the three-parameter selection model (Iyengar & Greenhouse, 1988; McShane, Böckenholt, & Hansen, 2016)–generally performs better than trim-and-fill, *p*-curve, *p*-uniform, PET, PEESE, or PET-PEESE, and that some of these other methods should typically not be used at all. However, it is unknown whether the success of the three-parameter selection model is due to the match between its assumptions and our modeling strategy, so future work is needed to further test its robustness. Despite this, we generally recommend that meta-analysts of data in psychology use the three-parameter selection model. Moreover, we strongly recommend that researchers in psychology continue their efforts on improving the primary literature and conducting large-scale, pre-registered replications.

Manuscript submitted for publication, version 0.1, 2017-05-26. If you cite this preprint, please check on https://osf.io/rf3ys whether the final version of the paper has been published.

Keywords: meta-analysis, publication bias, *p*-hacking, questionable research practices, bias-correction.

Statistical tools for analyzing the results from a set of studies in aggregate–often called meta-analysis–are popular in psychology and many other scientific disciplines. As compared to results from individual studies, meta-analytic results have higher statistical power and yield more precise estimates. Meta-analyses are frequently used to provide high-powered tests of hypotheses, to test for moderators of the effect size across studies, and to suggest effect size estimates for power analysis.

However, just as the results from individual studies can be marred by bias, meta-analytic results can be made far less useful, or even completely misleading, when influenced by various forms of bias. To address this, researchers have developed statistical techniques designed to identify and correct for bias, and several simulation studies have compared the performance of some of these tools (e.g., Moreno et al., 2009; Rücker, Carpenter, & Schwarzer, 2011; T. Stanley & Doucouliagos, 2014; Simonsohn, Nelson, & Simmons, 2014; McShane et al., 2016). Importantly, nearly all such efforts come from outside of psychology, and these simulations have examined effect sizes, sample sizes, and numbers of studies atypical of psychological research. As a result, psychological scientists are faced with an ever-growing menu of meta-analytic tools but little information about which tools are likely to work for the data they commonly encounter.

Here, we present a systematic simulation study to compare the performance of statistical techniques intended to correct for the influence of bias on meta-analytic estimates. We assess the performance of these techniques on data that are designed to be typical of psychological science.

Meta-analysis

Meta-analytic techniques involve synthesizing a set of studies investigating the same empirical phenomenon. For

example, meta-analysis is often used to produce a single summary estimate of the hypothetical true underlying effect, δ, that each study in the dataset purportedly measured. This is usually called fixed-effect meta-analysis (Cooper, Hedges, & Valentine, 2009), and can be modeled as $T_i = \delta + e_i$, where T_i is the observed treatment effect for study i that differs from the true underlying effect, δ, by some amount of sampling error e_i, which is normally distributed with a mean of 0 and a variance of v_i.

A more complex–and very likely more realistic–model known as random-effects meta-analysis (Cooper et al., 2009) holds that each study measures a different, related true effect, δ_i. This approach allows for the possibility that researchers attempting to study the same phenomenon may nonetheless be studying different underlying effects that vary as a function of, for example, different operationalizations of the independent variable or different populations. In this model, $\delta_i = \mu + u_i$, where μ is the mean of the true effect sizes that are estimated by the individual studies and the ith study's deviation from this mean, u_i, is normally distributed with a mean of 0 and a variance of τ^2. Applying the random-effects model to an observed set of studies provides an estimate of the average true underlying effect, μ, and the amount of between-study heterogeneity, τ^2.

Because meta-analyses are usually applied to studies with dependent variables measured on different scales, effect size estimates are typically standardized. In our simulation, we use Cohen's d with its associated variance v_d (Borenstein, Hedges, Higgins, & Rothstein, 2011).

Bias

We use the term bias to refer to the systematic over- or underestimation of meta-analytic estimates. Meta-analytic bias is caused by factors that affect the analysis and reporting of the individual studies that go into a meta-analysis. We consider two primary sources of meta-analytic bias in our simulation study: *publication bias* and *questionable research practices*.

Publication bias is said to occur when the probability of results entering the published record is affected by the results themselves (Rothstein, Sutton, & Borenstein, 2006). For example, if researchers strongly believe that an effect is real and positive, statistically non-significant or negative estimates of that effect may never be submitted for publication or may be rejected by reviewers and editors (Greenwald, 1975; Sterling, Rosenbaum, & Weinkam, 1995; Rothstein et al., 2006; Ferguson & Heene, 2012). In other words, statistically nonsignificant results, or those results that counter accepted theory, are left in the "file-drawer." Since the data set collected by the meta-analyst depends on the availability of studies on the topic of interest, and published data are much easier to find, publication bias can result in a meta-analytic sample that over-represents statistically significant, theory-consistent studies. This can lead to misleading meta-analytic results.

Another form of bias is the undisclosed use of questionable research practices (QRPs; also called "researcher degrees of freedom" or "p-hacking") whereby researchers choose from a variety of potential analyses based on the results they yield. These analytic choices may be justifiable, yet simultaneously arbitrary and motivated (Simonsohn, Simmons, & Nelson, 2015). For example, researchers may encounter a flexible design that can be analyzed in several ways–with or without covariates, outliers, or certain outcomes–and choose an approach that yields the desired statistically-significant result. Such behavior is likely to overestimate the true effect size, as analyses that yield significant results are highlighted and analyses that do not yield such results are censored from report.

Our approach

Given the pernicious influence of publication bias and the undisclosed use of QRPs, meta-analytic methods that are robust to these sources of bias are sorely needed. As mentioned above, much of the work comparing such methods has not focused on conditions representative of psychological science. For example, many systematic studies of meta-analytic techniques are focused on effect size measures (e.g., odds ratios) that are relatively infrequently used in psychology (see Table S1 in Fanelli, Costas, & Ioannidis, 2017, which shows that, of 430 meta-analyses in psychology, 207 use a standardized mean difference whereas only 59 use an odds ratio). Differences between the parameters and effect size measures used in previous simulation studies and those in psychology may limit the applicability of those simulations' results. Thus, we focus on meta-analyses of simulated studies for which the data can be described in terms of the standardized mean difference effect size, Cohen's d.

Along similar lines, previous simulation studies have also examined a range of values for heterogeneity that may or

Correspondence concerning this article should be addressed to Evan Carter, Email: evan.c.carter@gmail.com. *These authors contributed equally to this work.

R scripts are available at Github (https://github.com/nicebread/meta-showdown). Furthermore, we provide interactive figures and tables which allow a detailed exploration of the results (http://www.shinyapps.org/apps/metaExplorer/). Supplemental material, which includes many of the technical details of our simulation, is available at OSF (https://osf.io/rf3ys). We declare that we have no conflicts of interest with respect to the authorship or the publication of this article. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. We would like to acknowledge Tyler Yost for helpful comments on an earlier version of the simulation code and manuscript.

Table 1
Simulation parameters

Experimental factors	Levels
True underlying effect (δ)	0, 0.2, 0.5, 0.8
Between-study heterogeneity (τ)	0, 0.2, 0.4
Number of studies in the meta-analytic sample (k)	10, 30, 60, 100
Percentage of studies produced under publication bias (PB)	0%, 60%, 90%
QRP environment (QRP)	None, medium, high

may not represent what is common in psychology. Therefore, we also selected values for τ (0, 0.2, 0.4) that were designed to be typical of meta-analysis in psychology: In an analysis of 187 between-study heterogeneity estimates from meta-analyses using Cohen's d or Hedges' g published in *Psychological Bulletin* from 1990–2013 (van Erp, Verhagen, Grasman, & Wagenmakers, 2017), 50% of all estimated τs were smaller than 0.2, and 80% were smaller than 0.4.

Furthermore, previous simulations have tended to use relatively large or uniform study-level sample sizes (e.g., McShane et al., 2016). Within psychology, sample sizes are chronically small and can vary widely within a given literature (Fraley & Vazire, 2014). Thus, results from previous simulation studies may not reflect the typical situations for researchers in psychology. To address this, our simulation determines study-level samples sizes by drawing directly on data from an empirical investigation of sample sizes used in psychological research (see below).

Previous simulations usually focused on a limited set of bias-correcting methods. We provide a comprehensive comparison of currently promising approaches by including trim-and-fill, meta-regression techniques (PET, PEESE, PET-PEESE), and selection model techniques (i.e., p-curve, p-uniform, and a three-parameter selection model; Simonsohn et al., 2014; van Assen, van Aert, & Wicherts, 2015; McShane et al., 2016).

Finally, our approach extends research on the influence of QRPs on meta-analysis and meta-analytic adjustment. Previous work in this area chiefly inspected the effects of QRPs on p-curve (Simonsohn et al., 2014; Bishop & Thompson, 2016). We expand on this work by simulating more diverse forms of QRPs and observing their effects on a wider family of bias-correction methods.

Methods

We simulated data for 432 unique combinations of five fully-crossed factors (Table 1).We simulated 1,000 meta-analyses for each of the 432 conditions [1]

All simulated individual studies had a two-group experimental design. Sample sizes per cell were drawn from an empirically-derived distribution of sample sizes reported in social psychology.

We obtained this distribution by accessing data from a recent study that quantified per-study sample size from articles published in six prominent social and personality psychology journals (Fraley & Vazire, 2014). To better represent our simulation's emphasis on between-subject designs, we excluded samples from personality journals, focusing on those from social psychology journals. We assumed that all studies included simple two-groups between-subjects designs. The resulting distribution of samples per cell was strongly right-skewed, with mean = 53, median = 43, SD = 44, minimum = 10, maximum = 388, and skewness = 4.5.

Independent samples were then randomly generated for the control and experimental group, where observations in the control group were drawn from a normal distribution with $\mu = 0$ and $\sigma = 1$, and observations in the experimental group were drawn from a normal distribution with $\mu = D_i$ and $\sigma = 1$. D_i was defined as the sum of δ and τ_i, where τ_i was drawn from a normal distribution with mean 0 and standard deviation τ. Note that D_i, therefore, represented a study-specific true effect that varied randomly if τ was greater than 0. Cohen's d and the variance v of the effect size were then calculated (see supplemental material) and a two-tailed independent-samples t-test was applied to generate a t-value and p-value.

In the conditions where studies were affected by publication bias, studies were produced as described above but were deleted if the results were statistically non-significant ($p \geq .05$) or if the effect size was negative ($d < 0$). Studies were continually produced until the target number of studies had been reached. For example, for a meta-analysis with $k = 10$ and 60% publication bias, 4 studies were produced without publication bias (producing either a significant or non-significant result), and then the simulation ran until 6 additional studies with $p < .05$ and $d > 0$ were produced.

We studied four forms of QRPs and simulated different levels of severity by combining multiple QRPs. These were organized as three types of *individual QRP strategies*: (1) pure (no use of QRPs), (2) moderate (optional dependent variables and some optional stopping), and (3) aggressive

[1] For a selection of conditions we also computed 10,000 simulations and compared the results to that of 1,000 simulations. These comparisons clearly demonstrated that 1,000 replications lead to sufficiently stable estimates (see supplementary material, https://osf.io/rf3ys).

(use of optional outliers, optional dependent variables, optional moderators, and extensive optional stopping).

As it is unlikely that every researcher in a field applies QRPs in the same fashion, we defined three *QRP environments* to describe possible prototypical research fields with a specific severity of QRP application. Each QRP environment was characterized by a mixture of simulated researchers with individual QRP strategies: (1) none (100% of simulated researchers adopted the pure strategy); (2) medium (30% pure, 50% moderate, and 20% aggressive); and high (10% pure, 40% moderate, and 50% aggressive). Critically, the QRP percentages and the publication bias percentages were independent, such that the QRP percentages held for the studies produced with and without the influence of publication bias. Further details on how we implemented QRPs are available in the supplemental material https://osf.io/rf3ys.

Meta-analytic methods

We examined the performance of eight estimators. Further details on each method available in the supplemental material https://osf.io/rf3ys.

Random-effects meta-analysis (RE). We applied the random-effects meta-analysis as described above using the `metafor` package in R (Viechtbauer, 2010). This approach makes no adjustment for publication bias or QRPs. We used the Dersimonian-Laird method for estimating between-study variance.

Trim-and-fill (TF). Trim-and-fill (Duval & Tweedie, 2000) is an adjustment for publication bias based on funnel plot asymmetry (a scatter plot of effect size estimates against the standard error of those estimates). Publication bias introduces clear rightward asymmetry in a funnel plot (see supplemental material) by censoring non-significant and negative observations. Trim-and-fill iteratively removes (i.e., trims) observations from one side of the funnel plot until a criterion for symmetry is met and then fills observations back into the funnel plot along with imputed observations reflected about the mean. Standard meta-analytic methods are then applied to a data set including both observed and imputed studies. We use the default algorithm provided by the `metafor` package.

PET. PET (T. Stanley & Doucouliagos, 2014) is a meta-regression approach to adjusting for publication bias. PET operates through adjustment for small-study effects: When there is a small study effect, the observed effect size gets smaller as the standard error shrinks. PET fits a linear regression line to this relationship, then extrapolates to estimate the effect size of a hypothetical study with a standard error of zero (i.e., a study with infinite sample size). The resulting PET estimate is an estimate of the true underlying effect after adjustment for small-study effects. Again, as small-study effects may have benign causes, this may represent a substantial overadjustment.

PET is the weighted-least-squares regression model where effect size is regressed on its standard error: $d_i = b_0 + b_1 se_i + e_i$, where b_0 and b_1 are the intercept and slope terms describing the linear relationship between the ith effect size estimate d_i and its associated standard error se_i. The regression model is weighted by the inverse of the variance (i.e., the squared standard errors) of the effect size estimates. The intercept b_0 represents the estimated effect size when the standard error is zero.

PEESE. PEESE (T. Stanley & Doucouliagos, 2014) is, like PET, a meta-regression model for the adjustment for small-study effects. Whereas PET fits a linear relationship between effect size and standard error, PEESE fits a quadratic relationship. The rationale for this quadratic relationship is this: Assuming there is some true effect, low-precision studies are poorly powered and publishable only when the effect is badly overestimated. On the other hand, high-precision studies will be well-powered and routinely publishable without such overestimation. Thus, publication bias (and the observed small-study effect) is stronger when the standard error is large and weaker when the standard error is small. A quadratic curve can model these differences in the degree of bias. PEESE is the weighted-least-squares regression model where effect size is regressed on the square of the standard error: $d_i = b_0 + b_1 se_i^2 + e_i$. As in PET, the weights are the inverse of the variances and the intercept is interpreted as an estimate of the true underlying effect that is uninfluenced by small-study effects.

PET-PEESE. Simulation studies find that PET outperforms PEESE when the true underlying effect is zero, as it underestimates the size of non-zero true effects, whereas PEESE outperforms PET when the true underlying effect is non-zero, as it overestimates the size of null effects (T. Stanley & Doucouliagos, 2014). In an attempt to offset the opposite biases of PET and PEESE, T. Stanley and Doucouliagos (2014) suggested the conditional estimator PET-PEESE. PET-PEESE considers the statistical significance of the PET estimate to decide whether PET or PEESE is taken as the final estimate. When the estimate from PET is statistically non-significant (i.e., the estimated true effect is not distinguishable from zero), the PET estimate is taken. In contrast, when the estimate from PET is statistically significant, the PEESE estimate is used as the value for the conditional PET-PEESE procedure.

***p*-curve.** A p-curve is the distribution of all statistically significant p-values from the set of studies of interest (i.e., ps < 0.05; Simonsohn et al., 2014). The shape of the p-curve is a function of the statistical power of the studies, which is itself a function of the sample sizes and the true effect size. When studies have no statistical power–that is, when the null is true–the distribution of significant p-values is uniform between .00 and .05. With increasing power (i.e., larger effects, larger samples), the p-curve becomes increas-

ingly right-skewed, with $.00 < p < .01$ becoming more probable than $.04 < p < .05$. Because the degree of right skew is a function of the average study power, *p*-curve can use the degree of right-skew to (a) test the absence of a real effect (H_0: $\delta = 0$), and (b) estimate the average study power, and thus, the average effect size in a fixed-effect model.

p-curve estimation does not provide confidence intervals. Therefore we could not apply the coverage metric (see below) to *p*-curve. From our initial simulations it became clear that *p*-curve provides very poor results when there are ≤ 3 significant and directionally consistent studies (see supplementary material). We marked results from *p*-curve as missing in this case.

***p*-uniform.** Like *p*-curve, the *p*-uniform method also considers only the statistically-significant results and uses the fact that the distribution of *p*-values is flat under the null (van Assen et al., 2015). It yields a fixed-effects estimate of the true effect by finding the value $d*$ for $H_0 : \delta = d*$ which makes the distribution of *p*-values as flat as possible.

p-uniform provides a hypothesis test, an estimate of the bias-corrected effect size, and a confidence interval around that estimate. Computationally, *p*-curve and *p*-uniform only differ by an alternative implementation of the estimation algorithm, and so *p*-curve and *p*-uniform are expected to have similar strengths and weaknesses (McShane et al., 2016).

Again, we applied *p*-uniform only when four or more significant, directionally consistent studies were available. On rare occasions *p*-uniform failed to produce a lower-bound to its 95% CI. We marked results from *p*-uniform as missing in these cases.

Three-parameter selection model (3PSM). Selection models, first introduced by Hedges (1984) and later extended by Iyengar and Greenhouse (1988) and Hedges and Vevea (1996), attempt to model the process by which results are either published or file-drawered. We employed the three-parameter selection model (3PSM) as developed by Iyengar and Greenhouse (1988) and recommended by McShane et al. (2016). This model's three parameters represent the population average effect size μ, the heterogeneity of the random effect sizes τ^2, and the probability that a nonsignificant effect enters the literature The joint likelihood function of these three parameters is then maximized given the observed data. In contrast to trim-and-fill or PET-PEESE, this method provides an explicit model for publication bias. However, it does not model the influence of QRPs, so it is unclear whether 3PSM performs well when data have been flexibly analyzed.

Performance metrics

For the hypothesis test of whether the true underlying effect provided by each meta-analytic method differs from zero, we evaluated the false positive (Type I error) rate at $\delta = 0$ and the true positive rate (i.e., the statistical power) at $\delta = 0.2$, 0.5, and 0.8.

Following the recommendations of Burton, Altman, Royston, and Holder (2006), we measured the bias-adjusting performance of each method in terms of mean error (ME), root mean squared error (RMSE), and 95% coverage probability.

ME (often called bias) is the average of the deviations of each estimate from the true effect (i.e., the errors). Nonzero ME indicates that the expected value of the estimate does not converge on the true value in the long run, being instead too high or too low. ME is not sensitive to variance in estimates, so it is possible that a method produces low ME by equivalently strong over- and under-estimation. Such a case would yield estimates that are accurate *on average*, but any individual estimate could be very far from the truth.

RMSE incorporates information about average error as well as the variance (i.e., the efficiency) in the estimates. It is possible to observe low RMSE (i.e., high efficiency) even when a method produces estimates that are consistently biased in one direction. Imagine a very narrow distribution of estimates that is centered a bit above the true value. On average, the estimates are too high, but the variability of these estimates will be low. Thus, a method's estimation performance must be considered in terms of both ME and RMSE. For both ME and RMSE, values as close to zero as possible are desirable.

We also examined the performance of the 95% confidence intervals (CIs) associated with each method (except *p*-curve, which does not provide a CI) by examining 95% coverage probability. This coverage probability is the proportion of each method's 95% confidence intervals that included the true value of δ. Optimally, the coverage probability is at the nominal level of 95%. Low coverage is obviously problematic as most intervals do not contain the true value, but coverage rates higher than the nominal 95% are also cause for concern and may indicate exceedingly-wide intervals.

Presentation of results

We simulated 1,000 meta-analyses under 432 unique conditions (Table 1) and analyzed each with eight different meta-analytic methods. Here, we avoid an exhaustive presentation of the results and focus instead on providing a more general overview. As shorthand, we will refer to effects of size $\delta = 0.2$, 0.5, and 0.8 as small, medium, and large, respectively.

Some methods had computational difficulties in some conditions where $\delta = 0$. For example, in conditions without publication bias and a true effect size of $\delta = 0$, most studies that enter a meta-analysis are non-significant. It very rarely happens that ≥ 4 studies out of $k = 10$ reach statistical significance, and therefore *p*-curve and *p*-uniform could not be applied in most of the 1000 simulation runs for those conditions. We report the summaries of the successful runs, but readers should be aware that this implies a conditional interpretation: In many cases this method did not provide an

estimate at all; but when an estimate was provided, it had the reported ME, RMSE, and error rates. In Figures 1 and 2, symbols indicate if a method did not yield an estimate in more than 25%, 50% or 75% of the 1000 simulation runs.

In the following section, we discuss estimators' performance across all conditions and performance metrics, but we provide figures only for conditions in which $\delta = 0$ or $\delta = 0.50$. Additionally, rather than providing exact values for ME, RMSE, and coverage, we display the distributions of effect size estimates in terms of means and the 95% quantile ranges. In our view, these values are more intuitive, and they allow a visual assessment of ME (the means) and RMSE (the range).

Furthermore, initial simulations indicated that some methods occasionally provided bias-corrected estimates that were both statistically significant and negative. Because it seems highly improbable that one could recover a true effect of opposite sign, the following results are given with all negative meta-analytical effect size estimates ($d < 0$) redefined as zero ($d = 0$) and counted as failing to reject the null hypothesis.

All of our findings, including exact values for ME, RMSE, and coverage probability for all conditions, calculated with both negative estimates set to zero and not, are available in supplemental material (https://osf.io/rf3ys). We also provide several interactive figures and tables (http://www.shinyapps.org/apps/metaExplorer/) as a means of encouraging researchers to explore combinations of conditions that they find to be particularly relevant to their own work.

Results

Figure 1 shows both the false positive rates (when $\delta = 0$) and the statistical power (when $\delta = 0.50$) of each method. Figure 2 displays the mean and 95% quantiles of effect size estimates from each method. In discussing these results, we first focus on conditions where QRP environment was set as "None". The influence of QRPs will be discussed subsequently.

No publication bias (0%), no QRPs

Hypothesis testing. As expected, random-effects meta-analysis was conservative under the null (false positive rates of 2-5%). Its power was good under homogeneity, detecting $\delta = 0.2$ with 80% power at $k = 10$. Heterogeneity impaired power to detect small effects; 80%+ power required $k = 30$ when $\tau = 0.2$ and $k = 60$ when $\tau = 0.4$.

Trim-and-fill was less conservative under the null and heterogeneity, experiencing false positive rates of up to 16% as k and τ increased. Its power was similar to that of random-effects meta-analysis but slightly poorer under heterogeneity.

PET was conservative under the null but became less so with increasing k and τ up to an 8% false positive rate. PET's power to detect effects was much reduced compared to random-effects meta-analysis, especially when $\delta = 0.2$ and when $\tau \geq 0.2$. PEESE behaved similarly to PET, with similar false positive rates but substantially better power. PET-PEESE's performance was better than that of PET but worse than that of PEESE.

Under homogeneity, When the null was true and there was no publication bias, *p*-curve and *p*-uniform rarely had enough statistically significant results to meta-analyze. When there were enough results to meta-analyze, Type I error was sometimes as high as 9%. Moreover, power to detect $\delta = 0.2$ was poor, reaching only 60% even at $k = 100$. Power was good for medium and large effect sizes. Heterogeneity increased false positive rates, which became quite high given increasing τ and k (e.g., 100% at $k = 100$ and $\tau = 0.4$).

3PSM was excessively conservative, with false positive rates below the nominal threshold (1-2%). Similarly, its power was weaker than that of random-effects meta-analysis or trim-and-fill, particularly with increasing τ.

ME and RMSE. Not surprisingly, random-effects meta-analysis was the most consistently unbiased and efficient estimator in the absence of publication bias. Trim-and-fill returned generally unbiased estimates, although large effects were very slightly underestimated in the presence of heterogeneity. Trim-and-fill yielded a very slight loss of efficiency relative to random-effects meta-analysis.

PEESE performed similarly to these two methods, but suffered from a loss of efficiency and downward bias that grew worse with increasing δ and τ. PET-PEESE and PET showed a similar but more severe pattern.

p-curve and *p*-uniform were only unbiased and efficient when δ was medium or large and $\tau = 0$. As δ decreased and τ increased, estimates from *p*-curve and *p*-uniform gained a (sometimes severe) upward bias.

3PSM performed similarly to trim-and-fill. It was very slightly biased upwards for null effects when k was small and τ was large, overestimating these as $\delta = 0.05$. 3PSM was slightly less efficient than trim-and-fill.

95% CI coverage. Random-effects meta-analysis demonstrated coverage rates near the nominal level (91-97%) across all levels of δ, τ, and k. Trim-and-fill tended to show undercoverage, particularly with increasing heterogeneity and high k, dipping as low as 55%. PET and PEESE showed acceptable coverage under homogeneity (92-97%), but poorer coverage for large effects, high heterogeneity, and large k.

p-uniform was often ineligible for application due to too few significant results (e.g., when the null was true). For non-null homogeneous effects, coverage was good (93-97%). Coverage was poor under heterogeneity, presumably due to *p*-uniform's upward bias.

3PSM showed good coverage under homogeneity (94-96%) and acceptable coverage (89-96%) under heterogeneity, with the poorest coverage observed for $k = 10$ under

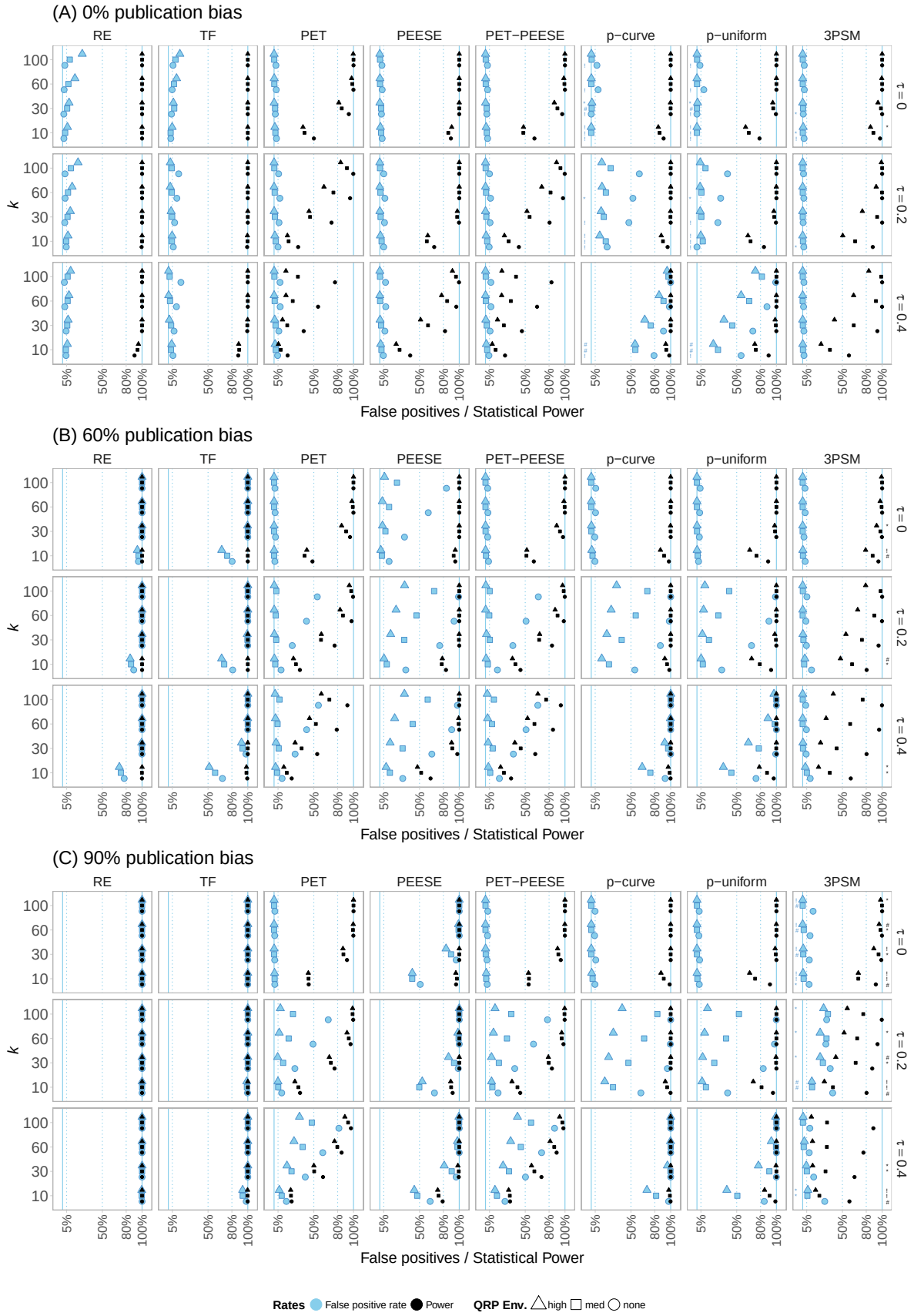

Figure 1. False positive rates (when $\delta = 0$) and statistical power (when $\delta = 0.5$) for all methods across all conditions. k = meta-analytic sample size. τ = heterogeneity. Symbols on the left and right border of each panel indicate when a method computationally failed in a substantial proportion of the 1000 simulations: *: < 750/1000, #: < 500/1000, !: < 250/1000 successful computations. Figure available at https://osf.io/rf3ys, under a CC-BY4.0 license.

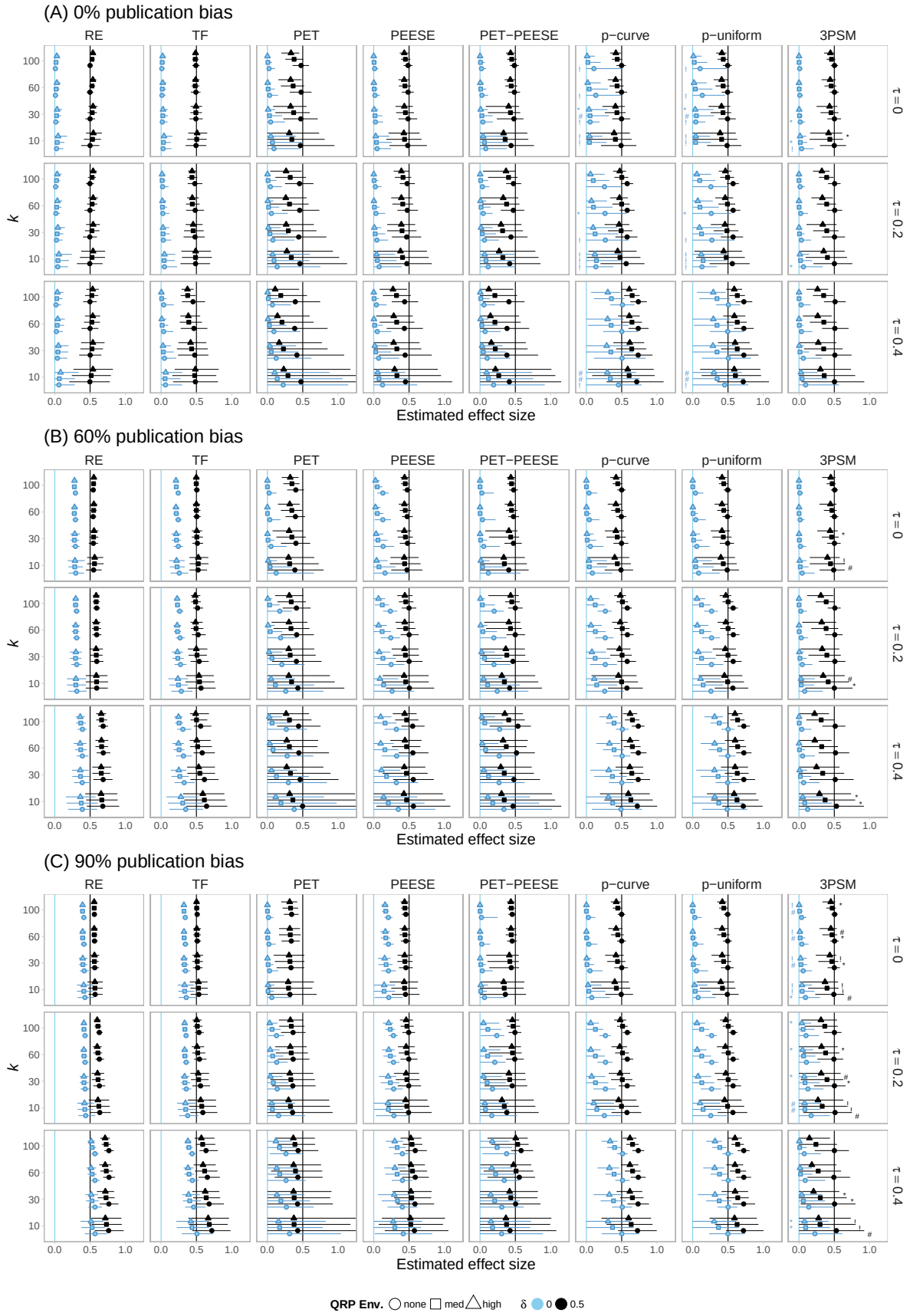

Figure 2. Means (points) and inner 95% quantile ranges (whiskers) of effect size estimate distributions when $\delta = 0$ and when $\delta = 0.5$ for all methods across all conditions. k = meta-analytic sample size. τ = heterogeneity. Symbols left and right of the whiskers indicate when a method computationally failed in a substantial proportion of the 1000 simulations: *: < 750/1000, #: < 500/1000, !: < 250/1000 successful computations. Figure available at https://osf.io/rf3ys, under a CC-BY4.0 license.

heterogeneity.

Partial publication bias (60%), no QRPs

Hypothesis testing. Not surprisingly, publication bias led to false positive rates near 100% for random-effects meta-analysis. Trim-and-fill had similarly unacceptable false positive rates, indicating an inability to adjust for publication bias.

PET was conservative under homogeneity with false positive error rates of 1-2%. Heterogeneity, however, drastically increased its false positive rate to as much as 56%, a pattern that was exacerbated by increasing sample sizes. PET's power to detect small effects was poor, requiring large sample sizes (e.g., $k = 100$ when $\tau = 0.4$) to approximate 80% power. For medium and large effects, power was acceptable given $k \geq 30$ when $\tau = 0.2$ or $k > 60$ when $\tau = 0.4$.

PEESE suffered false positive error rates in excess of the nominal rate (at best, 12%), which grew much worse with increasing k and heterogeneity (up to 99%). PEESE had $\geq$ 80% power to detect nonzero effects for $k \geq 30$.

PET-PEESE provided a compromise between the strengths and weaknesses of PET and PEESE, although its performance was more akin to that of PET (e.g., increasing false positive rates with increasing heterogeneity and sample size).

p-curve and *p*-uniform were conservative under homogeneity, with false positive rates of 2-6%. These rates became much worse with increasing heterogeneity, however (e.g., $\geq$ 99% at $\tau = 0.4$ and $k \geq 30$). Power to detect $\delta \geq 0.5$ was good even at $k = 10$, but to detect $\delta = 0.2$ required 60 studies or more. *p*-curve tended to reject the null, regardless of the value of δ, slightly more often than *p*-uniform.

When $\tau = 0$, 3PSM had conservative false positive rates (1-3%) and was better-powered to detect effects than all other bias-correcting methods (other than trim-and-fill). Performance degraded with increasing heterogeneity. With heterogeneity and a few studies ($k = 10$), false positive rates could exceeded the nominal threshold (up to 11%). When $\tau = 0.2$, power remained good for $\delta \geq 0.5$, but $\delta = 0.2$ required 60 studies. When $\tau = 0.4$, medium effects required $k = 30$ for good power, and even 100 studies only reached 70% power to detect a small effect.

ME and RMSE. When publication bias was set to 60% and $\tau = 0$, the random-effects meta-analysis estimate was biased upwards. This bias was, of course, strongest when $\delta = 0$ and smallest when $\delta = 0.8$. Trim-and-fill reduced this bias only slightly. When publication bias was minimal (e.g., when the true effect was large), trim-and-fill slightly underestimated the effect. As heterogeneity increased, upward bias in both random-effects meta-analysis and trim-and-fill increased, presumably because studies on the high side of τ were more likely to get published. Efficiency also decreased due to the added variability caused by τ.

Under homogeneity, PET was generally accurate and efficient at retrieving $\delta = 0$, but, as expected, demonstrated a slight downward bias when $\delta > 0$. Increasing τ systematically decreased efficiency for PET. The effect of heterogeneity on PET's bias, however, was more complicated: When δ was 0 or 0.2, τ caused upward bias in PET, but when δ was 0.5, τ reduced PET's downward bias, and when δ was 0.8, τ exacerbated PET's downward bias.

PEESE overestimated $\delta = 0$ and $\delta = 0.2$, was generally unbiased for $\delta = 0.5$, but slightly underestimated $\delta = 0.8$. This pattern became slightly more severe with increasing τ. PEESE became consistently less efficient with increasing τ.

Again, PET-PEESE seemed to perform somewhere between PET and PEESE, tending to have better estimates than PEESE when $\delta = 0$ and better estimates than PET when $\delta \geq 0.2$.

p-curve and *p*-uniform were both generally unbiased and efficient when $\tau = 0$. Increases in τ, however, caused a large upward bias in *p*-curve and *p*-uniform–a bias so large, in fact, that the uncorrected random-effects meta-analysis was somewhat less biased when $\tau = 0.4$.

3PSM tended to equivalent or superior to all other methods across all levels of δ and τ, providing less bias and greater efficiency. In this sense, 3PSM was clearly the best choice, but in some cases, its performance was not quite ideal (e.g., 3PSM lost efficiency with increasing τ at lower levels of k).

95% CI coverage. Coverage of random-effects meta-analysis was very poor due to the effects of publication bias, and this was made worse with increasing τ. Trim-and-fill's performance was similarly unacceptable.

PET had good coverage for small or null effects under homogeneity (93-96% coverage), but coverage was poor for medium and large effects, particularly at large k. Heterogeneity impaired PET's coverage for small and null effects and improved it for medium and large effects. PEESE suffered from undercoverage across all conditions, particularly as k and τ increased. PET-PEESE only performed somewhat well when $\delta = 0$ and $\tau = 0$. Otherwise, PET-PEESE showed generally unacceptable coverage.

p-uniform showed good to over-coverage under homogeneity (93-98%) but poor coverage under heterogeneity due to the bias inflicted by those conditions.

3PSM likewise showed good to over-coverage under homogeneity (93-97%), but was far less sensitive to increases in τ than *p*-uniform. In some cases, (e.g., small k and large τ), coverage could fall as low as 83%.

Widespread publication bias (90%), no QRPs

When publication bias was set such that at least 90% of reports were statistically significant, results resembled an intensified version of 60% publication bias. Because publication bias was stronger, sets of meta-analyzed results often consisted of only significant results, especially when k was

small. Critically, this prevented the application of 3PSM in many data sets.

Hypothesis testing. Under these circumstances, random-effects meta-analysis and trim-and-fill nearly always returned a significant result, regardless of the underlying true effect.

PET was too conservative under the null and homogeneity (1% false positive rate), and power for small effects was poor even at $k = 100$ (at most 73%). However, medium and large effects could be detected with 80% power at $k = 30$. Increasing heterogeneity and sample size could drastically increase false positive rates. Heterogeneity had an asymmetrical effect on PET's power, increasing power when $\delta = 0.2$ but reducing power when $\delta \geq 0.5$.

PEESE, in contrast, reported statistical significance in nearly all circumstances, exhibiting false positive rates of 51% or more and power of 72% or more. As in other cases described above, PET-PEESE tended to perform better than PEESE when $\delta = 0$ and better than PET when $\delta > 0$.

p-curve and *p*-uniform were appropriately conservative (false positive rate of 3-6%) under homogeneity. With heterogeneity, however, false positive rates increased to unacceptable levels, particularly with large k and τ (between 39% and 100%). Power for medium-to-large effects was good at $k = 10$, but small effects required $k > 30$. Not surprisingly given the relationship between heterogeneity and overestimation by *p*-curve and *p*-uniform, power increased with increasing τ.

3PSM's false positive rates were in excess of the nominal threshold, ranging from 8-13% under homogeneity and reaching as high as 38% under heterogeneity with small k. Power was acceptable under homogeneity, requiring $k = 30$ for small effects and $k = 10$ for medium and large effects. Heterogeneity, however, reduced power, particularly for small effects, which were not detected with 80% power even at $k = 100$.

ME and RMSE. In general, the same patterns of bias and inefficiency observed when publication bias was set to 60% were exacerbated when publication bias was increased to 90%.

It is worth noting that *p*-curve and *p*-uniform are more efficient under this stronger publication bias as compared to their performance under weaker publication bias. As these methods consider only the statistically significant results, stronger publication bias means that more results are statistically significant, reducing the loss of information caused by ignoring null results. *p*-curve and *p*-uniform returned unbiased estimates so long as $\tau = 0$, but sometimes dramatically overestimated the effect when $\tau > 0$.

3PSM returned generally unbiased estimates, although some upward bias remained for $\delta = 0$ when heterogeneity was present and $k \leq 30$. We discuss this weakness and its likely causes in the discussion section.

95% CI coverage. As before, random-effects meta-analysis had unacceptable coverage due to publication bias, overestimating the magnitude of null-to-medium effects and overstating the precision of large effects. Trim-and-fill had comparably unacceptable coverage, although its coverage was slightly better than that of random-effects meta-analysis for medium effect sizes or under heterogeneity.

PET showed good coverage under the null and homogeneity (94-95%) but generally poor coverage otherwise, especially with increasing k. PEESE generally had poor coverage, particularly for small effects, large k, and/or large τ. Like PET, PET-PEESE had good coverage under the null and homogeneity, but poor coverage otherwise, rarely exceeding rates of 85%, and often performing much worse.

p-uniform showed good coverage under homogeneity (94-96%) but poor coverage (similar in degree to PET-PEESE) under any amount of heterogeneity.

3PSM showed good coverage for non-null effects under homogeneity (94-96%). Coverage was poorer when effects were small or null and/or when heterogeneity was present, with a minimum coverage of 62%.

The influence of QRPs

Without publication bias, increasing QRPs led to increasing false positive rates for random-effects meta-analysis, and trim-and-fill. For all other bias-correction methods, increasing QRPs generally led to decreasing false positive rates. The effect of QRPs on power was, in general, negative, although in some cases (e.g., PEESE under no publication bias), it seemed non-linear.

Interestingly, QRPs in the absence of publication bias led to only a slight upward bias and increase in inefficiency in the random-effects meta-analysis estimate. However, in the face of publication bias, QRPs did not seem to inflict any additional upward bias in the random-effects estimate, and actually may have very slightly reduced the degree of overestimation.

QRPs tended to reduce estimates from bias-correcting methods across most combinations of simulation parameters. In general, as the QRP environment grew more severe, so too did this downward adjustment. This downward adjustment was particularly pronounced in 3PSM when heterogeneity was also present.

Sometimes the downward bias canceled out an upward bias and resulted in more accurate estimates. For example, *p*-curve and *p*-uniform are upward-biased in the presence of heterogeneity, but downward-biased in the presence of QRPs. Likewise, PEESE is upward-biased when the true effect is very near zero, but downward-biased in the presence of QRPs. Under some combinations, these competing biases were of approximately equal and opposite magnitude, yielding less biased estimates.

Discussion

We inspected and compared the efficacy of meta-analytic adjustments for bias across hundreds of thousands of simulated literatures representing a range of true effect sizes, degrees of heterogeneity, degrees of publication bias, and degrees of questionable research practices. We assessed the results according to both the ability to reject a null effect/detect a true effect and the ability to estimate the mean of the distribution of true underlying effects.

Each bias-correcting method suffered from some degree of weakness. For example, trim-and-fill did not adjust enough for bias to recover $\delta = 0$ after any amount of publication bias (Figure 2) and small and null effects ($\delta \leq 0.2$) remained substantially overestimated (Figure 3). PET, PEESE, PET-PEESE, *p*-curve, and *p*-uniform showed (sometimes rapidly) increasing false positive rates as a function of changes in publication bias, heterogeneity, and k (Figure 2). PET, PEESE, and PET-PEESE all also showed a downward (and sometimes upward) bias under some conditions, whereas *p*-uniform and *p*-curve exhibited a very consistent upward bias with increasing heterogeneity (Figure 3 and Figure S3). The performance of the three-parameter selection model (3PSM) was the most promising. It was the only method to provide acceptable false positive rates in most scenarios, and its estimates tended to have the least bias and the most efficiency across conditions.

Recommendations for meta-analysts in psychology

Which condition is most like my data?. In our simulations, we fully crossed many types of parameters. Each method has conditions where it works well and some (or many) where it does not. Unfortunately, one cannot know for sure which of the 432 conditions best describes one's research environment, and consequently, which method would be the best choice. However, some of the conditions are more plausible than others, and informed guesses are possible.

First, always expect some degree of publication bias. This becomes evident when one considers that, in the field of psychology/psychiatry, more than 90% of all published hypothesis tests are significant (Fanelli, 2011) despite the average power being estimated as around 35% (Bakker, van Dijk, & Wicherts, 2012).

Second, heterogeneity is very likely (Higgins, 2008; McShane et al., 2016). Even in the ideal case of exact replications of computer administered experiments, multilab collaborations revealed non-negligible heterogeneity in a large proportion of the investigated effects.

Third, true effects are rather small. The median published effect size is around $d = 0.3$ to 0.4 (Richard, Bond, & Stokes-Zoota, 2003; Bosco, Aguinis, Singh, Field, & Pierce, 2015), and as these naive estimates are not corrected for publication bias, the true values are almost certainly lower.

Taken together, we think that the conditions with $\tau = 0.2$, publication bias $\geq 60\%$, effect sizes $\delta \leq 0.5$ under H_1, and some amount of QRPs ("medium") might be a realistic description of many fields in psychology. However, this should not be seen as a "one-size-fits-all" solution, and specific literatures might deviate from this typical situation.

What to avoid. We strongly recommend analysts not use a bias-correction method when one can be sure that there is no publication bias, such as in the case of analyzing only the data from a Registered Replication Report (RRRs; Simons, Holcombe, & Spellman, 2014), a series of registered reports (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014). Even in the presence of some QRPs, random-effects meta-analysis has less bias, more efficiency, and more power than all other methods we examined.

In any other setting, however, one should expect publication bias and not trust either random-effects meta-analysis or trim-and-fill. Given likely values of publication bias, both methods have false positive rates close to 100% and unacceptable upward bias.

The performance of PET, PEESE, and PET-PEESE is characterized by high variability, with estimates ranging from severe under-correction to severe over-correction. Given how difficult it is to determine which condition best represents one's particular situation, the safest route is to avoid using these methods with data from research on psychology.

Similarly, it seems that it is also best to avoid *p*-curve and *p*-uniform with data from research in psychology. With $k < 4$ significant and directionally consistent studies, estimates were highly variable. Furthermore, both methods have unacceptable false positive rates and upward biases under heterogeneity, which, unfortunately, is the norm in psychology (see also van Aert, Wicherts, & van Assen, 2016). Some applications of *p*-curve, such as evaluating the evidential value of journals or authors (e.g. Motyl et al., 2017) are necessarily based on heterogeneous data sets. Given our current results, such an application nearly guarantees a positive result, regardless of the true underlying effect.

What to do? Use 3PSM with caution. Overall, 3PSM showed the most promising results across conditions, and at the current state of bias-correcting methods, we generally recommend its use. However, at least five weaknesses in the method should be noted.

First, 3PSM shows a downward bias under QRPs (at least as we have modeled them), especially with heterogeneity. It is encouraging that the bias is downward, rather than upward, since it allows 3PSM to maintain conservative false positive rates. Still, this underestimation of effects is undesirable, as it may lead to false negatives in meta-analyses of strongly *p*-hacked studies.

Second, there's no proper implementation of 3PSM when all studies are statistically significant (see Appendix A in

McShane et al., 2016). Under these circumstances, 3PSM provides no confidence interval, sometimes fails to give a p-value, and returns estimates that may retain substantial upward bias and excessive false positive rates. When all results are statistically significant, we recommend ignoring results from 3PSM.

Third, 3PSM had unacceptable false positive rates (up to 38%) when there was heterogeneity and very strong publication bias. Under these conditions, 3PSM seems to have difficulty disentangling the joint likelihood function of the average effect, the heterogeneity, and the degree of publication bias. The heterogeneity is often underestimated and the average effect size overestimated. Larger k can help to estimate τ and reduce the bias in effect size estimates, but even so, bias may remain.

Fourth, 3PSM relies on the assumption that δ is normally distributed, which matches the assumptions underlying our simulations. Previous work has found that violating this assumption can lead to poorer performance of the selection model (Hedges & Vevea, 1996; see Lee & Thompson, 2008 for first approaches to specify non-normal distributions for the random effects). Unfortunately, given that the appropriateness of this assumption is essentially unverifiable, this may be a serious limitation. Of course, all the meta-analytic methods we have compared here rely on this or other unverifiable assumptions, so this warning applies to each of them in some sense.

Finally, 3PSM requires specification of significance cut-points. In this simulation, the 3PSM cut-point was set at $p < .05$, which is identical to the publication bias process we used for data generation in the simulation. Movement of this cut-point may degrade the performance of 3PSM.

To summarize, 3PSM it is relatively unbiased and efficient, although some bias and high false positive error rates remain when publication bias is strong, k is small, QRPs are severe, or heterogeneity is large. Although our results suggest that 3PSM is superior to all investigated competitors, 3PSM is imperfect. It would be better to improve how we perform research as a field than to place our trust in post-hoc corrections for biased and incomplete data.

Limitations

Alternative formulations of publication bias and QRPs. Our simulation relied on a simple selection filter based on a single $p < .05$ threshold for inducing publication bias. Given the increase in appreciation for well-powered null results, it is possible that selection filters are less rigid for large sample sizes. Well-powered results immune to publication bias may be expected to improve the performance of methods like PEESE, which emphasizes high-powered studies.

We simulated some forms of QRPs and found that they often had a downward bias on adjusted estimates. However, QRPs are a heterogeneous family of behaviors, which may each have their own effects on naive meta-analysis and meta-analytic adjustments. As the nature and prevalence of QRPs are difficult to estimate, it is unclear whether the QRPs in this simulation are a close match to those in the real world. Still, the simulated QRPs seem roughly appropriate and provide a useful starting point.

Distributions of true effects and study sample sizes. We simulated the true effect δ as following a normal distribution with mean μ and standard deviation τ. Several methods make this assumption, and it is not clear how the performance of these methods would change had we simulated a different distribution on δ. It would be valuable for future simulation studies to explore the robustness of meta-analytic methods to non-normal distributions of δ.

For all simulated meta-analyses, we sampled study sample sizes from a distribution taken from an empirical review of sample sizes in the literature (Fraley & Vazire, 2014). In reality, however, sample sizes within a meta-analysis could be more similar to each other than what we have modeled. If, for example, researchers plan sample sizes based on previous studies on the same topic, the body of work in that area will have less variation in sample sizes than we simulated. In particular, PET-PEESE relies on having both small and large samples to infer the magnitude of small-study effects. Hence, a reduced variance in sample sizes may reduce the performance of PET-PEESE and other methods.

Fraudulent and mistaken results. Fraudulent results are probably rare in psychology, but it is difficult to estimate their prevalence. The same can be said for honest errors in data collection or reporting that produce false findings. None of the collected techniques are intended to detect and adjust for the influence of either of these factors. Indeed, fraudulent and mistaken results at the study-level may lead to considerable degradation of the performance of the meta-analytic adjustments we examined. For example, a falsified report with both a strong effect size and high precision can be expected to have strong influence over the selection model or funnel plot and may prevent appropriate downward adjustment.

Ways forward

Concerning further method development in research synthesis, we see the best prospects in focusing on selection models (for some promising work in Bayesian selection models, see Guan & Vandekerckhove, 2015). We agree with McShane et al. (2016) that the three-parameter selection model, which accounts for heterogeneity and models different levels of publication bias, should be the minimal default model for applied meta-analyses. Researchers developing new bias-correcting methods are encouraged to use our simulated data sets as a sort of benchmark data[2], which allows

[2] https://github.com/nicebread/meta-showdown/tree/master/simParts

an easy comparison of a new method's performance with the methods investigated here.

Despite the relatively promising results of the 3PSM, we stress that meta-analysis in psychology is *difficult*. Observable issues such as small samples–in both the primary literature and at the level of the meta-analysis–interact with heterogeneity and bias, both of which are unknowable in terms of severity and functional form (e.g., do the true effects follow a normal distribution? Is publication bias applied as we have modeled it?). Thus, it is hard to interpret the results of a meta-analysis in psychology, just as it is difficult to interpret the results of any single replication study (Braver, Thoemmes, & Rosenthal, 2014; D. J. Stanley & Spence, 2014; Fabrigar & Wegener, 2016).

Along the same lines of what has been argued for replication results (e.g., D. J. Stanley & Spence, 2014), we therefore conclude that the field should modify its expectations about meta-analysis–researchers should not expect to produce a conclusive, debate-ending result by conducting a meta-analysis on an existing literature. Instead, we imagine meta-analyses may serve best to draw attention to the existing strengths and/or weaknesses in a literature (e.g., Carter, Kofler, Forster, & McCullough, 2015; van Elk et al., 2015; Hilgard, Engelhardt, & Rouder, n.d.), which can then inspire a careful re-examination of methodology and theory followed by, if necessary, large-scale, preregistered replication efforts (e.g., Hagger et al., 2016). Such replications can then be summarized with random-effects meta-analytic methods, which work extremely well in the absence of bias (Figures 1A and 2A).

Conclusion

Comparing adjustments for bias using effect sizes and sample sizes typical of psychology, we find that the majority of methods showed unacceptable performance. Trim-and-fill is evidently incapable of retrieving $\delta = 0$ and therefore suffers from excessive false positive rates. PET, PEESE, and PET-PEESE suffer from poor efficiency and a tendency to either underestimate or overestimate effects across various conditions. Heterogeneity, which will likely always be present, complicates the interpretation and degrades the performance of *p*-curve and *p*-uniform.

We find that the three-parameter selection model is likely to provide the relatively best performance. However, being the relatively best obviously does not imply good performance in absolute terms, and in some conditions, even the relatively best performance was not acceptable. However, we believe that the application of the three-parameter selection model can improve practice and interpretation in meta-analysis through the correction of bias. Application of the three-parameter selection model is preferable to the absence of correction or the application of other methods. Regardless, the greatest trust should be reserved for the results of multi-site preregistered studies and our field should continue its efforts to improve the primary literature.

References

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012, November). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. doi:10.1177/1745691612459060

Bishop, D. V. & Thompson, P. A. (2016). Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ*, *4*, e1715.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2011, August). *Introduction to meta-analysis*. John Wiley & Sons.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015, March). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. doi:10.1037/a0038047

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333–342.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, *25*(24), 4279–4292.

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: self-control does not seem to rely on a limited resource. American Psychological Association.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, *1*(1), 4–17.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Duval, S. & Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463.

Fabrigar, L. R. & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68–80.

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904.

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, 201618569.

Ferguson, C. J. & Heene, M. (2012). A vast graveyard of undead theories: publication bias and psychological sci-

ence's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561.

Fraley, R. C. & Vazire, S. (2014). The n-pact factor: evaluating the quality of empirical journals with respect to sample size and statistical power. *PLOS One*, *9*(10), e109019.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1.

Guan, M. & Vandekerckhove, J. (2015, July). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-015-0868-6

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., … Bruyneel, S., et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573.

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. & 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.

Higgins, J. P. T. (2008). Commentary: heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, *37*(5), 1158–1160.

Hilgard, J., Engelhardt, C. R., & Rouder, J. N. (n.d.). Overstated evidence for short-term effects of violent games on affect and behavior: a reanalysis of anderson et al. (2010). *Psychological Bulletin*.

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

Lee, K. J. & Thompson, S. G. (2008). Flexible parametric models for random-effects distributions. *Statistics in medicine*, *27*(3), 418–434.

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.

Moreno, S. G., Sutton, A. J., Ades, A., Stanley, T., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, *9*(1), 2.

Motyl, M., Demos, A., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., … et al. (2017, April). *The state of social and personality science: rotten to the core, not so bad, getting better, or getting worse?* Retrieved from https://papers.ssrn.com/abstract=2959799

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003, December). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331–363. doi:10.1037/1089-2680.7.4.331

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2006). *Publication bias in meta-analysis: prevention, assessment and adjustments*. John Wiley & Sons.

Rücker, G., Carpenter, J. R., & Schwarzer, G. (2011). Detecting and adjusting for small-study effects in meta-analysis. *Biometrical Journal*, *53*(2), 351–368.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014, September). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *9*(5), 552–555. doi:10.1177/1745691614543974

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.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: descriptive and inferential statistics on all reasonable specifications.

Stanley, D. J. & Spence, J. R. (2014). Expectations for replications: are yours realistic? *Perspectives on Psychological Science*, *9*(3), 305–318.

Stanley, T. & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*(1), 108–112.

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016, September). Conducting meta-analyses based 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 Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293.

van Elk, M., Matzke, D., Gronau, Q., Guan, M., Vandekerckhove, J., & Wagenmakers, E.-J. (2015). Meta-analyses are no substitute for registered replications: a skeptical perspective on religious priming. *Frontiers in Psychology*, *6*, 1365. doi:10.3389/fpsyg.2015.01365

van Erp, S., 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*. Retrieved from osf.io/myu9c

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.