General Article

Advances in Methods and Practices in Psychological Science
2019, Vol. 2(2) 115–144

DOI: 10.1177/2515245919847196
www.psychologicalscience.org/AMPPS

Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods

Evan C. Carter[1], Felix D. Schönbrodt[2], Will M. Gervais[3], and Joseph Hilgard[4]

[1]Human Research and Engineering Directorate, U.S. Army Research Laboratory, Aberdeen, Maryland; [2]Department of Psychology, Ludwig-Maximilians-Universität München; [3]Department of Psychology, University of Kentucky; and [4]Department of Psychology, Illinois State University

Abstract

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 correct for such overestimation. However, it is not clear which methods work best for data typically seen in psychology. Here, we present a comprehensive simulation study in which we examined how some of the most promising meta-analytic methods perform on data that might realistically be produced by research in psychology. We simulated several levels of questionable research practices, publication bias, and heterogeneity, and used study sample sizes empirically derived from the literature. Our results clearly indicated that no single meta-analytic method consistently outperformed all the others. Therefore, we recommend that meta-analysts in psychology focus on sensitivity analyses—that is, report on a variety of methods, consider the conditions under which these methods fail (as indicated by simulation studies such as ours), and then report how conclusions might change depending on which conditions are most plausible. Moreover, given the dependence of meta-analytic methods on untestable assumptions, we strongly recommend that researchers in psychology continue their efforts to improve the primary literature and conduct large-scale, preregistered replications. We provide detailed results and simulation code at https://osf.io/rf3ys and interactive figures at http://www.shinyapps.org/apps/metaExplorer/.

Keywords

meta-analysis, publication bias, *p*-hacking, questionable research practices, bias correction, open data, open materials

Received 5/26/17; Revision accepted 3/18/19

Statistical techniques for analyzing the results from a set of studies in aggregate—often called meta-analysis—are popular in psychology and many other scientific disciplines because they provide high-powered tests, the ability to examine moderators across studies, and precise effect-size estimates that are useful for planning future studies and making policy decisions. However, just as bias can make the results from individual studies completely misleading (e.g., Simmons, Nelson, & Simonsohn, 2011), it can do the same to meta-analytic results. To address this problem, researchers have developed statistical techniques designed to identify and correct for bias. In this article, we present a neutral comparison (Boulesteix, Wilson, & Hapfelmeier, 2017) of how several promising methods perform when applied to simulated data that could have plausibly been produced by research in psychology. Our goal is to help researchers in psychology know what to expect from different methods when conducting meta-analyses in the face of bias.

Meta-Analysis

Meta-analytic techniques involve synthesizing a set of results from studies investigating the same empirical

Corresponding Author:
Evan C. Carter, Human Research and Engineering Directorate, U.S. Army Research Laboratory, Aberdeen Proving Ground, Aberdeen, MD 21005
E-mail: evan.c.carter@gmail.com

Box 1. Glossary

δ. Under the fixed-effect model, δ is the hypothetical true underlying effect estimated by each study. Under the random-effects model, δ is the mean of the distribution of hypothetical true underlying effects.

***d**_i*, ***v**_i*. The observed effect size (d) and its associated variance (v) for the i^{th} study. We calculate d as $\frac{M_1 - M_2}{S}$, where M_1 and M_2 are the means of the two groups and S is the pooled standard error of the two groups. The variance of d can be calculated as $\frac{n_1 + n_2}{n_2 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)} \times \frac{n_1 + n_2}{n_2 n_2 - 2}$, where n_1 and n_2 are the sample sizes of the two groups.

τ. The standard deviation of the distribution of hypothetical true underlying effects assumed by the random-effects model. It is also referred to as measuring between-study heterogeneity.

k. The number of studies in a meta-analytic sample.

Mean error. The average of the deviations of the estimates from the true effect (often called bias). Nonzero mean error indicates that the expected value of the estimate does not converge on the true value in the long run—that the estimate is too high or too low. Mean error is not sensitive to variance in estimates, so it is possible for a method to produce low mean error as a result of large but equal over- and underestimation. Such a case would yield estimates that are accurate on average, but any individual estimate could be quite far from the truth.

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

95% coverage probability. The percentage of 95% confidence intervals that include the true value of δ. Optimally, a method's coverage probability is at the nominal level of 95%. Low coverage is problematic, as it means that most confidence intervals do not contain the true value. Coverage rates higher than 95% may indicate exceedingly wide confidence intervals.

phenomenon (Borenstein, Hedges, Higgins, & Rothstein, 2011). Most often, the results from the individual studies take the form of effect-size estimates, and because meta-analyses are usually applied to studies with dependent variables measured on different scales, effect-size estimates are typically standardized. The typical goal of a meta-analysis is to produce a single summary estimate of the hypothetical true underlying effect, δ, estimated by each effect size in the data set. This approach is usually called fixed-effect meta-analysis (Cooper, Hedges, & Valentine, 2009) and can be modeled as $d_i = \delta + e_i$, where d_i is the observed effect size for study i, which 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. (See Box 1 for a glossary of the statistical symbols and terms used in this article.)

Another common model, known as random-effects (RE) meta-analysis (Cooper et al., 2009), holds that each study provides an estimate, d_i, of a different, related true effect, T_i—that is, $d_i = T_i + e_i$. 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, the operationalization of the independent variable or the population sampled in the particular study. In this model, the study-specific true effect, T_i, is calculated as $\delta + u_i$, where δ is the mean of the true effects 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 RE model to an observed set of studies provides an estimate of the average true underlying effect, δ, and the amount of between-study heterogeneity, τ^2. In this article, we use RE meta-analysis as our baseline for "uncorrected" meta-analysis. It should be noted, however, that determining which uncorrected estimator for the average true underlying effect to use is an active area of study itself (Baker & Jackson, 2013; T. Rice, Higgins, & Lumley, 2017; Schmid, 2017; T. D. Stanley & Doucouliagos, 2015; Veroniki et al., 2016).

Bias

The effects being estimated by meta-analysis can be systematically over- or underestimated in the face of bias, which is caused by factors that affect the analysis and reporting of the individual studies in the meta-analytic data set. We considered two primary sources of meta-analytic bias in our simulation study: *publication bias* and *questionable research practices* (QRPs).

Publication bias occurs 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, reports of statistically

nonsignificant or negative estimates of that effect may never be submitted for publication or may be rejected by reviewers and editors (Ferguson & Heene, 2012; Greenwald, 1975; Rothstein et al., 2006; Sterling, Rosenbaum, & Weinkam, 1995). In other words, statistically nonsignificant results, or results that contradict accepted theory, are left in the "file drawer." Because 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 than nonpublished data, publication bias can result in a meta-analytic sample that overrepresents studies yielding statistically significant, theory-consistent results. This can result in misleading meta-analytic findings, such as inflated estimates of the average true underlying effect. And although we do not focus on heterogeneity here, it is important to note that such bias also affects estimates of heterogeneity in complex, nonlinear ways (e.g., Augusteijn, van Aert, & van Assen, 2019; Jackson, 2007).

A related but independent form of bias is the use of QRPs (also referred to as the undisclosed use of researcher degrees of freedom or *p*-hacking). QRPs are said to occur when researchers favor a specific analytic approach (e.g., removing outliers or covariates) from the variety of potential approaches on the basis of the results that it yields. Such choices may be justifiable, yet simultaneously arbitrary and motivated (Simonsohn, Simmons, & Nelson, 2016). As is the case with publication bias, QRPs can result in overestimates of the true effect, as analyses that yield significant results are highlighted and analyses that do not yield such results are censored. We note that all bias-correcting methods that we applied in our study were designed to address publication bias, not QRPs.

Simulation Studies of Bias Correction in Meta-Analysis

Many simulation studies have been conducted to compare the performance of methods that correct for bias in meta-analysis (e.g., Hedges & Vevea, 1996; McShane, Böckenholt, & Hansen, 2016; Moreno et al., 2009; Rücker, Carpenter, & Schwarzer, 2011; Simonsohn, Nelson, & Simmons, 2014; T. D. Stanley, 2017; T. D. Stanley & Doucouliagos, 2014; van Aert, Wicherts, & van Assen, 2016; van Assen, van Aert, & Wicherts, 2015). However, there is very little overlap among these studies in either the methods they have examined or the simulated conditions they have explored. Different simulation studies have implemented bias differently, have drawn sample sizes from different distributions, and have varied widely in the value and form of the simulated true underlying effects. This lack of overlap is not surprising given that there is an effectively infinite number of possible combinations of different conditions to explore and no way of determining which conditions actually underlie real-world data. In other words, not only is there an inherent dimensionality problem in these simulation studies, but there is also no ground truth. These problems are often not discussed in reports of simulation studies, and indeed, many of the reports just cited—explicitly or implicitly—recommended the use of a single method, despite the fact that each study examined performance of only a handful of correction methods in only a limited subset of possible conditions.

In this article, we do not identify a single method that we believe should be used in all situations. Instead, we aim to add to the existing literature by (a) exploring a further set of conditions that may plausibly represent real data from research in psychology; (b) comparing a larger set of meta-analytic methods that, to our knowledge, have yet to be directly compared; and (c) discussing how our results can facilitate sensitivity analysis in meta-analysis.[1]

Disclosures

R (R Core Team, 2016) scripts for our analyses and simulation are available at the Open Science Framework (https://osf.io/rf3ys). Furthermore, we have made available interactive figures and tables that allow a detailed exploration of the results (http://www.shinyapps.org/apps/metaExplorer/). Supplemental material, which includes a comprehensive presentation of our results, is also available at the Open Science Framework (https://osf.io/rf3ys). We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Method

Simulation

We simulated the number of meta-analyzed studies, *k*, as one of four values (10, 30, 60, 100). All simulated individual studies had a two-group experimental design, so all effect sizes took the form of a standardized mean difference, Cohen's *d*. Notably, there is reason to think that this may be the most commonly used effect-size measure in psychology (see Table S1 in Fanelli, Costas, & Ioannidis, 2017). Cohen's *d* is an estimate of the true underlying effect, δ, which we chose to simulate as taking one of four values (0, 0.2, 0.5, 0.8), corresponding to the null hypothesis and Cohen's rule-of-thumb values for small, medium, and large effects, respectively.

Heterogeneity. As mentioned, variation in the true underlying effect, δ, is described by the heterogeneity

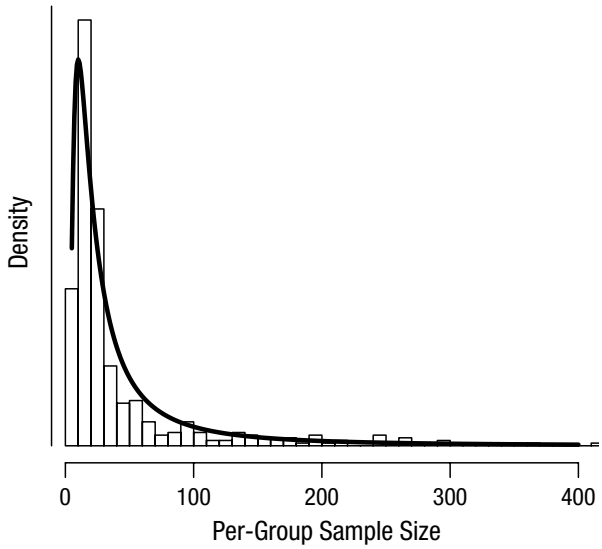

Fig. 1. Comparison of the empirical per-group sample-size distribution (histogram) from Marszalek, Barber, Kohlhart, and Holmes (2011) with the best-fitting inverse gamma curve (continuous line). The *x*-axis has been truncated at $n = 400$ for better visibility. This figure is available at https://osf.io/av285/, under a CC-BY4.0 license.

parameter, τ. We simulated three values for τ (0, 0.2, 0.4)[2] that may plausibly represent research in psychology: In an analysis of 187 meta-analyses that used standardized mean differences and were published in *Psychological Bulletin* from 1990 through 2013 (van Erp, Verhagen, Grasman, & Wagenmakers, 2017), 50% of all estimates of τ were smaller than 0.2, and 80% were smaller than 0.4.[3]

Study-level data. Independent samples were randomly generated for the control and experimental groups; observations in the control group were drawn from a normal distribution with a mean of 0 and standard deviation of 1, and observations in the experimental group were drawn from a normal distribution with a mean of T_i and standard deviation of 1. T_i was defined as the sum of δ and u_i, where u_i was drawn from a normal distribution with a mean of 0 and standard deviation of τ. Note that T_i, therefore, represented a study-specific true effect that varied randomly if τ was greater than 0. Cohen's *d* and the associated variance, *v*, were calculated for each simulated study, and a two-tailed independent-samples *t* test was applied to generate a *t* value and a *p* value.

Simulated sample sizes were based on an empirical distribution (Marszalek, 2011; Marszalek, Barber, Kohlhart, & Holmes, 2011) of per-group sample sizes from 1,225 studies published in 1995 and 2006 in four journals (*Journal of Abnormal Psychology, Journal of Applied Psychology, Journal of Experimental Psychology: Human Perception and Performance*, and *Developmental Psychology*). After we removed sample sizes smaller than 5, the strongly right-skewed per-group sample-size distribution had a median of 23 (25% quantile: 14, 75% quantile: 50). We found that an inverse gamma distribution (compared with negative binomial, log-normal, gamma, and Weibull distributions) clearly showed the best fit, according to the log likelihood. To sample per-group sample sizes in our simulations, we used a truncated inverse gamma distribution (truncated at $n = 5$ and $n = 1{,}905$, the latter being the largest observed per-group sample size in Marszalek et al.'s data set). The distribution had a shape of 1.153 and scale of 0.046. Figure 1 compares the empirical per-group sample-size distribution with the best-fitting curve.

Publication bias. For the simulation of publication bias, we used two censoring functions ("medium publication bias" and "high publication bias") that mapped a probability that a study was published to the study's one-tailed *p*-value. If the effect was in the "correct" direction, both functions returned a 100% probability of publication when $p_{one\text{-}tailed}$ was less than .025. Studies with "marginally significant" effects, $.025 \leq p_{one\text{-}tailed} < .05$, were published with an exponentially decreasing probability that reached 20% in the medium-publication-bias condition and 5% in the strong-publication-bias condition at $p_{one\text{-}tailed} = .05$. The probability of publication then remained constant for all *p* values up to, but not including, .5. If the effect was in the "wrong" direction ($p_{one\text{-}tailed} \geq .5$), the probability of publication was constant at 5% in the medium-publication-bias condition and 0% in the strong-publication-bias condition for all values of $p_{one\text{-}tailed}$ up to, but not including, .995. The probability of publication then increased exponentially until, at $p_{one\text{-}tailed} = .9995$, it reached a constant level of 50% in the medium-publication-bias condition and 20% in the strong-publication-bias condition (see Fig. 2). In our simulation, a random Bernoulli draw using the probability computed by these censoring functions determined whether a study was published.[4] Studies were continually simulated until the target number of *k* studies had been reached. In the no-publication-bias condition, all studies were included regardless of the value of $p_{one\text{-}tailed}$.

To our knowledge, this specific implementation of publication bias is comparable to, but different from, the implementations used in previous simulation studies (e.g., Bayarri & DeGroot, 1991; Guan & Vandekerckhove, 2016). Our primary reason for choosing this approach was that we did not want our publication-bias functions to exactly match those assumed by any of the bias-correcting methods being examined (e.g., Iyengar & Greenhouse, 1988; McShane et al., 2016), as this might result in an overly optimistic assessment of those methods' performance (Simonsohn, Simmons, & Nelson, 2017). Furthermore, we wanted to test whether results obtained previously with more straightforward

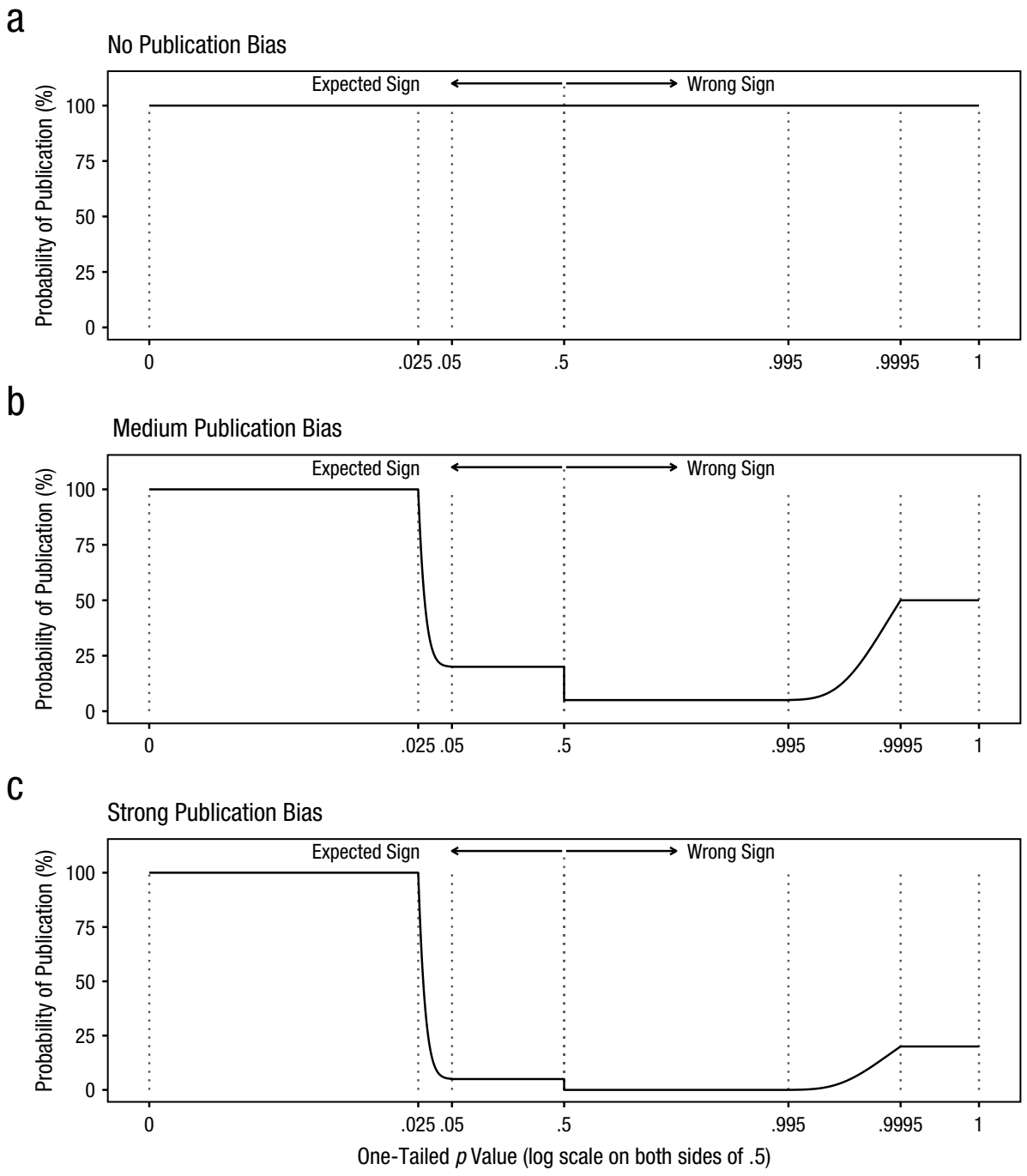

Fig. 2. Implementation of publication bias in the simulation. The graphs show the probability of publication as a function of the one-tailed p value of the simulated results in (a) the no-publication-bias condition, (b) the medium-publication-bias condition, and (c) the strong-publication-bias condition. The x-axes have a logarithmic scale on both sides of $p_{one\text{-}tailed} = .5$ to increase the visibility of the function at the high and low ends of the scale. This figure is available at https://osf.io/f6esc/, under a CC-BY4.0 license.

publication-bias functions would be robust to our more nuanced implementation.

Questionable research practices. We studied four forms of QRPs: (a) optional removal of outliers, (b) optional selection between two dependent variables, (c) optional use of moderators, and (d) optional stopping. Each data set that had QRPs applied to it was designed to simulate a study with a 2 (group: experimental vs. control) × 2 (level of the moderator: Level 1 vs. Level 2) design and

two dependent variables. Each dependent variable was measured across n observations. The moderator divided the simulated data set in half in a way that was independent of the dependent variable (i.e., the moderator had no main effect on the dependent variable) and the treatment (i.e., there was no collinearity between moderator and treatment). The Pearson's correlation between the two dependent variables was .20.

QRPs were applied so as to simulate the behavior of a researcher fishing for statistical significance. To test different levels of severity, we created three *individual QRP strategies* a simulated researcher could adopt: (a) pure (no use of QRPs), (b) moderate usage (use of optional dependent variables and the addition of 3 observations per cell for up to three data-collection efforts), and (c) strong usage (optional removal of outliers, use of optional dependent variables and optional moderators, and the addition of 3 observations per cell for up to five data-collection efforts).

In the case of the strong-usage strategy, the simulated researcher first tested the main effect of experimental manipulation on the first dependent variable. If this effect was not statistically significant and positive, the simulated researcher removed outliers (defined as observations with a z score greater than $|2|$).[5] If this second test was not positive and significant, the simulated researcher moved to the second dependent variable and repeated these steps. If no positive and significant effect was found, the researcher moved back to the first dependent variable and tested for an interaction effect between the experimental manipulation and the moderator. If there was a significant interaction, the researcher compared the experimental and control groups in only the subgroup defined by the first level of the moderator. This examination was conducted first with and then without outliers. If no positive and significant effect was found, the subgroup defined by the second level of the moderator was assessed in the same way. If a positive and significant effect was not found at the second level, the researcher moved to the second dependent variable and repeated the same procedure.

Additionally, simulated researchers could collect some additional amount of data (see two paragraphs earlier). After each additional collection effort, the QRPs just described were repeated. Thus, for each data-collection effort, simulated researchers could potentially apply 12 comparisons. If none of these analyses produced a positive and significant effect, the first test (comparison of the experimental group and the control group on the first dependent variable, with outliers untouched and no subgroups created with the moderator) was taken as the final result. The moderate-usage strategy represented a subset of the approach described in the previous paragraph, in combination with less additional data collection.

Given the sample sizes of our simulated primary studies, the moderate QRP strategy resulted in an inflated false-positive rate of 9% (computed in conditions without heterogeneity and publication bias, and counting only directionally consistent results), and the strong QRP strategy resulted in a false-positive rate of 27%. Note that more aggressive p-hacking beyond our strong setting is easily possible, for example, by examining even more dependent variables or excluding only directional outliers. Indeed, Simmons et al. (2011) reported a false positive rate of 61% produced by combining certain types of p-hacking.

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 characterized by different specific severities of QRP application, according to the strategies of individual researchers. In the *no-QRP* environment, 100% of the simulated researchers adopted the pure strategy. In the *medium-QRP* environment, 30% of the simulated researchers adopted the pure strategy, 50% adopted the moderate strategy, and 20% adopted the strong strategy; this mixture led to a false-positive rate of 11%. In the *high-QRP* environment, 10% of the simulated researchers adopted the pure strategy, 40% adopted the moderate strategy, and 50% adopted the strong strategy; this mixture led to a 17% false-positive rate.

Not all QRPs have the same distorting impact on a meta-analysis. Furthermore, some QRPs lead bias-correcting techniques to overestimate the true effect, whereas others lead to underestimation (van Aert et al., 2016). Our goal was not to investigate the differential impact of distinct QRPs on bias correction, but rather was to investigate some combinations of QRPs that may be plausible in real settings (John, Loewenstein, & Prelec, 2012). As there are infinite possibilities of how QRPs can be implemented, and an infinite number of ways in which these individual researcher strategies can be combined in QRP environments, our study is best considered a sensitivity analysis that explored the effect of a range of three plausible QRP environments on meta-analytic results. Our results do not necessarily generalize to other implementation of QRPs.

Design. To summarize, we simulated data for 432 unique combinations of five fully crossed factors (Table 1). We simulated 1,000 meta-analytic data sets for each of the 432 conditions. For a random selection of conditions, we also simulated 10,000 meta-analytic data sets and computed the Monte Carlo simulation error (Koehler, Brown, & Haneuse, 2009). These comparisons clearly

Table 1. Simulation Parameters

Experimental factor	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
Publication bias	None, medium, strong
Questionable-research-practices (QRPs) environment	No QRPs, medium use of QRPs, high use of QRPs

demonstrated that 1,000 simulated meta-analytic data sets lead to sufficiently stable estimates (see the supplemental material at https://osf.io/rf3ys).

Performance metrics

For each meta-analytic method, to test the hypothesis that the estimate provided differed 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 estimation performance of each method in terms of mean error, root mean squared error (RMSE), and 95% coverage probability (see Box 1).

Meta-analytic methods

We examined the performance of seven estimators: our baseline estimator, RE meta-analysis, and six estimators that adjust for bias. Further details on our specific implementations are available in the supplemental material at https://osf.io/rf3ys.

Random-effects meta-analysis. We applied the RE meta-analysis as described earlier using the *metafor* package in R (Viechtbauer, 2010). This approach makes no adjustment for publication bias or QRPs. We used the restricted maximum likelihood method for estimating between-study variance.

Trim-and-fill method. The trim-and-fill method (Duval & Tweedie, 2000) was introduced as a diagnostic test for publication bias and is based on the asymmetry of a funnel plot (a scatterplot showing effect-size estimates as a function of the standard error of those estimates). Publication bias introduces clear rightward asymmetry in a funnel plot (see the supplemental material) because nonsignificant and negative observations are censored. The trim-and-fill method involves iteratively removing (i.e., trimming) observations from one side of the funnel plot until a criterion for symmetry is met, and then "filling" these observations back into the funnel plot along with imputed observations that are identical to the trimmed observations but on the opposite side of the mean along the horizontal axis. Standard meta-analytic methods can then be applied to a data set including both observed and imputed studies.

Several previous simulation studies suggest that, although the trim-and-fill method can correct for bias in some cases, it tends to be outperformed by other methods and generally fails as heterogeneity increases (e.g., Idris & Ruzni, 2012; Moreno et al., 2009; Peters, Sutton, Jones, Abrams, & Rushton, 2007; Simonsohn, Nelson, & Simmons, 2014; Terrin, Schmid, Lau, & Olkin, 2003). For example, Terrin et al. (2003) examined the coverage probability of this method both with and without heterogeneity. They observed that coverage decreased as heterogeneity increased, primarily because the method imputed studies that were not missing when effect sizes from large studies (i.e., those near the top of the funnel plot) were far from the average overall effect. More recently, it has been suggested that in addition to unnecessarily correcting for bias, the trim-and-fill method does not correct enough for bias that does exist (Simonsohn, Nelson, & Simmons, 2014; Simonsohn, Simmons, & Nelson, 2014; van Assen et al., 2015).

Overall, no conclusion has been reached on the best way to implement the trim-and-fill method, as its performance can vary widely depending on the version of the algorithm and the conditions in which it is used (Moreno et al., 2009; Peters et al., 2007). Therefore, we used the default algorithm provided by the *metafor* package. Notably, results for this method did not always converge. Across all conditions, it returned a valid estimate in 95% of data sets. Nonconvergence happened mostly when k was at or above 60 and publication bias was strong.

Weighted average of the adequately powered studies. T. D. Stanley and Doucouliagos (2017) proposed the use of an intercept-only weighted-least-squares (WLS) metaregression estimator as a replacement for the naive fixed-effect and RE meta-analytic models. Simulation studies (T. D. Stanley, 2017; T. D. Stanley & Doucouliagos, 2017; T. D. Stanley, Doucouliagos, & Ioannidis, 2017) suggested that the WLS estimator performed on par with

the fixed-effect and RE models when the assumptions underlying those models were true, but outperformed both of them when the assumptions were violated (e.g., in the presence of publication bias).

Researchers have suggested extending this WLS estimator to reduce the impact of potential publication bias (Ioannidis, Stanley, & Doucouliagos, 2017; T. D. Stanley et al., 2017). In this extension, one first performs a WLS meta-analysis on all primary studies to obtain a (potentially biased) estimate of the true underlying effect. Then, one performs a second WLS meta-analysis on only those studies that had 80% statistical power to detect an effect of the size estimated by the first WLS meta-analysis, so as to obtain a weighted average of adequately powered (WAAP) studies. This approach is intended to avoid bias by discarding underpowered studies, which must overestimate the true effect to find statistical significance. If there are no adequately powered studies or only one adequately powered study in the data set, the WLS estimate for the entire data set is used. Thus, this conditional estimator, called WAAP-WLS, applies WAAP when there are at least two adequately powered studies and WLS otherwise.

Across all conditions, the WAAP-WLS method returned 77% WAAP and 23% WLS estimates. In small-*k*, small-δ conditions, there were not enough adequately powered studies, and 100% of estimates used WLS. In large-*k*, large-δ conditions, 100% of estimates used WAAP.

Previous simulation studies have suggested that the WAAP-WLS method is comparable to the WLS method, standard fixed-effects meta-analysis, and RE meta-analysis in the absence of heterogeneity and publication bias; however, as those conditions changed, the WAAP-WLS method has outperformed both WLS and standard meta-analysis (T. D. Stanley et al., 2017). The same simulation study suggested, however, that in terms of efficiency and overall bias, the WAAP-WLS method is outperformed by the precision-effect test/precision-effect estimate with standard error (PET-PEESE) method (described later in this section).

p-*curve.* A *p*-curve is the distribution of all statistically significant *p* values from the set of studies of interest (i.e., *p*s < .05; Simonsohn, Nelson, & Simmons, 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 (i.e., when the null is true), the distribution of significant, independent *p* values is uniform between .00 and .05. With increasing power, this distribution becomes increasingly right skewed. Because the degree of right skew is a function of the average study power, one can use the degree of right skew in a *p*-curve to (a) test the absence of a real effect and (b) estimate the average effect size corrected for publication bias.

Simonsohn, Nelson, and Simmons (2014) demonstrated that some typical QRPs cause the *p*-curve method to underestimate the true effect size. Later work by van Aert et al. (2016), however, suggested that bias can be upward or downward, depending on the specific type of QRPs. Additional work demonstrated that the *p*-curve method overestimates the average true underlying effect when there is heterogeneity (Simonsohn, Nelson, & Simmons, 2014; van Aert et al., 2016).

Simonsohn, Nelson, and Simmons (2014) interpreted the *p*-curve estimate as "the average effect size one expects to get if one were to rerun all studies included in the *p*-curve" (p. 667; see also Simmons, Nelson, & Simonsohn, 2018). In our view, however, meta-analysts generally aim to recover the average of the distribution of all true effects related to the phenomenon of interest (i.e., δ). Indeed, that is the purpose of the other estimators we examined. For consistency, therefore, we interpret *p*-curve results in the same fashion; however, in the supplemental material, we also assess this method's ability to recover the average true effect size of the studies submitted to it.

We implemented the *p*-curve method as recommended by Simonsohn, Nelson, and Simmons (2014), with the following settings. Only statistically significant and directionally consistent studies were submitted to the analysis. Any studies with significant but negative effects were discarded. Consequently, when no studies with significant positive effects were in a set, this method did not return an estimate (0.8% of all simulations). Across all conditions, the method returned an estimate in 99.2% of all simulated data sets. Not surprisingly, the method failed to produce an estimate almost exclusively when the set (*k*) consisted of 10 studies, the true effect (δ) was 0, and there was no publication bias.

In some cases, the *p*-curve method can return an estimate with a negative sign even though all included studies yielded effects with positive signs. It was our understanding that one should interpret only nonnegative effect-size estimates from the *p*-curve method, because a negative estimate based on a series of *p* values just below .05 is likely to indicate that the null hypothesis is true and there has been intensive *p*-hacking, rather than that there is a true effect in the opposite direction. Negative effect-size estimates obtained with the *p*-curve method were set to zero in our study (see recommendations from van Aert et al., 2016).[6]

In testing for the presence of an effect, we relied on the test for evidential value (i.e., the test for right skewness) for the full *p*-curve (Simonsohn, Simmons, & Nelson, 2015). This test is conceptually—but not

statistically—equivalent to a test for $\mu > 0$. Furthermore, *p*-curve estimation does not provide confidence intervals, so we could not assess this aspect of estimation for this method.

p-*uniform*. As does the *p*-curve method, the *p*-uniform method considers only the statistically significant results. It is based on the idea that the distribution of *p* values is uniform conditional on the population effect size (van Assen et al., 2015). Hence, it focuses on the *p*-value distribution under the alternative hypothesis, and it yields a fixed-effects estimate of the true effect by finding the value d^* that makes the conditional distribution of *p* values as uniform as possible.

The *p*-uniform method provides a hypothesis test, an estimate of the bias-corrected effect size, and a confidence interval around that estimate. Computationally, the *p*-curve and *p*-uniform methods differ only in that they use different implementations of the estimation algorithm, so in general they are expected to have similar strengths and weaknesses (McShane et al., 2016). For the computation of the *p*-uniform estimate, we used the Irwin-Hall estimator as implemented in the *puniform* package (van Aert, 2017) and recommended by van Aert et al. (2016). We also followed van Aert et al.'s recommendation to set the estimate to zero if the average of all significant *p* values was larger than .025, because the average *p* value is lower than .025 when there is a true positive effect.[7]

As is the case for the *p*-curve method, the *p*-uniform method does not return an estimate if there are no studies with significant positive effects (0.8% of all simulations). Across all conditions, the *p*-uniform method returned an estimate in 99.2% of all simulated data sets; in 10.4% of all simulations, the estimate was replaced by zero.

PET, PEESE, and PET-PEESE. The precision-effect test (PET; T. D. Stanley & Doucouliagos, 2014) is a metaregression approach to adjusting for small-study effects (see the supplemental material; see also the closely related Egger's test for publication bias—Egger, Smith, Schneider, & Minder, 1997). Small-study effects are said to exist when the observed effect size gets larger as the standard error grows (i.e., as the sample size shrinks). One cause of this pattern is publication bias, although other benign causes also exist (T. D. Stanley & Doucouliagos, 2014; Sterne, Gavaghan, & Egger, 2000).

The PET method uses a weighted-least-squares regression in which 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 *i*th effect-size estimate, d_i, and its associated standard error, se_i. The regression model is weighted by the inverse of the variances (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; it is an estimate of the true underlying effect that has been corrected for publication bas and other small-study effects. Of course, if small-study effects have many benign causes, there may be substantial overcorrection.

A closely related approach computes what is called the precision-effect estimate with standard error (PEESE; T. D. Stanley & Doucouliagos, 2014). In this method, a quadratic relationship between effect size and standard error is fitted to the data. The rationale is that if 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 are well powered and routinely publishable without such overestimation. Thus, publication bias (and the observed small-study effect) is expected to be stronger when the standard error is larger. A quadratic relationship can model such differences in bias across standard errors. The PEESE method uses a weighted-least-squares regression model, in which effect size is regressed on the square of the standard error: $d_i = b_0 + b_1 se_i^2 + e_i$. As in the PET method, 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.

Simulation studies suggest that the PET method outperforms the PEESE method when the true underlying effect is zero, whereas the PEESE method outperforms the PET method when the true underlying effect is nonzero (T. D. Stanley & Doucouliagos, 2014). In an attempt to offset the opposite biases of these methods, T. D. Stanley and Doucouliagos (2014) proposed the conditional PET-PEESE estimator. In this approach the statistical significance of the PET estimate is used to decide whether the PET or the PEESE estimate is taken as the final estimate. When the PET estimate is statistically nonsignificant (i.e., the estimated true effect is not distinguishable from zero) in a one-tailed test with $\alpha = .05$, the PET estimate is taken as the PET-PEESE estimate. In contrast, when the PET estimate is statistically significant, the PEESE estimate is used as the PET-PEESE estimate. For brevity, we focus only on describing the performance of the conditional PET-PEESE estimator, but in our discussion of sensitivity analysis, we recommend using all three methods: PET, PEESE, and PET-PEESE.

Although initial simulation results indicated that the PET-PEESE estimator's performance was promising (T. D. Stanley & Doucouliagos, 2014), two later

simulations revealed some weaknesses. In one, the standard RE meta-analysis estimator outperformed the PET and PEESE estimators in some ways; for example, it provided greater estimation efficiency (lower mean squared error) when heterogeneity was present (Reed, 2015). The other simulation showed unacceptable performance of the PET-PEESE estimator under conditions that seem common in psychology—a small number of studies, small samples across all studies, and high heterogeneity (T. D. Stanley, 2017).

Selection model. Selection-model approaches to mitigation of bias in meta-analysis have been in use for some time (Hedges, 1984; Hedges & Vevea, 1996; Iyengar & Greenhouse, 1988). We employed the three-parameter selection model (3PSM) as developed by Iyengar and Greenhouse (1988) and recently discussed by McShane et al. (2016). This model's three parameters represent the average true underlying effect, δ; the heterogeneity of the random effect sizes, τ^2; and the probability p_1 that a nonsignificant effect enters the literature. The last parameter, p_1, is modeled by a step function with a single cut point at $p = .025$ (one-tailed), which corresponds to a two-tailed p value of .05. This cut point divides the range of possible p values into two bins: significant and nonsignificant. The three parameters are estimated using maximum likelihood.

We implemented 3PSM using the default function in the *weightr* package (Coburn & Vevea, 2017). If no p value is present in one of the bins, the probability p_1 cannot be estimated. In this case, the *weightr* package uses a plug-in value of .01, which makes it possible to estimate the model (Vevea & Woods, 2005). However, even with this plug-in value, some models could not be estimated because of nonconvergence. Across all conditions, the 3PSM method returned an estimate in 91.5% of all simulated data sets. Estimation failed mostly when there were 100 studies in the set, the true effect (δ) was 0.2, and there was at least medium publication bias or a medium QRP environment (or both).

Several simulation studies of selection models have been conducted previously (e.g., Hedges & Vevea, 1996; Terrin et al., 2003). However, to our knowledge, only one examined the specific method we implemented: McShane et al. (2016) compared the 3PSM method with the p-uniform and p-curve methods, both of which can be understood as single-parameter selection models (i.e., only δ is estimated, publication bias is set to 100%, and heterogeneity is set to 0). In that study, the 3PSM approach clearly provided the best estimation and hypothesis testing (a) when δ was no greater than 0.30 and τ exceeded 0 and (b) when incomplete bias allowed some nonsignificant studies to be published.

Results

We simulated 1,000 data sets under 432 unique conditions (Table 1) and analyzed each with seven different meta-analytic methods. Here, we avoid an exhaustive presentation of the results and provide instead a more focused report. However, all of our findings, including information on convergence probabilities and exact values for mean error, RMSE, and coverage probabilities for all conditions, are available in Table 2 in the supplemental material (https://osf.io/rf3ys). We also have made several interactive figures available (http://www.shinyapps.org/apps/metaExplorer/) so researchers can explore combinations of conditions that they find to be particularly relevant to their own work.

In the following sections, we provide figures only for conditions in which $\delta = 0$ (i.e., the null hypothesis is true) or $\delta = 0.5$ (i.e., the alternative hypothesis that the effect size is 0.5 is true). The figures display the effects of increasing heterogeneity (from $\tau = 0$ to $\tau = 0.4$) and of increasing numbers of studies (from $k = 10$ to $k = 100$). Figure 3 shows both the false-positive rates (when $\delta = 0$) and the statistical power (when $\delta = 0.50$) of each method. In Figure 4, rather than providing exact values for mean error, RMSE, and coverage, we display the distributions of mean effect-size estimates with 95% quantile ranges. In our view, these values are more intuitive, and they allow a visual assessment of the observed mean errors (the means) and variability (the ranges).

In our summary of results, we first discuss the no-publication-bias and strong-publication-bias conditions, focusing on cases where $\tau = 0$ ("no heterogeneity") and $\tau = 0.2$ ("addition of heterogeneity") and where $k = 10$ and $k = 60$. The influence of QRPs is discussed separately at the end of this section.

Some methods did not always return an estimate. In some cases, this was intended (i.e., when no studies with significant, directionally consistent effects were available for the p-curve and p-uniform estimators); in other cases, the method failed to produce an estimate, for example, because of nonconvergence. We report the summaries of all computations that did return an estimate, but readers should be aware that this implies a conditional interpretation: The reported mean errors, RMSEs, and error rates are conditional on a result being provided. If a method performed well when it returned an estimate, but did not return an estimate the majority of the time, this should be taken into consideration when comparing that method with others. Figures 3 and 4 indicate when a method did not return an estimate in more than 25%, 50%, or 75% of the 1,000 simulation runs. RE meta-analysis and the PET-PEESE and WAAP-WLS methods always returned an estimate. Table 2

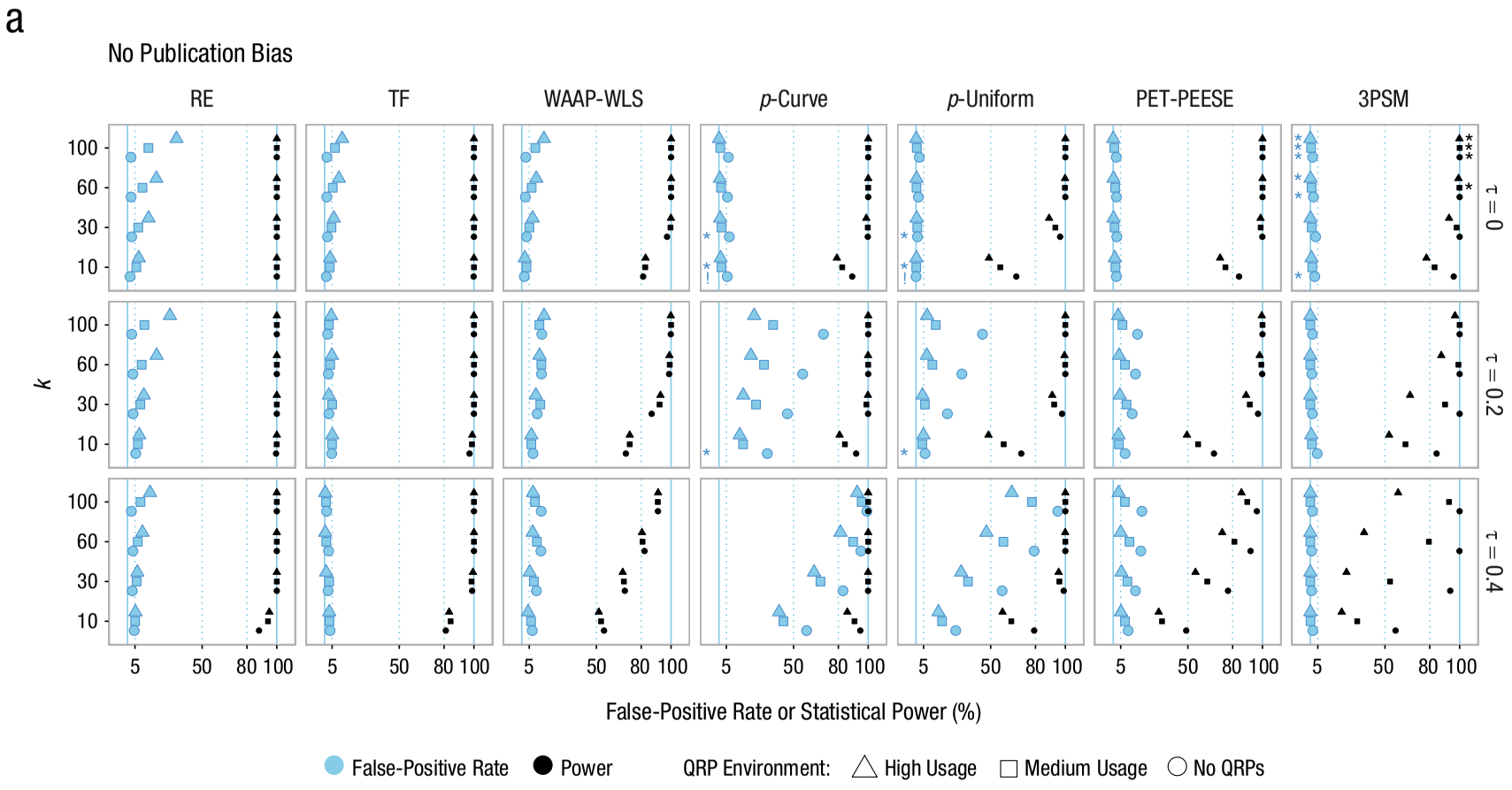

Fig. 3. *(continued on next page)*

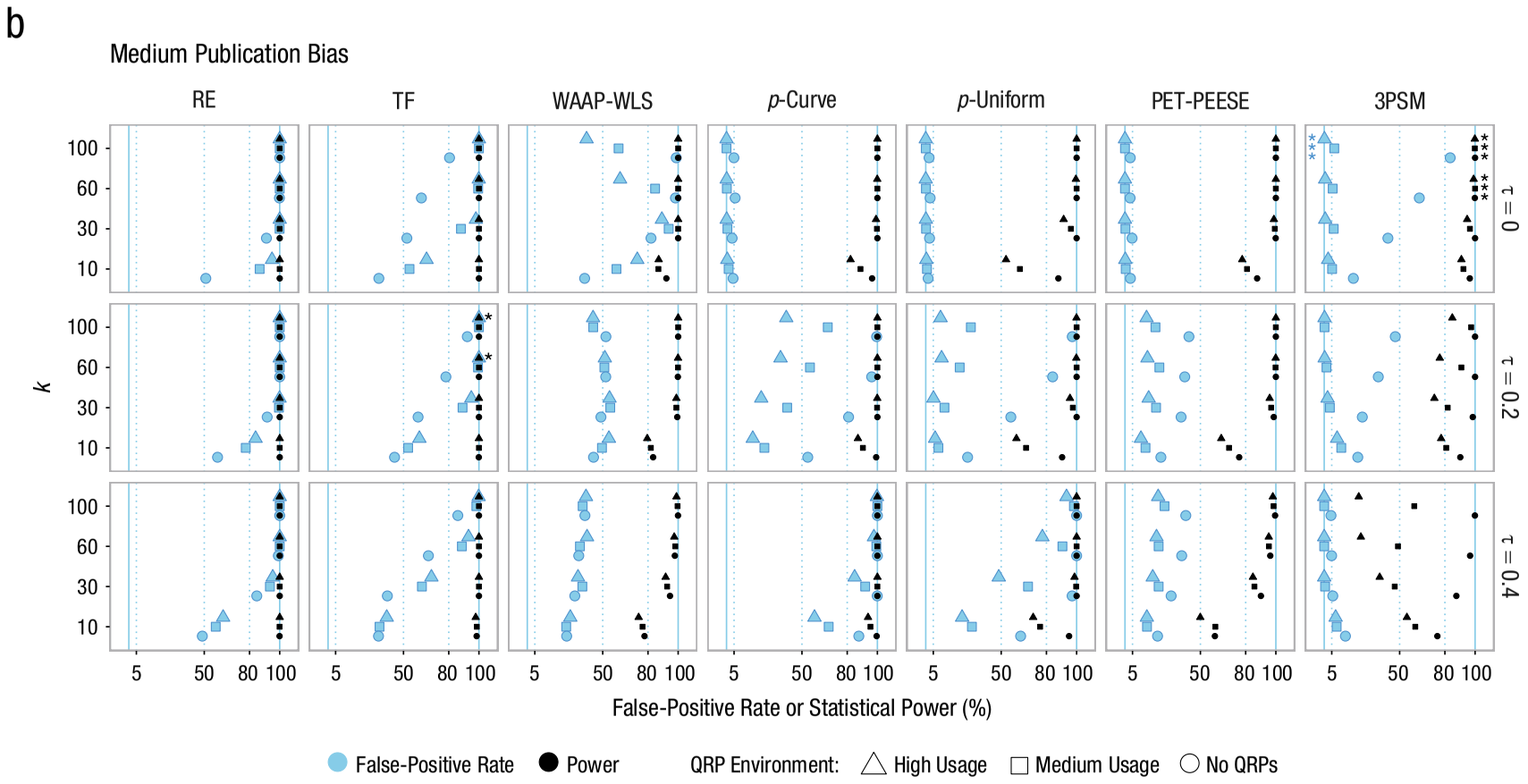

Fig. 3. *(continued on next page)*

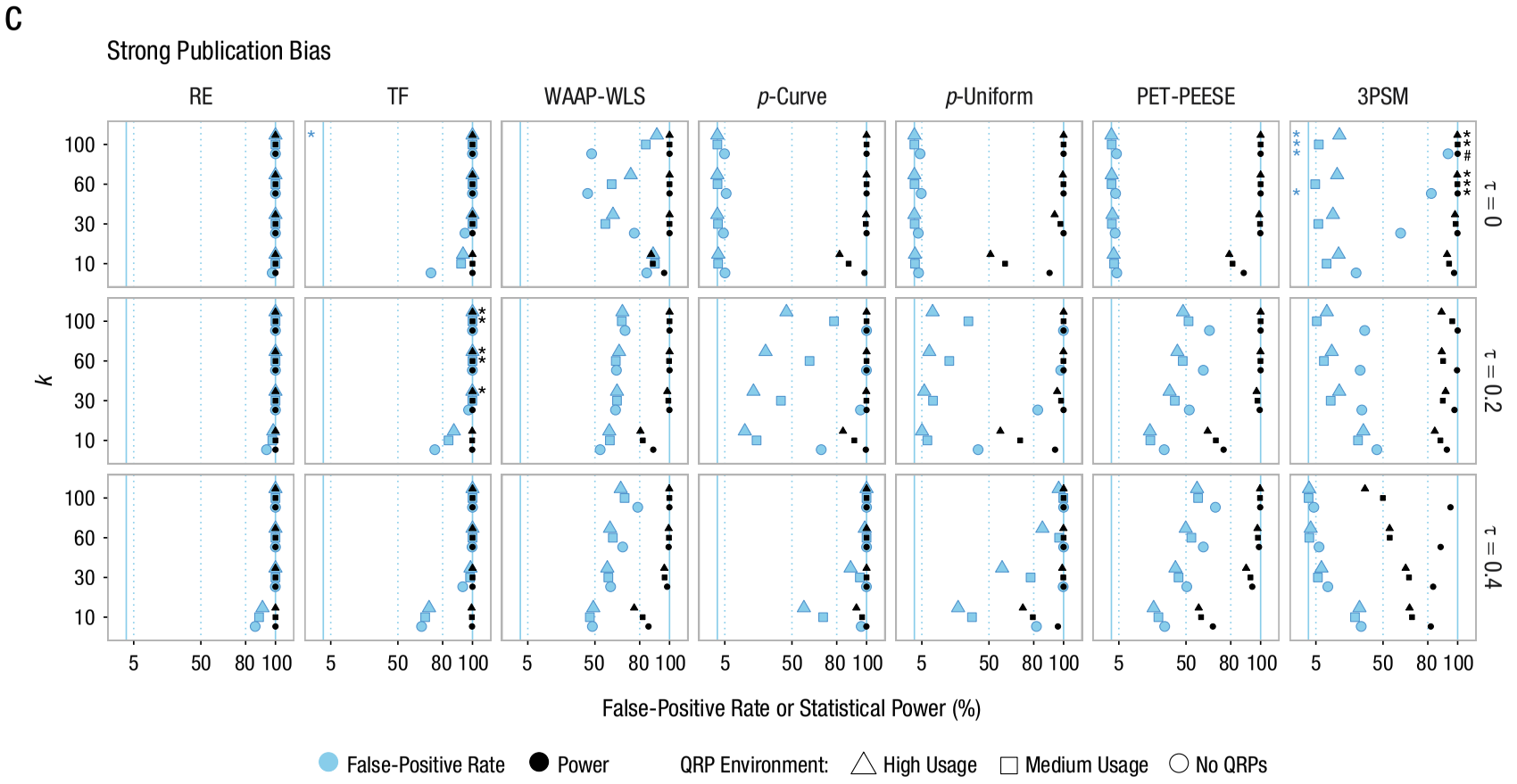

Fig. 3. False-positive rates (when δ = 0) and statistical power (when δ = 0.5) for all seven methods in (a) the no-publication-bias condition, (b) the medium-publication-bias condition, and (c) the strong-publication-bias condition. In each graph, results are shown separately for the four meta-analytic sample sizes (*k*) and the three questionable-research-practices (QRP) environments. Results for the three levels of between-study heterogeneity (τ) are in separate graphs. RE = random-effects meta-analysis; TF = trim-and-fill method; WAAP-WLS = weighted average of adequately powered studies/weighted-least-squares estimator; PET-PEESE = precision-effect test/precision-effect estimate with standard error; 3PSM = three-parameter selection model. Symbols on the left and right borders of the graphs indicate when a method computationally did not return a result in a substantial proportion of the 1,000 simulations: < 750 (*), < 500 (#), < 250 (!) successful computations. This figure is available at https://osf.io/379p2/, under a CC-BY4.0 license.

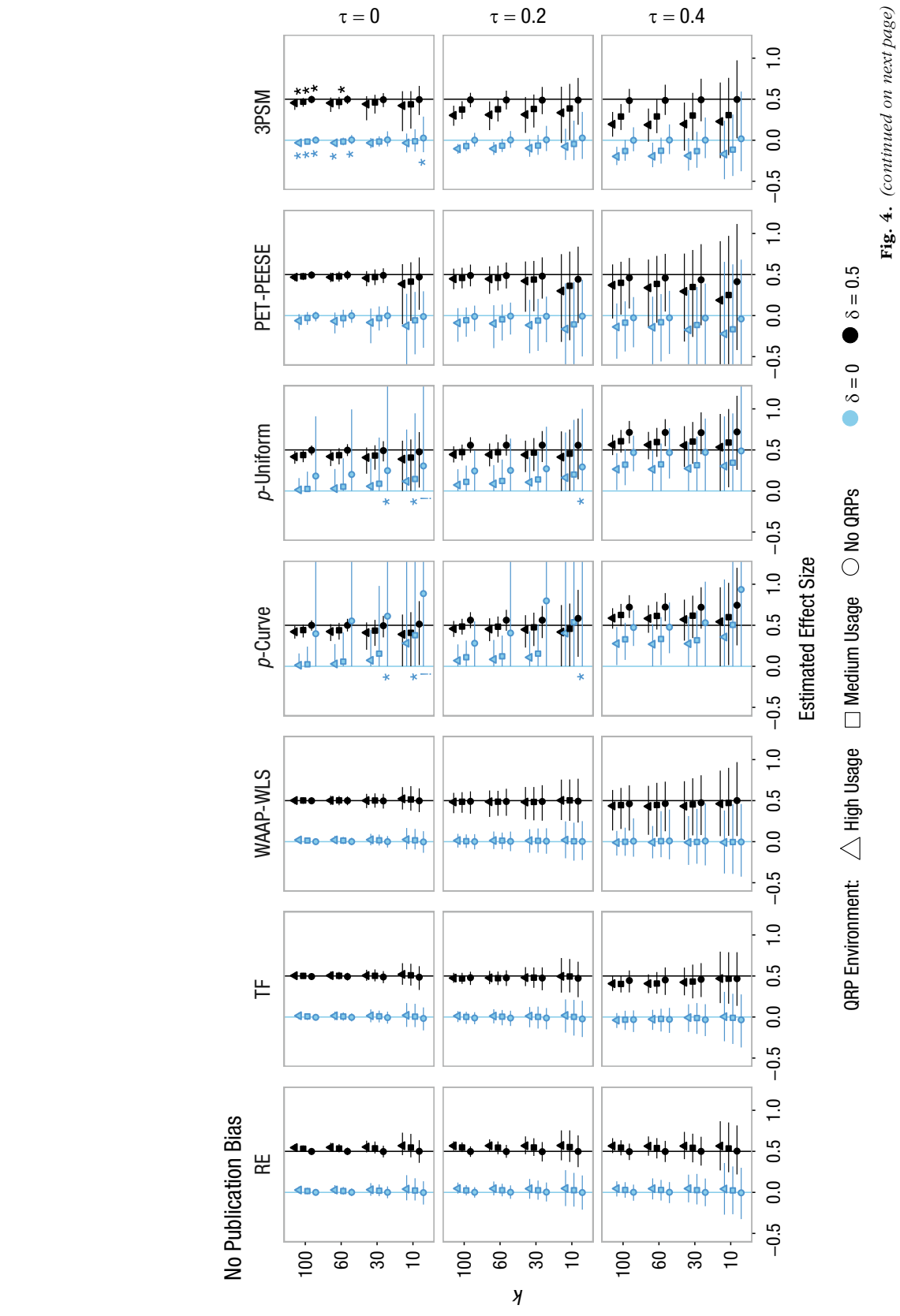

Fig. 4. *(continued on next page)*

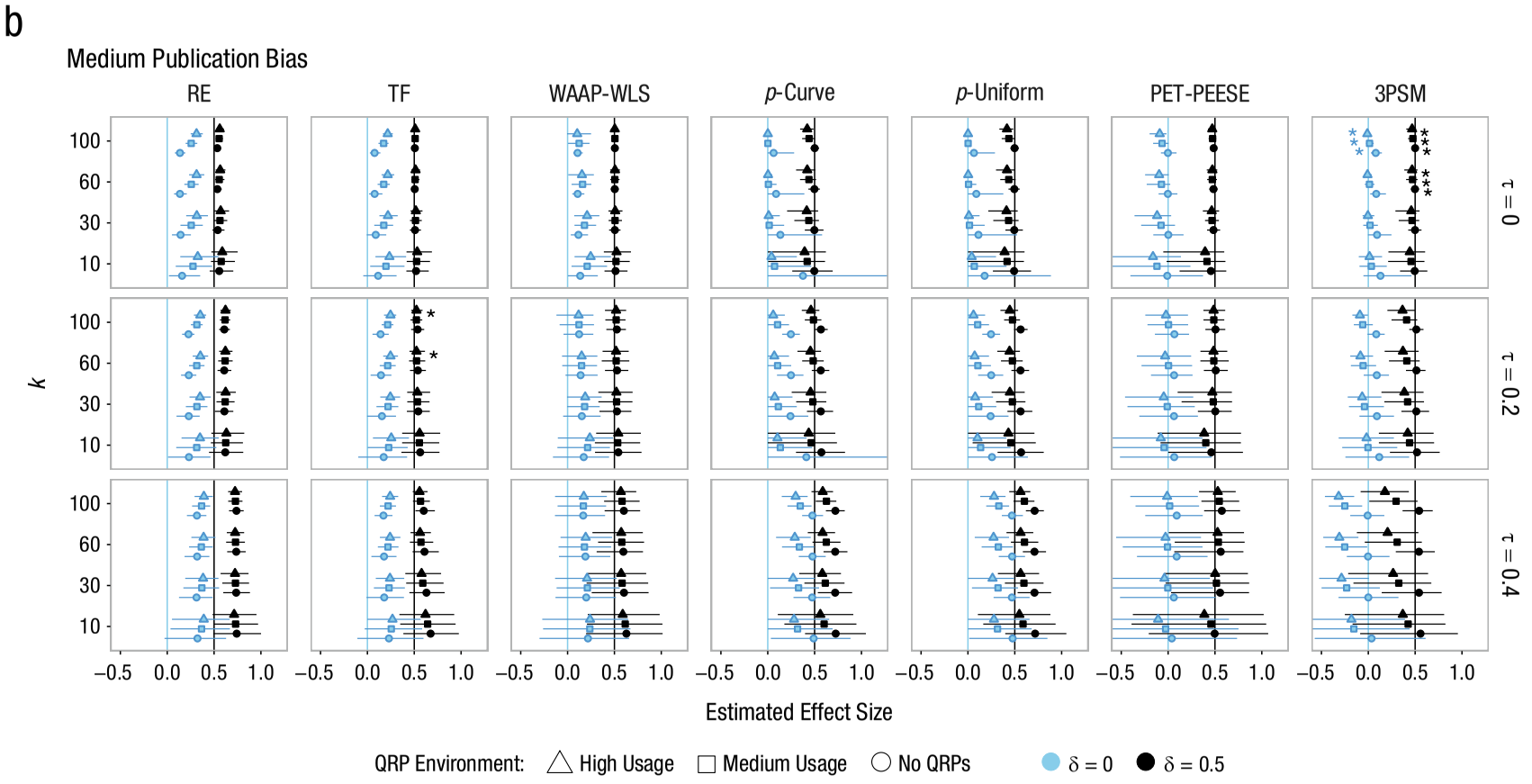

Fig. 4. *(continued on next page)*

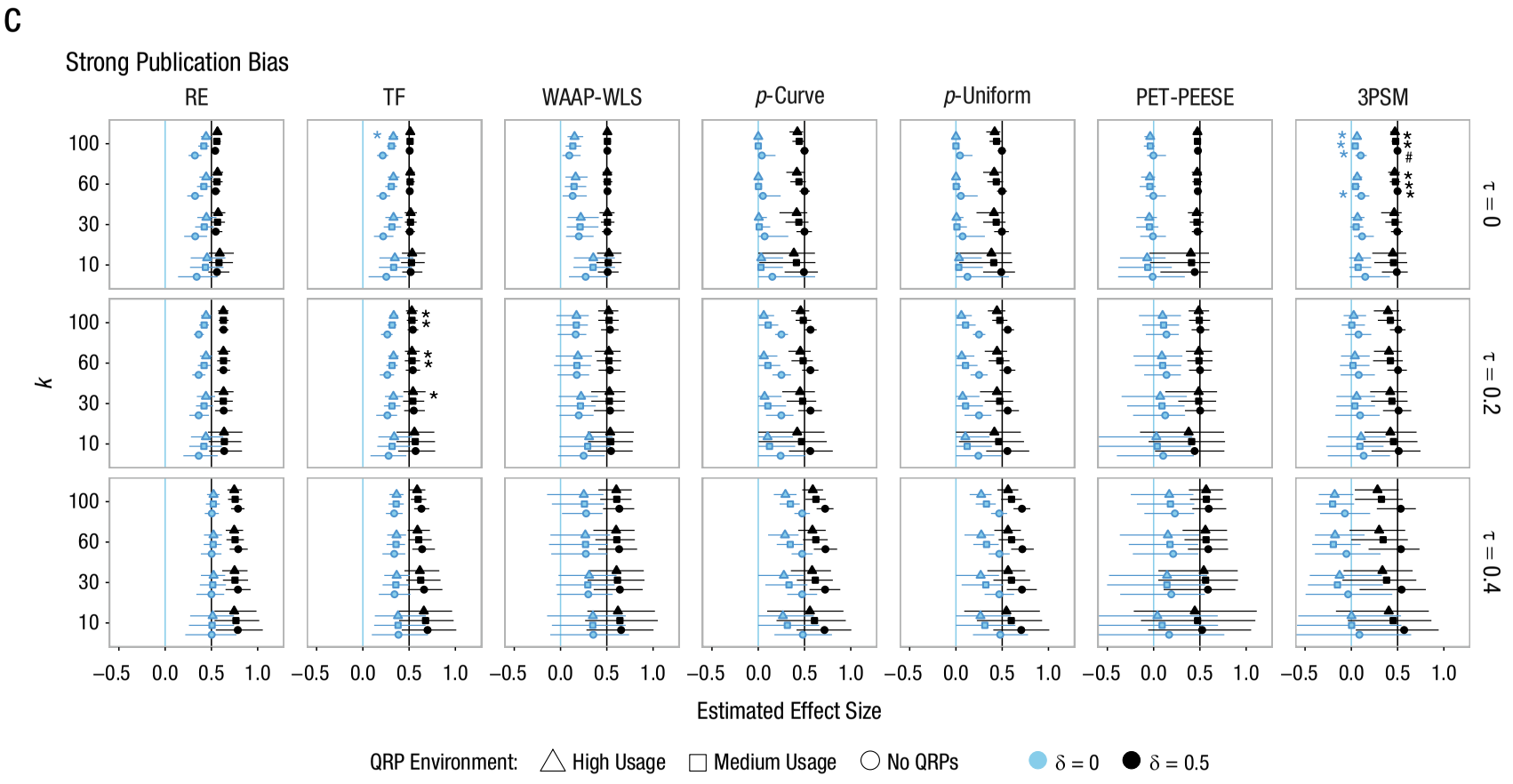

Fig. 4. Means (points) and inner 95% quantile ranges (whiskers) of the effect-size estimate distributions for all seven methods when δ = 0 and when δ = 0.5 in (a) the no-publication-bias condition, (b) the medium-publication-bias condition, and (c) the strong-publication-bias condition. In each graph, results are shown separately for the four meta-analytic sample sizes (*k*) and the three questionable-research-practices (QRP) environments. Results for the three levels of between-study heterogeneity (τ) are in separate graphs. RE = random-effects meta-analysis; TF = trim-and-fill method; WAAP-WLS = weighted average of adequately powered studies/weighted-least-squares estimator; PET-PEESE = precision-effect test/precision-effect estimate with standard error; 3PSM = three-parameter selection model. Symbols to the left and right of the whiskers indicate when a method computationally did not return an estimate in a substantial proportion of the 1,000 simulations: < 750 (*), < 500 (#), < 250 (!) successful computations. This figure is available at https://osf.io/hjzfq/, under a CC-BY4.0 license.

in the supplemental material reports the exact percentages of returned estimates for each of the other four methods in each condition.

No publication bias, no QRPs

Type I error rate. When the null hypothesis was true and there was no heterogeneity, most methods had appropriate Type I error rates, although the error rates for the 3PSM and *p*-uniform methods were more conservative (0–3%) than the nominal alpha rate. The *p*-curve and *p*-uniform methods sometimes failed to provide an estimate because of the scarcity of statistically significant results, especially at $k = 10$.

The addition of heterogeneity led to slightly higher Type I error rates (< 10%) for RE meta-analysis and the trim-and-fill method. Type I error rates rose moderately (10–30%) for the WAAP-WLS and PET-PEESE methods and increased considerably for the *p*-curve and *p*-uniform methods (8–56%). Adding heterogeneity did not increase Type I error rates for the 3PSM method, which remained excessively conservative.

Power. With a sample size (k) of 10 and no heterogeneity, RE meta-analysis offered the best power, followed closely by the trim-and-fill and 3PSM methods. Power was poorer for the *p*-curve, PET-PEESE, and WAAP-WLS estimators, and markedly poorer for the *p*-uniform estimator. Despite this, all the methods achieved 100% power at $k = 60$.

The addition of heterogeneity at $k = 10$ slightly increased the power of the *p*-curve and *p*-uniform methods. The power for RE meta-analysis was unchanged, and the power of the trim-and-fill method was slightly reduced. The power of the 3PSM, WAAP-WLS, and PET-PEESE methods fell by more, 5 to 15 percentage points. Despite this, all the methods had greater than 95% power at $k = 60$.

Mean error. The methods were generally unbiased, with some exceptions. With a sample size (k) of 10 and no heterogeneity, the trim-and-fill and PET-PEESE methods demonstrated slight downward bias (about –0.02); these downward biases were mitigated at $k = 60$. The *p*-curve and *p*-uniform methods both exhibited upward bias when the null hypothesis was true. For the *p*-uniform method, the mean estimate was 0.30 at $k = 10$ and 0.20 at $k = 60$. For the *p*-curve method, the mean estimate was 0.89 at $k = 10$ and 0.55 at $k = 60$. However, both methods were unbiased when the null was false. The upward bias when the null was true was likely caused by these methods' truncation at zero—there were no negative underestimates to cancel out positive overestimates.

Adding heterogeneity had little effect on the bias of RE meta-analysis and the WAAP-WLS and 3PSM methods. It added some upward bias to the *p*-curve and *p*-uniform estimators and a very slight downward bias to the trim-and-fill and PET-PEESE estimators.

RMSE. In the absence of heterogeneity, RE meta-analysis provided the most efficient results and the smallest RMSE. RMSE was slightly greater for the trim-and-fill and WAAP-WLS methods, moderately greater for the 3PSM and PET-PEESE methods, and noticeably greater for the *p*-uniform and *p*-curve methods. RMSE was particularly large for the *p*-curve and *p*-uniform methods when the null was true, a pattern that is consistent with their upward bias in that condition, as well as their use of only statistically significant results, of which there are fewer when the null is true. The 3PSM method also had relatively high RMSE when the null was true and k was 10.

Adding heterogeneity increased the RMSE of RE meta-analysis only slightly. The increases were slightly larger for the trim-and-fill, WAAP-WLS, 3PSM, and PET-PEESE methods. Heterogeneity tended to slightly increase the RMSE of the *p*-curve and *p*-uniform methods, presumably by causing bias. However, when the null was true or k was small (i.e., $k = 10$), heterogeneity improved the RMSE of the *p*-curve and *p*-uniform methods, presumably by increasing the number of significant results that were drawn upon.

Coverage of 95% confidence intervals. In the absence of heterogeneity, coverage rates for RE meta-analysis and the PET-PEESE and trim-and-fill methods were ideal at 95%. The WAAP-WLS, *p*-uniform, and 3PSM estimators also had approximately correct coverage rates (±2%). (Recall that the *p*-curve method does not give a confidence interval.)

Adding heterogeneity caused slightly poorer coverage for all the methods, particularly *p*-uniform. For RE meta-analysis and the 3PSM method, coverage rates recovered somewhat at $k = 60$. For the PET-PEESE, *p*-uniform, and WAAP-WLS methods, coverage rates grew worse as k increased from 10 to 60.

Strong publication bias, no QRPs

In the face of strong publication bias, sets of meta-analyzed results often consisted of only significant results, especially at $k = 10$.

Type I error rate. In the absence of heterogeneity, RE meta-analysis suffered from false-positive rates of 98% and higher. The trim-and-fill method had slightly lower, but still unacceptable, Type I error rates in excess of 70%.

The WAAP-WLS method had poor Type I error rates at $k = 10$ (> 85%), but its Type I error rates decreased with increasing k (45% at $k = 60$), which was not the case for RE meta-analysis and the trim-and-fill method. The 3PSM method had lower Type I error rates (31%) at $k = 10$, but these error rates increased with increasing k (82% at $k = 60$). The *p*-curve, *p*-uniform, and PET-PEESE estimators had approximately conservative Type I error rates ranging from 3% to 10%.

The addition of heterogeneity slightly reduced the Type I error rate for RE meta-analysis, but error rates still approached 100% with increasing k. The trim-and-fill method was generally not affected. For the WAAP-WLS method, heterogeneity reduced Type I error at $k = 10$ but increased Type I error at $k = 60$; error rates were still quite high for all sample sizes (50+%). Heterogeneity substantially increased the Type I error rates of the *p*-curve and *p*-uniform methods, leading to error rates of 40+% at $k = 10$ and 98+% at $k = 60$. Heterogeneity also substantially increased the Type I error rate of the PET-PEESE estimator (by 33 percentage points or more). Heterogeneity increased the 3PSM method's Type I error rates at $k = 10$ (+14 percentage points) but reduced the Type I error rate substantially at $k = 60$ (−48 percentage points).

Power. In the absence of heterogeneity, RE meta-analysis and the trim-and-fill estimator had approximately 100% power at all levels of k. Power was slightly lower but still good (89+%), even at $k = 10$, for the other methods: PET-PEESE, *p*-uniform, WAAP-WLS, 3PSM, and *p*-curve (in order of ascending power). Note that "good power" in this case was accompanied by high Type I error; that is, the null was frequently rejected incorrectly because the methods overestimated the true underlying effect.

Adding heterogeneity had little influence on power. At $k = 10$, it reduced the power of the PET-PEESE estimator by 13 percentage points, the power of the WAAP-WLS estimator by 7 percentage points, and the power of the 3PSM estimator by 5 percentage points. It slightly increased the power of the *p*-curve (+1 percentage point) and *p*-uniform (+4 percentage points) methods. Power of all the methods approximated 100% at $k = 60$.

Mean error. When there was no heterogeneity and the null hypothesis was true, many methods were badly biased upward. RE meta-analysis estimated the null effect as about 0.33. The trim-and-fill method did not substantially reduce this bias, estimating the effect as about 0.22. The WAAP-WLS estimates were quite biased at $k = 10$ (0.27) and became less biased at $k = 60$ (0.13). At $k = 10$, the *p*-curve, *p*-uniform, and 3PSM methods overestimated the null effect, yielding estimates of approximately 0.15, but their overestimation was smaller than that of the other methods just mentioned, and the bias decreased at $k = 60$ (estimates ranging from 0.05 to 0.11). The *p*-curve and *p*-uniform methods showed the strongest benefits of increasing the sample size. The PET-PEESE method was unbiased.

When the null was false, RE meta-analysis still slightly overestimated the true effect (by ~0.05). The trim-and-fill and WAAP-WLS methods both reduced this small bias, yielding estimates that were only very slightly biased. The *p*-curve, *p*-uniform, and 3PSM methods were unbiased in estimating the true effect. The PET-PEESE method tended to slightly underestimate the true effect, although this tendency was ameliorated with increasing k (the effect was underestimated by 0.06 at $k = 10$ and by 0.02 at $k = 60$).

Adding heterogeneity to a true effect of zero tended to increase upward bias. RE meta-analysis and the trim-and-fill, WAAP-WLS, and PET-PEESE methods all demonstrated slight increases in upward bias under heterogeneity. The *p*-curve and *p*-uniform methods demonstrated substantial upward bias (~0.25) under heterogeneity. In contrast, adding heterogeneity did not increase the bias of the 3PSM estimator.

Adding heterogeneity to a true nonzero effect also tended to move estimates slightly upward. The upward bias of RE meta-analysis and the trim-and-fill and WAAP-WLS estimators was slightly increased; the downward bias of the PET-PEESE method was slightly reduced; and the *p*-curve and *p*-uniform methods showed slight upward bias (~0.06). The 3PSM method remained unbiased.

RMSE. In the absence of heterogeneity, when the null was true, RMSE was considerably elevated for RE meta-analysis. All six adjustment methods led to some improvement in RMSE, with the exception of the *p*-curve method at $k = 10$. The improvements were modest for the WAAP-WLS and trim-and-fill methods and successively greater for the *p*-uniform, *p*-curve, 3PSM, and PET-PEESE methods. These improvements with the latter four methods were particularly pronounced at $k = 60$.

When there was a true effect, the RMSE of RE meta-analysis was acceptable (0.08 at $k = 10$, 0.05 at $k = 60$). Most of the other methods again provided some improvement. At $k = 10$, the *p*-uniform and *p*-curve methods did not improve the RMSE, and the PET-PEESE method increased the RMSE. The 3PSM, WAAP-WLS, and trim-and-fill estimators provided modest improvements in the RMSE. At $k = 60$, all six adjustment methods yielded improvements in the RMSE; the trim-and-fill, WAAP-WLS, and 3PSM methods were slightly more efficient than the *p*-uniform, *p*-curve, and PET-PEESE methods.

Adding heterogeneity to a true effect of zero caused modest increases in RMSE for all the methods, and all the adjustment methods provided some benefit relative to RE meta-analysis. The greatest improvement was provided by the 3PSM method, followed in order by the PET-PEESE, *p*-uniform, *p*-curve, and trim-and-fill methods. The WAAP-WLS estimator improved as k increased, performing worse than the *p*-curve and *p*-uniform methods at $k = 10$ but better than those methods at $k = 60$.

Adding heterogeneity to a true nonzero effect caused moderate increases in RMSE for all the methods. Again, all the adjustment methods provided some benefit relative to RE meta-analysis, and these benefits were comparable across methods; the one exception was that the PET-PEESE estimator caused an increase in the RMSE at $k = 10$.

Coverage of 95% confidence intervals. Because of the considerable publication bias, when the null hypothesis was true and there was no heterogeneity, 95% coverage was very poor (< 3%) without adjustment. All the adjustment methods led to some improvement in coverage. The benefits of the trim-and-fill method were slight (28% coverage at $k = 10$, 0% coverage at $k = 60$). The benefits of the WAAP-WLS method increased with k, but coverage did not improve beyond 55%. The benefits of the 3PSM method, on the other hand, decreased with k: Coverage was 68% at $k = 10$ but 18% at $k = 60$. The *p*-uniform and PET-PEESE estimators were the only ones to yield good coverage rates across all sample sizes (90–94%).

Given a true effect and no heterogeneity, the coverage of RE meta-analysis was much better. However, as k increased, coverage fell, presumably because of some combination of bias and insufficient width of the confidence interval (e.g., at $k = 10$, coverage was 86%, but at $k = 60$, coverage was only 42%). All the adjustment methods achieved approximately correct coverage rates across all sample sizes (90–97%), with the exception that the PET-PEESE method demonstrated undercoverage that grew worse with increasing k.

When the null hypothesis was true and heterogeneity was added, the coverage for RE meta-analysis was slightly better but still very poor. All the adjustment methods slightly improved coverage. The greatest benefits were provided by the 3PSM method (52% coverage at $k = 10$, 65% coverage at $k = 60$) and the PET-PEESE method (60% coverage at $k = 10$, 36% coverage at $k = 60$). However, these are still suboptimal coverage rates.

Given a true effect, the addition of heterogeneity substantially worsened the coverage for RE meta-analysis. All the adjustment methods somewhat improved coverage. At $k = 10$, the 3PSM, WAAP-WLS, and *p*-uniform methods brought the greatest improvements, reaching coverage of 78%, 81%, and 87%, respectively. At $k = 60$, the coverage of the *p*-uniform and WAAP-WLS methods fell noticeably (~68% coverage), whereas the coverage of the 3PSM method improved to 93%.

The influence of QRPs

Influence of QRPs on naive meta-analysis. QRPs generally led to an increase in bias in RE meta-analysis, provided that the null was true and there was some publication bias. This increase was greatest when there was medium publication bias and the null was true; under these circumstances, high use of QRPs doubled mean error from 0.15 to 0.32. When there was strong publication bias and the null was true, the effect of QRPs was smaller; high use of QRPs increased mean error from about 0.32 to 0.44. The damage was presumably smaller because strong publication bias had already caused considerable mean error. When there was a true effect of 0.5, or when there was no publication bias, the influence of QRPs on the mean error and RMSE was minimal.

QRPs also generally led to an increase in Type I error. With a sample size (k) of 10 and medium publication bias, QRPs approximately doubled the Type I error rate from 51% to 95%. However, at $k = 30$ or more, when there was at least medium publication bias, Type I error rates were generally at ceiling (90+%), so QRPs did little to further influence Type I error. In the absence of publication bias, QRPs did lead to noticeable increases in Type I error given large-enough k (i.e., 60 or 100). For example, when there was no true effect and no heterogeneity, high use of QRPs increased Type I error to 19%, even though the mean error remained a mere 0.03. QRPs similarly tended to harm the coverage of 95% confidence intervals.

In summary, for RE meta-analysis, QRPs exacerbated the effects of publication bias when there was no effect; however, the effects of QRPs on mean error were modest when (a) there was no publication bias, (b) publication bias was strong, or (c) there was a true effect. Thus, QRPs, as we implemented them, seem to play a small role in meta-analytic bias on their own. In the company of moderate publication bias, however, QRPs can considerably amplify problems.

Influence of QRPs on bias-corrected meta-analysis. The effect of QRPs in our simulation varied as a function of both meta-analytic method and performance metric. In this section, we focus chiefly on bias as the metric because the effects of QRPs on bias are the most straightforward and communicable. The effects of QRPs on RMSE, error rates, and coverage were generally a

function of whether QRPs caused an increase or decrease in bias: When QRPs reduced the absolute value of the mean error, the RMSE and coverage probability generally improved; when QRPs increased the absolute value of the mean error, the RMSE, coverage rates, and Type I and II error rates were accordingly poorer. In some cases, a curvilinear effect was observed; as QRPs increased, an initial positive bias was reduced and then became negative; thus, these metrics first improved and then deteriorated. The influence of QRPs was generally strongest when there was a null or small effect, presumably because studies with medium or large true effects required less *p*-hacking to be published.

The effects of the QRPs we modeled on performance of the trim-and-fill method were similar to their effects on performance of RE meta-analysis: Bias increased when the null was true and there was medium or strong publication bias. This bias also led to elevated Type I error rates (except when there was heterogeneity and no publication bias, in which case Type I error decreased slightly). The effects of the QRPs on performance of the WAAP-WLS method were similar, but increases in bias were smaller than those with the trim-and-fill method. A curious exception is that QRPs reduced Type I error when the WAAP-WLS method was used and there was medium publication bias and a large sample size ($k = 60$, $k = 100$), although even in these conditions the Type I error rates were still unacceptable (≥ 40%). Perhaps, in these cases, the WAAP-WLS method switched from the better-powered WLS test to the poorer-powered WAAP test.

In contrast, QRPs caused downward bias in the *p*-curve and *p*-uniform methods. In the context of homogeneity, in which these methods are typically unbiased, QRPs led to underestimation of the effect size and an increase in the Type II error rate. In the context of heterogeneity, in which these methods tend to overestimate the effect size, QRPs led to less overestimation of the effect size, a decrease in Type I error rates, and an increase in Type II error rates. We consider this pattern to reflect two simple effects of opposite sign: Heterogeneity caused upward bias in the mean error, and QRPs caused downward bias, so the absolute value of the mean error was smaller when both were present than when only one or the other was present. QRPs also helped to reduce the upward bias in the average *p*-curve and *p*-uniform estimates when the null was true, perhaps by increasing the number of significant studies available.

The QRPs nudged PET-PEESE estimates downward. When these estimates were biased upward in our simulation, as in the case of small or null effects in the context of publication bias and heterogeneity, QRPs reduced bias and improved Type I error rates slightly. When PET-PEESE estimates were unbiased or biased downward, as in the case of nonzero true effects, QRPs led to greater downward bias. This downward bias was sometimes quite strong when the null was true. The PET-PEESE method yielded statistically significant effects of opposite sign in many analyses; the Type I error rates tended to grow with increasing use of QRPs, increasing publication bias, and larger sample sizes, with rates ranging from 9% (medium use of QRPs, strong publication bias, $k = 10$) to 62% (high use of QRPs, medium publication bias, $k = 100$). Researchers have, at times, considered a significant and negative PET-PEESE estimate as evidence that the estimate is incorrect, choosing instead to interpret results from less extreme adjustments, such as the trim-and-fill method (see, e.g., Bediou et al., 2018). A statistically significant PET-PEESE estimate in the unexpected direction probably is incorrect, but researchers should be aware that when they obtain such an estimate, there is likely to be some combination of QRPs and publication bias and, perhaps, a null effect.

The QRPs we simulated generally led to a slight downward bias in 3PSM estimates. This bias was stronger when heterogeneity was present. At worst, given high heterogeneity ($\tau = 0.4$), the mean error caused by QRPs was as large as –0.32. QRPs therefore tended to reduce Type I error and increase Type II error for this method.

What was the worst that happened with each estimator as a consequence of the QRPs we implemented? In the case of RE meta-analysis and the trim-and-fill and WAAP-WLS methods, QRPs exacerbated the effects of publication bias. In the medium-publication-bias condition, QRPs increased the bias in these estimators; mean error increased from 0.14 to 0.31 for RE meta-analysis, from 0.08 to 0.22 for the trim-and-fill method, and from 0.11 to 0.15 for the WAAP-WLS method. These changes in bias caused corresponding increases in Type I error rates, increasing them by as much as 40 percentage points. When publication bias was strong, QRPs increased bias by as much as 0.12, but this increase was less dramatic than that caused by QRPs when there was medium publication bias. The PET-PEESE, *p*-curve, and *p*-uniform methods each demonstrated downward bias of up to –0.14 when QRPs were simulated. This bias caused a loss of power of up to 17 percentage points for the first two methods and 37 percentage points for the *p*-uniform method. High application of QRPs also caused the PET-PEESE method to frequently mistake null effects for significant negative effects (up to 62% of the cases with no heterogeneity, high use of QRPs, and medium publication bias). The 3PSM method underestimated the true effect of 0.5 by as much as 0.19, and power was reduced by up to 32 percentage points.

The reactions of the estimators to QRPs may be broadly considered to fall into two clusters. In the case of RE meta-analysis and the trim-and-fill and WAAP-WLS methods, QRPs caused overestimation, particularly of null effects. In the case of the PET-PEESE, *p*-curve, *p*-uniform, and 3PSM methods, QRPs caused underestimation of true effects and noticeable loss of power.

It is important to bear in mind that the results we have described are specific to the way we simulated QRPs. Our approach could likely be changed in a variety of ways that would provide very different results. We see this topic as an important area for future research.

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 QRP usage. We assessed the results according to both the ability to reject a null effect or detect a true effect and the ability to estimate the mean of the distribution of true underlying effects. We begin our discussion of the results with a coarse summary of the three overall patterns we observed, as well as some general recommendations.

First, RE meta-analysis and the trim-and-fill and WAAP-WLS methods showed alarmingly high false-positive rates (Fig. 3) and overestimation (Fig. 4) in the face of publication bias in combination with a zero or small true effect size. Generally, the WAAP-WLS method outperformed both RE meta-analysis and the trim-and-fill method.

Second, the *p*-curve and *p*-uniform methods had reasonable false-positive rates and little bias under homogeneity. With increasing heterogeneity, however, both showed increasing false-positive rates (Fig. 3) and overestimation (Fig. 4), particularly for a zero or small true effect size. This poor performance was actually mitigated by increasing use of QRPs, and was primarily independent of changes in publication bias and sample size. Again, we note that the original developers of the *p*-curve method have argued that its performance should not be evaluated using the average true underlying effect, as is the usual approach in the meta-analytic literature, but rather should be evaluated using the average of the effects submitted to the analysis (Simonsohn, Nelson, and Simmons, 2014). The performance of the *p*-curve method evaluated in this way is described in the supplemental material, but in general, it should be noted that this method performed well when estimating the average of the true underlying effects of the submitted studies. If no QRPs were present, this estimator recovered that quantity with very low mean error, regardless of the level of publication bias and the value of δ. QRPs as we modeled them induced a downward bias in *p*-curve estimates, particularly when δ was small and there was little or no heterogeneity. These results are consistent with previous simulation results (Simonsohn, Nelson, and Simmons, 2014).

Third, the PET-PEESE and 3PSM methods both showed mostly reasonable false-positive rates, but they suffered from reduced power when sample sizes were small, heterogeneity was high, there was little publication bias, or there was heavy use of QRPs (Fig. 3). These two methods also showed similar patterns of estimation error (Fig. 4): Both methods tended to underestimate effects when sample sizes were small, heterogeneity was high, or there was heavy use of QRPs. Although the two methods produced similar overall patterns of results, the 3PSM estimator almost always outperformed the PET-PEESE estimator.

Furthermore, it is worth noting that RE meta-analysis and the WAAP-WLS and PET-PEESE methods always returned at least some estimate, whereas the other methods sometimes failed to converge or could not be applied because of lack of significant studies. Information on the percentages of valid estimates is available in Table 2 in the supplemental material and should be considered alongside the performance information we have reported here. For example, it may be that consistent failures to converge in certain simulated conditions indicate that the method will be less applicable to real-world data than other methods are.

On the basis of our results, we believe we can confidently make five general recommendations:

1. If publication bias is highly unlikely (e.g., data are from a multilaboratory preregistered replication), rely on RE meta-analysis rather than any of the other methods we examined.
2. When there may be publication bias, do not rely on RE meta-analysis alone. Publication bias can quickly accumulate in even small sets of published studies, leading to overestimated effects and high Type I error rates.
3. Recognize that the popular trim-and-fill adjustment, although efficient, reduces bias and Type I error rates only slightly. To achieve stronger reductions in bias, consider the PET-PEESE, *p*-curve, *p*-uniform, and 3PSM adjustments. However, keep in mind that these adjustments are often inefficient and a given individual estimate may be poor, even if these adjustments are unbiased in the long run.
4. Do not use the *p*-curve or *p*-uniform methods for estimating δ if heterogeneity is expected or if many studies yielded nonsignificant results.

5. Given that adjustments for publication bias are only partially successful, we offer a final recommendation that must be implemented not by meta-analysts, but by primary researchers and journal editors: Take steps to ensure the completeness and transparency of the original literature. An ounce of registered report is worth a pound of bias correction.

Limits on generalizability

Simulation studies necessarily require making assumptions that might limit the generalizability of their results to real data. Although we simulated a data-generation process that might plausibly underlie real-world research in psychology, several limits to our study should be kept in mind when considering our findings. First, we simulated the data-generation process in the absence of QRPs as a two-group design, despite the fact that real research designs are rarely this simple. However, the vast majority of meta-analyses use effect-size measures, such as correlations and standardized mean differences, that ignore such design complexities (see, e.g., Table S1 in Fanelli et al., 2017). For example, to meta-analyze data from an experiment with a 2 × 2 factorial design, one would typically calculate a standardized mean difference by either discarding or collapsing across the factor that is not of primary interest. So for a 2 × 2 design with a per-group sample size of 20, the comparison entered into the meta-analysis would have either a total N of 40 (i.e., 2 × 20; when the second factor is discarded) or a total N of 80 (i.e., 2 × 2 × 20; when data are collapsed across the second factor). Therefore, although most designs are more complex than the two-group case we simulated, data in meta-analyses are often reduced to this simple form. As a result, our findings generalize to meta-analyses in which more complex designs are handled by discarding nonfocal factors. In the case of collapsing across these factors, our simulation likely underestimates sample size on average. However, it should be noted that the choice to collapse across factors is problematic given the required assumption that the nonfocal factors do not interact with the comparison of interest (i.e., there is a true interaction effect of 0). Thus, our findings generalize to the least problematic case.

A second, related issue arises for real-world data generated by single-sample designs (e.g., correlational studies). If such studies tend to have larger or smaller sample sizes than those with factorial designs, our simulation might under- or overestimate sample size, respectively. Critically, the generalizability issue here is related only to sample sizes, not to the fact that different study designs tend to be summarized with different effect-size measures. At the level of the study, one can translate between most effect-size measures without changing statistical significance or the direction of the effect. Given that bias acts at the study level through these two features, the generalizability of our results holds regardless of whether data originally take the form of standardized mean differences or correlations.

A third point to consider is whether our implementation of publication bias mirrors bias in real-world data. We implemented publication bias using specific functions with specific parameter values. Of course, it would be entirely possible to use different functions or different parameter values. What is unclear, however, is the degree to which different choices at this level would produce different results. For example, we intentionally modeled publication bias in a way that differed from the bias that the 3PSM, p-curve, and p-uniform methods are designed for, so it may be that these methods would show improved performance under different specifications of publication bias. Ultimately, this is an empirical question and should be the focus of future research. Additionally, our implementation of QRPs was extremely specific and might limit the generalizability of our results. Because the kinds of QRPs that can be applied depend entirely on the design of the specific study, there are an infinite number of possible ways to simulate QRPs (Hartgerink, van Aert, Nuijten, Wicherts, & Van Assen, 2016). Thus, our results for QRPs likely will not generalize to designs that are dramatically different from those we simulated.

Fourth, it is impossible to perfectly mirror how real data are generated. However, it is our hope that researchers can use our framework to close this gap and tackle some of the issues we have mentioned. It would be relatively easy, for example, to modify our code to use larger or smaller sample sizes and then assess whether this substantially changes how the methods perform.

Finally, bias correction in meta-analyses is an active field of research, and multiple new methods were published after our simulations were completed. These include an extension of the p-uniform method, called p-uniform*, that estimates heterogeneity and includes nonsignificant results (R. C. M. van Aert & van Assen, 2018) and a Bayesian fill-in method called BALM (Du, Liu, & Wang, 2017).

Method performance checks and sensitivity analysis for meta-analysis in psychology

Several authors have suggested that sensitivity analysis can be a valuable tool for evaluating the robustness of

conclusions from a meta-analysis (e.g., APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008; McShane et al., 2016; van Aert, 2017; van Aert et al., 2016). If results do not substantially change across a range of different methods and assumptions, the conclusions can be considered to be robust. However, the set of methods employed in a sensitivity analysis should include only those that can be expected to perform reasonably well. Put differently, if a method is known to perform poorly under the conditions that apply to a meta-analysis at hand, it should not be included in a sensitivity analysis, or it should at least be treated with skepticism and given less weight than other methods when the results are evaluated.

To encourage and facilitate sensitivity analysis in meta-analysis, we suggest a two-step procedure: The first step is to evaluate which bias-correcting methods can be expected a priori to perform reasonably well in research conditions that are plausible for the meta-analysis at hand (*method performance check*). The second step is to compute meta-analytic estimates using all the included methods and compare them in order to evaluate the variability (or robustness) of conclusions (*sensitivity analysis*). This evaluation should respect the results from the method performance check and weigh the methods accordingly.

For a sensible sensitivity analysis, we recommend that meta-analysts and consumers of meta-analyses focus on the following question: "Do my conclusions depend on a meta-analytic method that performs poorly in plausible conditions?" If the answer is "yes," then those conclusions should clearly be revisited. To help analysts and consumers answer this question, we have provided an interactive app (http://www.shinyapps.org/apps/metaExplorer/) that, for a given method and a given definition of "performs poorly," identifies all of the conditions in our simulation for which the answer to the question is "yes." In the following, we provide an illustration of how one might perform this form of method performance check and discuss how this step guides the subsequent sensitivity analysis.

A real-world example: two data sets on ego depletion. We use data from studies on the topic of ego depletion for this example because it is relatively easy to understand and there are data from both meta-analyses of the literature and a large-scale preregistered replication.

Briefly, the limited-strength model of self-control holds that any act of self-control will result in subsequent acts of self-control being less likely to succeed—a state referred to as ego depletion (Muraven, Tice, & Baumeister, 1998). Typically, experiments aimed at testing this hypothesis involve participants completing a sequence of two tasks—a manipulation task and an outcome task. Prior to the outcome task, participants in the depletion condition are given a version of the manipulation task that is designed to require more self-control than the version given to participants in the control condition. Support for ego depletion is claimed when participants in the depletion condition perform worse on the subsequent outcome task than participants in the control condition do. Following convention, we represent this effect as a standardized mean difference (*d*); higher values indicate a greater depletion effect. In the following example, we analyze both a preexisting meta-analytic sample of 116 ego-depletion studies (Carter, Kofler, Forster, & McCullough, 2015) and a data set of 23 studies from a Registered Replication Report (Hagger et al., 2016). We apply each of our methods to these two data sets (see Tables 2 and 3).

Imagine that researchers agree to the logic that a depletion effect (δ) less than or equal to 0.15 should be considered practically equivalent to zero (Carter & McCullough, 2018). Unfortunately, in this case, different meta-analytic methods would lead to different conclusions (see Table 2). On the basis of the results from the RE model and the WAAP-WLS, trim-and-fill, *p*-curve, *p*-uniform, and 3PSM methods, a group of researchers could conclude that the depletion effect is practically significant (i.e., $\delta > 0.15$). In contrast, on the basis of the results from the PET, PEESE, and PET-PEESE methods, a separate group of researchers could conclude that the depletion effect is practically nonsignificant (i.e., $\delta \leq 0.15$). Hence, a naive sensitivity analysis would be inconclusive, as the variability in results is so large that either conclusion can be drawn. To help overcome this inconclusiveness, we recommend that researchers ask, "Do my conclusions depend on a meta-analytic method that performs poorly in plausible conditions?" Such a method performance check gives guidance as to which results should be given more weight and credibility.

What are "plausible conditions"? For a method performance check, one must define "plausible conditions." As in an a priori power analysis, considerations should relate to the specific research environment at hand. Only if no specific prior knowledge is available can general knowledge about the field be used as an approximation. For example, some degree of bias seems possible: In the fields of psychology and psychiatry, more than 90% of all published hypothesis tests are significant (Fanelli, 2011), despite estimates that average power is around 35% (Bakker, van Dijk, & Wicherts, 2012), and whereas reported effects tend to be statistically significant, unreported effects tend not to be (Franco, Malhotra, & Simonovits, 2016). Moreover, there is both direct (Franco et al., 2016; LeBel et al., 2017) and self-report (John et al., 2012) evidence of

Table 2. Meta-Analytic Estimates and Method Performance Checks for Carter, Kofler, Forster, and McCullough's (2015) Data

Method	Estimate (with 95% confidence interval)	Poor performance[a]		
		δ = 0	δ = 0.20	δ = 0.50
Random-effects	0.43 [0.34, 0.52]	Yes	No	No
Trim-and-fill	0.24 [0.13, 0.34]	Yes	No	No
WAAP-WLS	0.35 [0.26, 0.43]	Yes	No	No
p-curve	0.55 [NA, NA]	Yes	No	No
p-uniform	0.55 [0.33, 0.71]	Yes	No	No
PET	–0.27 [–0.52, 0.00]	No	No	No
PEESE	0.00 [–0.14, 0.15]	Yes	No	No
PET-PEESE	–0.27 [–0.52, 0.00]	Yes	No	No
3PSM	0.33 [0.19, 0.47]	No	Yes	No

Note: WAAP-WLS = weighted average of adequately powered studies/weighted-least-squares estimator; PET = precision-effect test; PEESE = precision-effect estimate with standard error; 3PSM = three-parameter selection model; NA = not applicable.
[a]These columns report whether or not each model performed poorly in the plausible conditions of the simulation study when the true population effect (δ) had the indicated value.

the use of QRPs, and several studies have found evidence of small-study effects consistent with the presence of publication bias (Bakker et al., 2012; Fanelli et al., 2017; Kühberger, Fritz, & Scherndl, 2014).[8]

In addition to bias, a degree of heterogeneity seems very likely when diverse experimental paradigms are summarized in a meta-analysis (e.g., as opposed to a multilab registered report; T. D. Stanley, Carter, & Doucouliagos, 2018; Tackett, McShane, Bockenholt, & Gelman, 2017; van Erp et al., 2017). Finally, it seems that the typical true effect in psychology research is rather small: The median published effect size (*d*) is around 0.3 to 0.4 (Bosco, Aguinis, Singh, Field, & Pierce, 2015; Richard, Bond, & Stokes-Zoota, 2003), and as this estimate is not corrected for publication bias, the typical true effect is likely smaller. This general observation, of course, does not preclude the possibility that some effects in psychology are indeed large. As Hagger, Wood, Stiff, and Chatzisarantis's (2010) meta-analysis of ego-depletion research revealed an overall uncorrected *d* of 0.62, which most likely is inflated, we consider plausible conditions for the sensitivity analysis of

Table 3. Meta-Analytic Estimates and Method Performance Checks for Hagger et al.'s (2016) Data (Registered Replication Report)

Method	Estimate (with 95% confidence interval)	Poor performance[a]		
		δ = 0	δ = 0.20	δ = 0.50
Random-effects	0.04 [–0.06, 0.14]	No	No	No
Trim-and-fill	0.04 [–0.06, 0.14]	No	No	No
WAAP-WLS	0.04 [–0.08, 0.15]	No	No	No
p-curve	0.00 [NA, NA]	Yes	No	No
p-uniform	0.00 [NA, NA]	Yes	No	No
PET	–0.03 [–0.80, 0.74]	No	No	No
PEESE	0.00 [–0.41, 0.40]	No	No	No
PET-PEESE	–0.03 [–0.80, 0.74]	No	No	No
3PSM	0.01 [–0.10, 0.12]	No	No	No

Note: WAAP-WLS = weighted average of adequately powered studies/weighted-least-squares estimator; PET = precision-effect test; PEESE = precision-effect estimate with standard error; 3PSM = three-parameter selection model; NA = not applicable.
[a]These columns report whether or not each model performed poorly in the plausible conditions of the simulation study when the true population effect (δ) had the indicated value.

the data in Carter et al. (2015) to be those in which the true effect size (δ) is less than or equal to 0.5.

In summary, given the conditions we simulated, we define the most plausible conditions for this sensitivity analysis as those with medium to strong publication bias and QRPs, heterogeneity (τ) of 0.2 or more, and true effect sizes (δ) less than or equal to 0.5. We evaluate the performance of all the estimators we examined under these plausible conditions at $k = 100$, the simulated value closest to the observed k of 116.

What is "poor performance" of a meta-analytic method? Given this definition of plausible conditions, the next step for a method performance check is to identify defensible choices for the definition of "poor performance." For simplicity, in this example we consider only one metric, mean error,[9] and ask whether each method is likely to be biased enough that a true null effect is mistaken for a practically significant effect or, conversely, that a true effect is mistaken for a practically null effect. For each possible true effect, a given method's performance is considered poor if it leads to either of these mistakes. Thus, if δ = 0, an upward bias in mean error of 0.15 or more is poor performance, as the method, on average, would indicate a practically significant effect even though there is a true null effect (because 0 + 0.15 = 0.15). If δ = 0.2, a downward bias in mean error of 0.05 or more is poor performance, as a practically significant true effect would be underestimated as practically nonsignificant (i.e., 0.2 – 0.05 = 0.15). If δ = 0.5, a downward bias in mean error of 0.35 or more would lead to the same mistake (because 0.5 – 0.35 = 0.15).

Considering method performance in a sensitivity analysis. With these definitions for poor performance and plausible conditions in hand, our interactive app (http://www.shinyapps.org/apps/metaExplorer/) can be used to judge whether meta-analytic conclusions rely on methods that perform poorly in plausible conditions. The app indicates that the conclusion that the depletion effect is practically significant is indefensible, as the RE, trim-and-fill, WAAP-WLS, 3PSM, *p*-curve, and *p*-uniform methods, which lead to this conclusion, all perform poorly in at least one of the defined plausible conditions (primarily when δ = 0; Table 2). In contrast, the conclusion that the depletion effect is practically nonsignificant—which is based on results from the PET, PEESE, and PET-PEESE methods—appears to be reasonable because the PET method does not perform poorly in any of the plausible conditions we examined. If one gives more weight to methods that a priori perform well, rather than poorly, under the hypothesized plausible conditions, one would lean toward the conclusion that the depletion effect is practically nonsignificant.

For the data from Hagger et al. (2016), a very different set of conditions seem plausible. Because Hagger et al.'s data come from a Registered Replication Report, there is no reason to think that the results were influenced by publication bias or QRPs to any substantial degree. Furthermore, one can expect significantly less heterogeneity in these data as compared with those from Carter et al. (2015) because data collection at each location was conducted using the identical study design and a preregistered, automated script. Furthermore, there is evidence that heterogeneity in these kinds of multilab registered reports is generally low (Klein et al., 2018). Thus, from among the conditions we simulated, we imagine that researchers would view the most plausible conditions as having no publication bias, no QRPs, no heterogeneity (τ = 0), and effect sizes (δ) less than or equal to 0.5. Using the same definition of poor performance as for the meta-analysis of Carter et al., we evaluate each method's performance in these conditions at $k = 30$ (the simulated value closest to the observed k of 23 studies).

The results from our Web app are shown in Table 3. Unlike the results obtained with the Carter et al. (2015) data (Table 2), these results uniformly suggest that the depletion effect is practically nonsignificant. None of the methods—except *p*-curve and *p*-uniform—perform poorly in any of the plausible conditions we defined.

In summary, we performed two method performance checks in two different plausible research environments. In both cases, the estimates produced by those methods that a priori could be expected to perform reasonably well in plausible conditions suggested that the true ego-depletion effect is not practically or significantly different from zero.

Further considerations for method performance checks and sensitivity analyses. We conclude our discussion of our proposed approach by mentioning some important considerations. First, the form of our method performance check depends on the specifics of our simulation design. Because our simulation covered only a limited set of possible data-generation processes, it is possible that our approach does not generalize to real-world situations that meta-analysts and consumers of meta-analysis will encounter. Indeed, because of the infinite possible processes that might generate real-world data, generalizability will always be a concern.

Second, we suspect that some readers will worry that the method performance check we have described is subject to a kind of "assumption hacking" whereby researchers who are partial to a certain view can pick and choose the definitions of plausible conditions and poor performance that provide the result they want. This concern is technically correct, but the key strength

of our approach is that it is explicit and transparent. Researchers will need to clearly state their assumptions to run this method performance check, and consumers of the results can assess whether such assumptions appear reasonable to them. If potentially hacked assumptions do not appear to be reasonable, our Web app can easily be used to run an alternative method performance check, thereby preventing the possibility of effective assumption hacking. Furthermore, we suggest that analysts preregister a method performance check prior to data collection and define which methods will be given the greatest weight if different methods provide conflicting results. Finally, we encourage researchers to report results from all meta-analytic methods that reasonably can be considered, even if they did not pass some of the method performance checks, because other researchers might want to apply a different emphasis in their subjective evaluation.

Ways forward

On the basis of our results, we emphasize that meta-analysis in psychology is difficult. Observable factors such as small samples—both in the primary literature and at the level of the meta-analysis—interact with heterogeneity and bias, both of which have unknowable severity and functional form (e.g., do the true effects follow a normal distribution?). 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; Fabrigar & Wegener, 2016; D. J. Stanley & Spence, 2014).

Meta-analysts might hope that different bias-correcting methods will all converge on a true value. However, our simulations show that different methods often do not converge. For example, in the case of a true null effect and strong publication bias, the PET and trim-and-fill methods will virtually never give the same answer because the latter performs so poorly. For this reason, we caution against ideas of "triangulation" or basing conclusions on a "majority vote" of multiple methods. One should instead think carefully about which method (or methods) can be expected to perform well. We think a good approach is the combination of a method performance check with a subsequent sensitivity analysis, either as we have defined sensitivity analysis or as put forward by other researchers (e.g., Copas, 2013; Copas & Shi, 2000; Kim, Bangdiwala, Thaler, & Gartlehner, 2014; Vevea & Woods, 2005).

Furthermore, we conclude that the field should modify its expectations about meta-analysis (a similar argument has been made regarding replication results, e.g., by D. J. Stanley & Spence, 2014). Researchers in psychology should not expect to produce conclusive, debate-ending results by conducting meta-analyses on existing literatures. Instead, we think that meta-analyses may serve best to draw attention to the existing strengths and weaknesses in a literature (e.g., Carter et al., 2015; Hilgard, Engelhardt, & Rouder, 2017; van Elk et al., 2015). Meta-analytic results can then inspire a careful reexamination of methodology and theory, perhaps followed by large-scale, preregistered replication efforts (e.g., Hagger et al., 2016). Such efforts can then be summarized with RE meta-analytic methods, which show the best performance in the absence of bias (Figs. 3 and 4).

Conclusion

In simulations using effect sizes, sample sizes, QRPs, and degrees of publication bias that plausibly represent real data in psychology, we compared the performance of seven meta-analytic methods, including six intended to correct for publication bias. We found that each of the seven methods showed unacceptable performance in at least some conditions. This is not an entirely surprising result given previous simulation studies (e.g., Hedges & Vevea, 1996; McShane et al., 2016; Moreno et al., 2009; Rücker et al., 2011; Simonsohn, Nelson, & Simmons, 2014; T. D. Stanley, 2017; T. D. Stanley & Doucouliagos, 2014; van Aert et al., 2016). However, it highlights an important conclusion that we believe needs to be more widely acknowledged: Meta-analysts in psychology and consumers of those meta-analyses should not expect to come to definitive conclusions. Instead, we believe that the most productive outcomes will be generated by method performance checks, sensitivity analyses, and a willingness to carefully design and conduct preregistered replications.

Action Editor

Daniel J. Simons served as action editor for this article.

Author Contributions

E. C. Carter, F. D. Schönbrodt, and J. Hilgard developed the code for and managed the simulation study. F. D. Schönbrodt programmed and maintains the Web app. E. C. Carter, F. D. Schönbrodt, J. Hilgard and W. M. Gervais planned the project and wrote the manuscript. E. C. Carter and F. D. Schönbrodt contributed equally to this work.

ORCID iD

Joseph Hilgard https://orcid.org/0000-0002-7278-4698

Acknowledgments

We would like to acknowledge Tyler Yost for helpful comments on an earlier version of the simulation code and manuscript.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices

Open Data: https://osf.io/rf3ys/
Open Materials: https://osf.io/rf3ys/, http://www.shinyapps.org/apps/metaExplorer/
Preregistration: not applicable
All data have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/rf3ys/. All materials have been made publically available at https://osf.io/rf3ys/ and http://www.shinyapps.org/apps/metaExplorer/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245919847196. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Prior Versions

An earlier version of this manuscript was posted as a preprint at https://osf.io/rf3ys.

Notes

1. It is worth noting that our original intent with this study was in fact to identify, if possible, a single best method across many conditions. Further consideration and helpful comments from our peers changed our minds about this goal.
2. In terms of the heterogeneity metric I^2, the τ values of 0.2 and 0.4, in combination with the specific primary sample sizes we simulated, are approximately equal to what Pigott (2012) proposed as "medium" (I^2 = 50%) and "large" (I^2 = 75%) heterogeneity. RE meta-analysis of our simulated data in the condition with no publication bias and no QRPs, with results aggregated over k and δ, yielded an average observed I^2 of 46% (SD = 17%) for τ = 0.2 and 77% (SD = 10%) for τ = 0.4.
3. One should keep in mind that the estimates of v_i reported by van Erp et al. (2017) may be over- or underestimates as a result of bias (Augusteijn et al., 2019).
4. If the probability of publication was 25%, for example, we drew one random sample from a Bernoulli distribution with p = .25. If the sample value was 1, the simulated result was published.
5. We also could have deleted only outliers in one direction, which would have made the p-hacking more efficient.
6. Technically, we constrained the numerical optimizer to values of 0 or greater. In 10.3% of all cases, the estimate was less than 0.0001.
7. Although in this special case no p-value and no confidence interval are provided, we treated these cases as indicating that the null should not be rejected. Hence, this special case was utilized in the computation of the false-positive error rate, but not the coverage probability.
8. It should be noted that, in contrast to the work just cited, two meta-meta-analyses suggest that the influence of bias in psychology is relatively small (T. D. Stanley, Carter, & Doucouliagos, 2018; van Aert, 2018). Because of the specifics of each of these studies, it is difficult to reconcile their general conclusions. For our analysis here, we decided to take the conservative route and err on the side of assuming the existence of bias, but we recommend that, when applying our approach, meta-analysts consider the degree to which bias exists and explicitly describe their reasoning.
9. Note that by using this metric, we focus on the point estimate provided by each method, not the upper or lower bounds of the confidence interval. An evaluation of methods should also consider RMSE, error rates, and other performance metrics, which all are included in our online app. Furthermore, we want to note that these metrics take only statistical properties of the estimators into account. It has been argued that other dimensions of quality, such as the presence or absence of proper randomization, should go into the weighting of primary studies in a meta-analysis (Detsky, Naylor, O'Rourke, McGeer, & L'Abbé, 1992). But this is a topic for another article.

References

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*, 839–851. doi:10.1037/0003-066X.63.9.839

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*, 116–134.

Baker, R. D., & Jackson, D. (2013). Meta-analysis inside and outside particle physics: Two traditions that should converge? *Research Synthesis Methods*, *4*, 109–124.

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

Bayarri, M. J., & DeGroot, M. H. (1991). *The analysis of published significant results (Technical Report 91-21)*. Retrieved from http://www.stat.purdue.edu/docs/research/tech-reports/1991/tr91-21.pdf

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin*, *144*, 77–110.

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

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

Boulesteix, A.-L., Wilson, R., & Hapfelmeier, A. (2017). Towards evidence-based computational statistics: Lessons from clinical research on the role and design of real-data

benchmark studies. *BMC Medical Research Methodology*, *17*, Article 138. doi:10.1186/s12874-017-0417-2

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*, 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*, 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. *Journal of Experimental Psychology: General*, *144*, 796–815.

Carter, E. C., & McCullough, M. E. (2018). A simple, principled approach to combining evidence from meta-analysis and high-quality replications. *Advances in Methods and Practices in Psychological Science*, *1*, 174–185. doi:10.1177/2515245918756858

Coburn, K. M., & Vevea, J. L. (2017). Weightr: Estimating weight-function models for publication bias (R package Version 1.1.2) [Computer software]. Retrieved from https://CRAN.R-project.org/package=weightr

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

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

Copas, J. B. (2013). A likelihood-based sensitivity analysis for publication bias in meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62*, 47–66.

Detsky, A. S., Naylor, C. D., O'Rourke, K., McGeer, A. J., & L'Abbé, K. A. (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology*, *45*, 255–265.

Du, H., Liu, F., & Wang, L. (2017). A Bayesian "fill-in" method for correcting for publication bias in meta-analysis. *Psychological Methods*, *22*, 799–817. doi:10.1037/met0000164

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*, 455–463.

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.

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*, 891–904.

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences, USA*, *114*, 3714–3719.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*, 555–561.

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological & Personality Science*, *7*, 8–12.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review*, *23*, 74–86.

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

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, *136*, 495–525. doi:10.1037/a0019486

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Distributions of p-values smaller than .05 in psychology: What is going on? *PeerJ*, *4*, Article e1935. doi:10.7717/peerj.1935

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

Hedges, L. V., & 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*, 299–332.

Hilgard, J., Engelhardt, C. R., & Rouder, J. N. (2017). Overstated evidence for short-term effects of violent games on affect and behavior: A reanalysis of Anderson et al. (2010). *Psychological Bulletin*, *143*, 757–774. doi:10.1037/bul0000074

Idris, N., & Ruzni, N. (2012). Performance of the trim and fill method in adjusting for the publication bias in meta-analysis of continuous data. *American Journal of Applied Sciences*, *9*, 1512–1517.

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, *127*, F236–F265.

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

Jackson, D. (2007). Assessing the implications of publication bias for two popular estimates of between-study variance in meta-analysis. *Biometrics*, *63*, 187–193.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:10.1177/0956797611430953

Kim, N. Y., Bangdiwala, S. I., Thaler, K., & Gartlehner, G. (2014). SAMURAI: Sensitivity analysis of a meta-analysis with unpublished but registered analytical investigations (software). *Systematic Reviews*, *3*(1), Article 27. doi:10.1186/2046-4053-3-27

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490. doi:10.1177/2515245918810225

Koehler, E., Brown, E., & Haneuse, S. J.-P. A. (2009). On the assessment of Monte Carlo error in simulation-based

statistical analyses. *The American Statistician*, *63*, 155–162. doi:10.1198/tast.2009.0030

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLOS ONE*, *9*(9), Article e105825. doi:10.1371/journal.pone.0105825

LeBel, E. P., McCarthy, R., Earp, B. D., Elson, M., & Vanpaemel, W. (2017). *A unified framework to quantify the credibility of scientific findings*. Retrieved from https://doi.org/10.31234/osf.io/uwmr8

Marszalek, J. M. (2011). *Sample size in psychological research over the past 30 years* [Data file]. Retrieved from https://mospace.umsystem.edu/xmlui/handle/10355/62220

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*, 331–348.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*, 730–749.

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., 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*, Article 2. doi:10.1186/1471-2288-9-2

Muraven, M., Tice, D. M., & Baumeister, R. F. (1998). Self-control as a limited resource: Regulatory depletion patterns. *Journal of Personality and Social Psychology*, *74*, 774–789.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, *26*, 4544–4562.

Pigott, T. D. (2012). *Advances in meta-analysis*. New York, NY: Springer Science & Business Media.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reed, W. R. (2015). A Monte Carlo analysis of alternative meta-analysis estimators in the presence of publication bias. *Economics: The Open-Access, Open-Assessment E-Journal*, *9*, Article 30. doi:10.5018/economics-ejournal.ja.2015-30

Rice, K., Higgins, J. P. T., & Lumley, T. (2017). A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*, 205–227.

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 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*. Chichester, England: John Wiley & Sons.

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

Schmid, C. H. (2017). Heterogeneity: Multiplicative, additive or both? *Research Synthesis Methods*, *8*, 119–120.

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*, 1359–1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018, January 8). P-curve handles heterogeneity just fine [Blog post]. Retrieved from http://datacolada.org/67

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*, 666–681.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2014, December 3). Trim-and-fill is full of it (bias) [Blog post]. Retrieved from http://datacolada.org/30

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better *p*-curves: Making *p*-curve analysis more robust to errors, fraud, and ambitious *p*-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*, 1146–1152. doi:10.1037/xge0000104

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2016). *Specification curve: Descriptive and inferential statistics on all reasonable specifications*. doi:10.2139/ssrn.2694998

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2017, June 15). Why p-curve excludes ps > .05 [Blog post]. Retrieved from http://datacolada.org/61

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

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological & Personality Science*, *8*, 581–591. doi:10.1177/1948550617693062

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*, 1325–1346. doi:10.1037/bul0000169

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

Stanley, T. D., & Doucouliagos, H. (2015). Neither fixed nor random: Weighted least squares meta-analysis. *Statistics in Medicine*, *34*, 2116–2127.

Stanley, T. D., & Doucouliagos, H. (2017). Neither fixed nor random: Weighted least squares meta-regression. *Research Synthesis Methods*, *8*, 19–42.

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. A. (2017). Finding the power to reduce publication bias. *Statistics in Medicine*, *36*, 1580–1598.

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*, 108–112.

Sterne, J. A., 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*, 1119–1129.

Tackett, J. L., McShane, B. B., Bockenholt, U., & Gelman, A. (2017). *Large scale replication projects in contemporary*

psychological research. Retrieved from http://arxiv.org/abs/1710.06031

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, *22*, 2113–2126.

van Aert, R. C. M. (2017). puniform: Meta-analysis methods correcting for publication bias (R package Version 0.0.3) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/puniform/index.html

van Aert, R. C. M. (2018). *Meta-analysis: Shortcomings and potential* (Doctoral dissertation, Tilburg University). doi:10.31222/osf.io/eqhjd

van Aert, R. C. M., & van Assen, M. A. L. M. (2018). *Correcting for publication bias in a meta-analysis with the* P-*uniform* method*. doi:10.31222/osf.io/zqjr9

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on *p* values: Reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspectives on Psychological Science*, *11*, 713–729.

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*, 293–309.

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*, Article 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*. doi:10.31234/osf.io/myu9c

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*, 55–79.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*, 428–443.

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