**Comparing methods for reducing the influence of bias on meta-analytic estimates in psychology**

**Introduction**

Statistical tools for analyzing the results from a set of studies in aggregate—often called meta-analysis—are popular in many scientific disciplines, including psychology. This popularity likely stems in part from the fact that, as compared to results from individual studies, meta-analytic results have higher statistical power and more precise estimates. However, just as the results from individual studies can be marred by bias, meta-analytic results can be made far less useful, or even completely misleading, when influenced by bias. Researchers have developed statistical techniques designed to identify and correct for bias, and several simulation studies have compared the performance of some of these tools (e.g., Moreno et al., 2009a; Rücker, Carpenter, & Schwarzer, 2011; Stanley & Doucouliagos, 2013; Simonsohn, Nelson, & Simmons, 2014). Unfortunately, nearly all such efforts come from outside of psychology, and have examined effect sizes, sample sizes, and numbers of studies atypical of psychological research. As a result, psychological scientists are faced with an ever-growing menu of meta-analytic tools but little information about which tools are likely to work for the data they commonly encounter.

Here, we present a systematic simulation study of statistical techniques intended to correct for the influence of bias on meta-analytic estimates, and we assess the performance of these techniques on data that are designed to be typical of the research done in psychology.

**Meta-analysis**

Meta-analytic techniques involve synthesizing a set of studies investigating the same empirical phenomenon. For example, meta-analysis is often used to produce a single summary estimate of the hypothetical true underlying effect, *δ*, that each study in the meta-analytic data set purportedly measured. This is usually called fixed-effect meta-analysis (Cooper, Hedges, & Valentine, 2009), and can be modeled as

where *Ti* is the observed treatment effect for study *i* that differs from the true underlying effect δ by some amount of sampling error, *ei*. Sampling error is assumed to be normally distributed with a mean of 0 and a variance of *vi*. A more complex, and, arguably, more realistic model known as random-effects meta-analysis (Cooper et al., 2009) holds that each study measures a different, related true effect, *δi*. This approach allows for the possibility that researchers attempting to study the same phenomenon may nonetheless be studying different underlying effects that vary as a function of, for example, different operationalizations of the independent variable or different populations. Formally,

and *δi* can be further modeled as

Thus, *µ* is the mean of the unknown true effect sizes that are estimated by the individual studies. The *i*th study’s deviation from this mean is *ui*, which is normally distributed with a mean of zero and variance of *τ2*. The applying the random-effects model to an observed set of studies provides an estimate of the average true underlying effect, *µ*, and the amount of between-study heterogeneity, *τ2*. Estimates of these two parameters, particularly the estimate of *µ* and whether it is statistically significantly different from zero, tend to be of primary interest when examining a particular literature with meta-analysis.

Each observation in a meta-analytic data set must include, at a minimum, an estimate of the effect being measured and an estimate of the variance of that effect estimate. Because meta-analyses are usually applied to studies with dependent variables measured on different scales, effect estimates are typically standardized. To synthesize the observed studies, one would need to first transform the results of each study into an effect size measure such as the standardized mean difference, or Cohen's *d*, given as

,

where *M1* and *M2* are the means of the two groups and *S*, the pooled standard error, is

,

where *n* and *v* are the sample sizes and the variances of the groups. The variance of a given Cohen's d is given as

.

**Bias**

Meta-analytic bias is inflicted through factors that affect the analysis and reporting of the individual studies that go into a meta-analysis, and can result in systematic overestimation of meta-analytic estimates and Type I error rates in excess of the nominal threshold. We consider two primary sources of meta-analytic bias: publication bias and questionable research practices.

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

Another form of bias is the undisclosed use of questionable research practices (QRPs; also called “researcher degrees of freedom” or “p-hacking”) whereby researchers—intentionally or not—choose from a variety of potential analyses based on the results they yield. These analytic choices may be justifiable, yet simultaneously arbitrary and motivated (Simonsohn, Simmons, & Nelson, 2015). For example, researchers may encounter a flexible design that can be analyzed in several ways—with or without covariates, with or without planned contrasts, with or without outliers, any of several outcomes—and choose an approach based on the results it produces. Such behavior, although potentially truly motivated by a desire to understand the effect of interest, opens the researcher to the possibility of biased findings. Several papers have thoroughly discussed how frequent QRPs might be and the detrimental effect they have on individual studies (Simmons, Nelson, & Simonsohn, 2011; Gelman & Loken, 2013; Young & Karr, 2011; John, Lowenstein, & Prelec, 2012; Franco, Malhotra, & Simonovits, 2015; Fiedler & Schwarz, 2015), as well as the effect on meta-analyses (Simonsohn, Nelson, & Simmons, 2014).

In practice, QRPs and publication bias are likely to be related processes—the same justifications and motivations associated with discounting statistically non-significant or counter-theory findings may be associated with the undisclosed application of data-driven analyses. This combination of biasing processes would have a strong effect on meta-analysis such that null results are either omitted from analysis or altered into positive results.

**Funnel plots.** The influence of bias in meta-analysis can sometimes be seen by comparing the effect size estimates to the standard errors of those estimates (or some other indicator of sample size) with a funnel plot (Light & Pillemer, 1984). In a typical funnel plot, the reported effect size is plotted on the x-axis and the standard error is plotted on the inverted y-axis. The most precise estimates (i.e., those with the smallest standard error and largest sample) will tend to converge on the true effect size, whereas the more imprecise estimates will spread evenly on either side of the true effect, with studies equally likely to overestimate as underestimate the true effect. That is, the amount of deviation from the true effect increases as estimates become more imprecise, leading to a funnel-like pattern (Figure 1A). In the presence of bias, fewer studies will be present in the lower corner of the funnel where results would be non-significant or of the wrong sign (Figure 1B). In this case, the funnel plot will appear asymmetrical, with more imprecise studies finding larger effects than more precise studies. In this way, a funnel plot can reveal patterns that may indicate bias.

Figure 1A shows funnel plots of simulated meta-analytic data sets. These data sets vary in the true values of the underlying effect, *δ,* and heterogeneity, *τ*. Note that in this panel, none of these meta-analyses have been affected by bias, and the difference between the random-effects model estimate (marked as a solid vertical line and a ‘X’ along the horizontal axis) and the true value is very close to zero. Figure 1B, in contrast, shows these same conditions under complete publication bias—that is, every study producing a statistically non-significant or negative estimate was removed. Note the clear rightward asymmetry of the funnel plots in Figure 1B as compared to in Figure 1A, as well as the resulting overestimation in the random-effects model estimates: Along the horizontal axis, each X has been shifted to the right of the true value.

**Small-study effects.** As is illustrated in Figure 1B, publication bias induces a relationship—in this case, a positive correlation—between effect size estimates and their standard errors. However, such a correlation can have benign causes. It may be, for example, that expensive, small-sample manipulations have stronger effects than inexpensive, large-sample manipulations. Similarly, when a literature contains both large and small effects, and researchers use power analyses to plan their samples sizes accordingly, the large effects will be studied with smaller, less-precise samples. Sequential designs can also induce this correlation (Lakens, 2014; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2015): Studies measuring a large effect can stop early for efficacy, whereas studies measuring a small effect can stop at later stages of the sequential design after continuing data collection. In these scenarios, effect size and standard error also would be correlated, but not because of bias.

The collection of both bias and non-bias-related factors that result in this correlation between sample size and effect size are called “small-study effects” (Sterne, Gavaghan, & Egger, 2000). This issue is an important one, as several of the methods we examine here technically correct for small-study effects generally, not bias specifically.

**Our approach.**

Given the dangerous influence of publication bias and the undisclosed use of QRPS, meta-analytic methods that are robust to this form of bias are sorely needed. Furthermore, as mentioned above, much of the work comparing such methods has not focused on conditions that obviously represent psychological science. For example, many systematic studies of meta-analytic techniques are focused on effect size measures (e.g., log-odds) that are relatively infrequently used in psychology, and given that not all effect sizes are identical, certain methods may perform differently when faced with different data or previous simulations may focus on parameters that are different enough in magnitude from studies in psychology to be of uncertain usefulness. Thus, we focus on meta-analyses of simulated studies for which the data can be described in terms of the standardized mean difference effect size, Cohen's *d*.

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

Previous simulations usually focused on a rather limited set of bias-correcting methods; we provide the most comprehensive evaluation to date of current approaches by including trim-and-fill, meta-regression techniques (e.g., PET-PEESE), *p*-curve and *p*-uniform, and a three-parameter selection model (McShane et al., 2016).

Finally, to our knowledge, our approach provides the most extensive study of the influence of a realistic set of QRPs on meta-analytic estimates, how these factors interact with publication bias, and how they are handled by bias-correcting methods.

**Methods**

All simulations and analyses were conducted in R (R Development Core Team, 2014). R scripts are available in the online supplemental material. Interactive figures for visualizing both our simulation approach and results are available here: xxxxxx.

**Simulation.**

We simulated data for 432 unique combinations of five fully-crossed factors: (1) the true underlying effect, δ (0, 0.2, 0.5, 0.8); (2) between-study heterogeneity, τ (0, 0.2, 0.4); (3) the number of studies in the meta-analytic sample, *k* (10, 30, 60, 100); (4) the percentage of studies in the meta-analytic sample produced under publication bias, *PB* (0%, 60%, 90%); and (5) the use of QRPs in the literature that produced the meta-analytic sample (none, medium, high). See Table 1 for an overview of all conditions. We simulated 1,000 meta-analyses for each of the 432 conditions[[1]](#footnote-1).

|  |  |  |
| --- | --- | --- |
| **Table 1.** | | |
| **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 |
| Percentage of studies produced under publication bias (*PB*) |  | 0%, 60%, 90% |
| QRP environment (*QRP*) |  | None, medium, high |
|  | | |

To simulate an individual study, we first randomly selected a per-group sample size, *n*, from our empirically-derived sample size distribution (Figure 2). DESCRIBE HOW WE GOT THE EMPIRICAL DISTRIBUTION.

Independent samples were then randomly generated for the control and experimental group, where observations in the control group were drawn from a normal distribution with μ = 0 and σ = 1, and observations in the experimental group were drawn from a normal distribution with μ = *D* i and σ = 1. *D* was defined as the sum of the δ and τ i, where τ i was drawn from a normal distribution with mean 0 and standard deviation τ. Note that *D* i, therefore, represented a study-specific true effect that varied randomly if τ was greater than 0. Cohen's *d* and the variance *v* of the effect size were then calculated as described above for the simulated study. A two-tailed independent-samples *t*-test was then applied to generate a *t*-value and *p*-value. Each simulated study, therefore, was represented by a value for *d*, *v*, *p*, and *t*.

In the case of publication bias, the studies produced without the influence of bias followed the procedure just described. For the studies affected by publication bias, studies were produced as described above, but were deleted if the results were statistically non-significant (*p* < .05) or if the effect size was negative (*d* < 0). Studies were continually produced until the target number of studies had been reached. For example, for a meta-analysis with *k* = 10 and 60% publication bias, 4 studies were produced without publication bias (producing either a significant or non-significant result), and then the simulation ran until 6 additional studies with *p* < .05 and *d* > 0 were produced. This disentangles the effects of publication bias from the number of published studies *k* – some previous simulations did not use this approach, and therefore, stronger publication bias reduced the published studies *k*, making it difficult to tell whether poorer performance was due to stronger bias or fewer studies.

We studied four forms of QRPs: (1) Optional removal of outliers, (2) Optional selection between two dependent variables, (3) optional use of moderators, and (4) optional stopping. Each data set that would have QRPs applied to it was designed to simulate a study with a two-by-two (experimental group vs. the control group; level one of the moderator vs. level two of the moderator) design and two dependent variables. Each dependent variable was measured *n* times. 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., no collinearity between moderator and treatment). The two dependent variables were correlated at *r* = 0.50.

QRPs were applied in an algorithmic process, with the goal to simulate QRP settings that are as realistic as possible. The simulated researcher first tested the main effect of experimental manipulation (i.e., experimental vs. control group) on the first dependent variable. If this effect was not statistically significant and positive, the simulated researcher removed outliers (defined as observations with an absolute value z-score greater than 2). If this second test was not positive and significant, the simulated researcher moved to the second dependent variable and repeated the above steps. If no positive and significant effect was found using the second dependent variable, the researcher moved back to the first dependent variable and tested for an interaction effect between the experimental manipulation and the moderator. In the presence of 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, and in the absence of a positive and significant effect, the researcher moved to the second dependent variable, first checking for an interaction effect. If no positive significant effect was found, the researcher moved to the subgroup defined by the second level of the moderator and repeated the above steps.

If none of these analyses produced a positive and significant effect, the first test (experimental vs. control on the first dependent variable with outliers untouched and no division by the moderator) was taken as the final result. However, the simulated researcher could also opt to collect some additional amount of data, with new observations split evenly between each of the four cells. After each additional collection effort, the QRPs described above were repeated. Thus, for each data collection effort, simulated researchers could potentially apply 12 comparisons.

We then created three types of *QRP strategies* a simulated researcher could adopt: (1) pure (no use of QRPs); (2) moderate (optional dependent variables and the addition of three observations per cell for up to three data collection efforts); and (3) aggressive (use of optional outliers, optional dependent variables, optional moderators, and the addition of three observations per cell for up to five data collection efforts). As it is unlikely that every researcher in a field applies QRPs in the same fashion, we defined three *QRP environments* to describe possible prototypical research fields with a specific severity of QRP application. Each QRP environment was characterized by a mixture of simulated researchers with individual QRP strategies: (1) none (100% of simulated researchers adopted the pure strategy); (2) medium (30% pure, 50% moderate, and 20% aggressive); and high (10% pure, 40% moderate, and 50% aggressive). Critically, the QRP percentages and the publication bias percentages were independent, such that the QRP percentages held for the studies produced with and without the influence of publication bias.

To summarize, given that little is known about how or how often QRPs affect psychological research, or the severity of publication bias, our aim was to develop a simulation with a broad enough scope to represent the possible realistic states of various literatures in psychology. As such, researchers with different interests in different literatures can explore our results given the parameters that they believe to be the most appropriate.

**Meta-analytic Methods**

We examined the performance of eight estimators: (1) random-effects meta-analysis (RE), (2) trim-and-fill (TF), (3) precision effect test (PET), (4) precision effect estimate with standard error (PEESE), (5) the conditional combination of PET and PEESE (PET-PEESE), (6) *p*-curve (PC), (7) *p*-uniform, and (8) a three-parameter selection model (3PSM).

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

**Trim-and-fill.** The trim-and-fill approach (Duval & Tweedie, 2000) is based on funnel plot asymmetry. Note how publication bias introduces clear rightward asymmetry in Figure 1 by censoring non-significant and negative observations. The trim-and-fill iteratively removes observations from one side of the funnel plot until a criterion for symmetry is met, and then observations are filled back into the funnel plot along with imputed observations of the opposite sign (Figure 3A). Standard meta-analytic methods are then applied to a data set including both observed and imputed studies. We use the default algorithm provided by the *metafor* package.

**PET.** The precision effect test (PET; Stanley & Doucouliagos, 2013) is derived from a weighted-least-squares regression model where effect size is regressed on the standard error of the effect. In other words,

where *b0* and *b1* are the intercept and slope terms describing the linear relationship between the *i*th effect size estimate *di* and its associated standard error *sei*. The regression model is weighted by the inverse of the variance (i.e., the squared standard errors) of the effect size estimates. Here, *b0* can be understood as an estimate of the effect size for a study where the standard error is zero—that is, an infinitely large study that, in principle, would represent the most accurate estimate of the average true underlying effect. Figure 3B shows an example of PET.

Note that a hypothesis test applied to *b1* in the above is the popular test for funnel plot asymmetry, or Egger's test (Egger, Davey Smith, Schneider, & Minder, 1997). A statistically significant value of *b1* (with a recommended ɑ-level of .10 rather than .05; Egger et al., 1997) is interpreted as evidence for small-study effects, and thus, potentially bias. Another conceptualization of PET is as an estimate of the true underlying effect after controlling for small-study effects. This idea highlights the fact that PET can sometimes erroneously correct for more than bias if other small-study effects influence the literature.

**PEESE.** The precision-effect estimate with standard error (PEESE; Stanley & Doucouliagos, 2013) is another weighted-least-squares meta-regression model; however, in PEESE, effect size is regressed on the variance rather than standard error. Thus,

The weights for this model are the same as for PET—the inverse of the variances. Also, as in PET, the intercept is interpreted as an estimate of the true underlying effect that is uninfluenced by small-study effects. An example of PEESE is shown in Figure 3B.

Notably, both PET and PEESE are examples of weighted-least squares meta-regression and are therefore distinct in some ways from the fixed- and random-effects meta-analysis models described above. The specifics of this difference are discussed in detail elsewhere (Thompson & Sharp, 1999; Stanley & Doucouliagos, 2015XXXX); however, in practice, the result of the difference is that the estimates from weighted-least squares meta-regression models will have relatively larger standard errors, and thus, relatively wider confidence intervals than standard meta-analysis models. This is not necessarily a negative in the face of statistical heterogeneity and publication bias, and authors have argued for the use of both types of models (that Moreno paper; Thompson & Sharp, 1999; Stanley & Doucouliagos, 2015).

**PET-PEESE.** The key difference between PET and PEESE is the way in which small-study effects are modeled—that is, either as a linear relationship (PET) or a quadratic relationship (PEESE). Simulation studies showed that PET is more appropriate when the true underlying effect is zero, whereas PEESE is more appropriate when the true underlying effect is non-zero (Stanley & Doucouliagos, 2013). This pattern of results has been observed in simulation studies, where PET produces underestimates when the true effect is non-zero and PEESE produces overestimates when the true effect is zero. In an attempt to offset the complementary biases of PET and PEESE, Stanely and Doucouliagos (2013) suggested the conditional estimator PET-PEESE. PET-PEESE considers the statistical significance of the estimate from PET to decide whether the PET or PEESE is taken as the final estimate. When the estimate from PET is statistically non-significant (i.e., the estimated true effect is not distinguishable from zero), the PET estimate is taken. In contrast, when the estimate from PET is statistically significant, the PEESE estimate is used as the value for the conditional PET-PEESE procedure.

***P*-curve.** A *p*-curve is the distribution of all statistically significant *p*-values from the set of studies of interest (i.e., *p*s < 0.05; 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 effect sizes and sample sizes of the studies. When studies have no statistical power—that is, when the null is true—the distribution of significant *p*-values is uniform between 0 and .05. With increasing power (i.e., larger effects, larger samples), the *p*-curve becomes increasingly right-skewed, with 0 < *p* < .01 becoming more probable than .04 < *p* < .05. Because the degree of right skew is a function of the average study power, *p*-curve can use the degree of right-skew to (a) test the absence of a real effect (H₀: δ=0), and (b) estimate the average study power, and thus, the average effect size in a fixed-effect model.

By considering only the statistically-significant effect sizes, *p*-curve exhibits some interesting strengths and weaknesses. As a strength, it provides an effect size estimate that is unaffected by publication bias filters; *p*-curve considers only the statistically-significant results, so it does not matter whether the published literature censors statistically-nonsignificant results. As weaknesses, it considers only a subset of the data, which may make it inefficient (McShane et al., 2016). In the presence of QRPs, its estimates may be biased upwards or downwards (van Aert et al., in press). Furthermore, *p*-curve is likely to overestimate the mean effect when there is heterogeneity. This is because studies are more likely to reach statistical significance, and thus be included in *p*-curve, when the true effect is large, compared to when it is small. Thus, large-effect studies are overrepresented relative to small-effect studies[[2]](#footnote-2). Finally, the *p*-curve estimation of the true effect does not provide confidence intervals[[3]](#footnote-3). Therefore we could not apply the coverage metric (see below) to *p*-curve.

***p*-uniform**. Comparable to *p*-curve, the *p*-uniform method focuses on significant results only and utilizes the fact that the distribution of *p*-values is flat under the null (van Assen, van Aert, & Wicherts, 2015). The underlying true effect then can be estimated as a fixed effect by finding the value *d\** for H0: δ = *d\** which makes the distribution of *p*-values as flat as possible[[4]](#footnote-4). *P*-uniform provides a hypothesis test, an estimate of the bias-corrected effect size, and a confidence interval. McShane et al. (2016) demonstrated that *p*-curve and *p*-uniform can be seen as special cases of selection models by assuming 100% publication bias and a homogeneous effect. Computationally, *p*-curve and *p*-uniform only differ by an alternative implementation of the estimation algorithm.

A single significant test statistic is sufficient to be able to compute a *p*-value and a bias-corrected effect size both in *p*-curve and *p*-uniform. We could not find recommendations about the minimum number of significant results that should be available for a *p*-curve or *p*-uniform analysis. In our own simulations, however, we found extremely biased estimates (up to the range of *d* = -0.6) and huge variances in the estimates when only few studies were available (see supplementary Online Appendix A). Based on these analyses, and to give these methods a fair comparison, we decided to perform *p*-curve and *p*-uniform only when >= 4 significant, directionally consistent studies were available.

**Three-parameter selection model (3PSM)**. Recently, McShane, Böckenholt, and Hansen (2016) evaluated the performance of selection models under publication bias (but without QRPs) against that of *p*-curve and *p*-uniform. Selection models, first introduced by Hedges (1984) and later extended by Iyengar and Greenhouse (1988), assume that ... TODO

We employed the three parameter selection model recommended by McShane et al. (2016), henceforward called 3PSM, which contains a parameter for the population average effect size (i.e., μ), one parameter for the heterogeneity of the random effect sizes (i.e., τ²), and one parameter for the probability that a non-significant effect enters the literature[[5]](#footnote-5). This method provides an explicit model for publication bias (in contrast to PET-PEESE) and therefore deals with one source of bias. As it does not model the impact of QRPs, it is unclear how this model performs under these conditions.

**Performance Metrics**

**Hypothesis test.** For the hypothesis test provided by each meta-analytic method, we evaluated the false positive (Type I) error rate at δ = 0 and the true positive rate (i.e., the statistical power) at δ = 0.5. Furthermore, both rates can be combined in a single index, the *pre-experimental rejection ratio* (Bayarri, Benjamin, Berger, & Sellke, 2016), which is computed as ((1 – Type II error rate) / Type I error rate). Hence, the rejection rate quantifies how often a method correctly indicates an effect, compared to how often it incorrectly indicates an effect. Using the conventions for α = 5% and power = 80%, an acceptable rejection ratio for a method can be set at 16:1. In other words, under H1, an adequate method should reject H0 16 times more often than incorrectly rejecting H0 at δ = 0. For rejection ratios below 16:1, the study is not worth running as even a statistically significant result is not very diagnostic (Bayarri et al., 2016).

**Bias-corrected estimation.** Following the recommendations of Burton, Altman, Royston, & Holder (2006), we measured the performance of each method in terms of mean error (ME), root mean squared error (RMSE), and 95% coverage probability. ME (often called bias) is the average of the deviations of each estimate from the true effect (i.e., the errors). ME is not sensitive to variance in estimates, so it is possible that a method produces low ME by equivalent over- and under-estimation: Imagine a very wide, symmetrical distribution of estimates centered on the true value—on average, the estimates are accurate/unbiased, but any individual estimate can be very far from the truth.

RMSE, the root of the average of the squared errors, incorporates information about average error as well as the variance in the estimates (that is, it is the combined error due to variance and bias). It is possible to observe low MSE when a method produces estimates that are consistently biased in one direction: Imagine a very narrow distribution of estimates that is centered away from the true value—variability of these estimates will be low, but each individual estimate deviates from the truth. Thus, a method's estimation performance must be considered in terms of both ME and MSE. For both ME and MSE, values as close to zero as possible are desirable.

We also examined the performance of the 95% confidence intervals associated with each method (except *p*-curve) by examining 95% coverage probability, or, the proportion of times that a given method's confidence interval included the true value of δ. Low coverage is obviously problematic, but coverage rates higher than the nominal 95% are also cause for concern and may indicate overly-wide intervals.

**Results**

We simulated 1,000 meta-analyses under 432 unique conditions (Table 1) and analyzed each with eight different meta-analytic methods. Each method was then examined in terms of its capability to detect a true effect (relative to its false positive rate), and its capability to provide a bias-adjusted effect size estimate.

**Detection of an Effect**. Often one wishes to make an inference regarding the presence or absence of an effect. This research question is usually considered through a hypothesis test against the null hypothesis δ = 0. Figure XXX shows the false positive (Type I error) rates of each technique under [XXX CONDITIONS]. Figure YYY shows the statistical power of each technique to reject the null hypothesis when it is false [YYY CONDITIONS]. Figure ZZZ presents the rejection ratio of each technique under [ZZZ CONDITIONS]. These rejection ratios represent the ratio of power to false positives; a higher rejection ratio is preferred, as it signifies a test that is more sensitive [and/or] more specific.

**Bias-adjusted Estimation of an Effect Size**. Besides making inferences regarding whether to reject or retain the null hypothesis, one also wishes to estimate the underlying effect size and describe the degree of uncertainty or variability of that effect size across studies.

We therefore avoid an exhaustive description of our results here, and instead aim for a general overview. To do so, we have implemented a simple, point-based ranking scheme based on estimation (i.e., ME, MSE, and coverage probability) supplements with a discussion of hypothesis testing (i.e., statistical power and Type I error rate). The point scheme is based on (1) the assumption that performance on each metric is equally important to the analyst, and (2) a set of cut-offs that define “good performance”, “acceptable performance,” and “poor performance” for each metric (Table 2). Points are assigned following the cut-offs for each metric—2 points for good performance, 1 point for acceptable performance, 0 points for poor performance—and a summed score calculated for each method under each simulated condition. Importantly, our scoring metric is arbitrary—one could assign any number of points based on any set of cut-offs—and we make use of it entirely as an aid in understanding overarching patterns in the raw results. Readers interested in a more nuanced view or the performance of certain methods under specific conditions will find all of our results in the supplemental material.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2** | | | | | | | | |
| Performance Category | . | Mean error (ME) | . | Mean squared error (MSE) | . | Coverage probability (CP) | . | Pts. |
| Good |  | |ME| < 0.10 |  | MSE < 0.102 |  | |CP - 0.95| < 0.01 |  | +2 |
| Acceptable |  | 0.10 < |ME| < 0.15 |  | 0.102 < MSE < 0.152 |  | 0.01 < |CP - 0.95| < 0.02 |  | +1 |
| Poor |  | 0.15 < |ME| |  | 0.152 < MSE |  | 0.02 < |CP - 0.95| |  | 0 |
| *Note.* Pts. = points. | | | | | | | | |

**No publication bias (0%)**

Across values of δ, with zero heterogeneity, and in the absence of QRPs, all methods except for p-curve performed well (Figure 3). The standard random-effects meta-analysis method showed nearly uniformly good performance, whereas the meta-regression methods—PET, PEESE, and PET-PEESE—were a close second, followed by the trim-and-fill. Point deductions in these conditions were essentially due to issues with undercoverage—that is, coverage probabilities lower than the nominal 95%. Somewhat surprisingly, these problems occurred with increasing δ and increasing *k* for the meta-regression methods and at all levels of these parameters for the trim-and-fill.

For all methods, increasing heterogeneity decreased performance, again because of undercoverage, although the random-effects method was robust to this problem at higher levels of *k*. Besides p-curve, the meta-regression methods, particularly PET, suffered the most from increasing heterogeneity, especially at the highest level of δ, due to poor performance on both ME (i.e., underestimation of δ) and MSE.

The random-effects and meta-regression methods suffered with the addition of QRPs due to problems with over- and under-estimation, respectively. Interestingly, in some cases, the addition of QRPs seemed to improve performance by the trim-and-fill and p-curve. For the trim-and-fill, these improvements were specifically due to increased coverage probabilities, whereas for p-curve, improvements occurred across the board. In general, problems caused by increasing QRPs were exacerbated by increases in heterogeneity.

We also examined performance in terms of statistical power and false positive rates. For conditions where δ = 0, the values in Figure 4 represent false positive rates. Ideally, these numbers are kept to 0.05 (equivalent to 95% coverage probability). In the absence of heterogeneity, the random-effects method and the meta-regression methods showed good performance, whereas trim-and-fill tended to falsely reject the null at higher rates and the p-curve essentially failed to produce confidence intervals due to a lack of statistically significant studies in the simulated meta-analyses. Increasing heterogeneity, *k*, and QRPs also increased the false positive rate for all methods, although the random-effects method was the most robust to these changes.

In the absence of heterogeneity and QRPs, statistical power increased with increasing *k* and δ. Increasing heterogeneity and QRPs, however, decreased power. This occurred most severely for the meta-regression methods (especially PET), whereas the random-effects method showed consistently adequate power (greater than 0.80) with trim-and-fill a close second.

**Partial publication bias (60%)**

When publication bias was set to 60%, a large amount of variation in performance appeared to be determined by the value of δ such that performance patterns when δ = 0 or δ = 0.20 were more similar to each other than patterns when δ = 0.50 or 0.80 (Figure 5). When δ = 0 or δ = 0.20, the meta-regression methods showed the best performance. For these lower ranges of δ and in the absence of QRPs, PET showed the best performance, although performance decreased sharply with increasing heterogeneity. PEESE and PET-PEESE began to dominate PET as QRPs increased, and PET-PEESE tended to outperform PEESE as heterogeneity increased. For these methods in these conditions, performance decreases tended to be driven by decreasing coverage first, then increasing MSE as opposed to increasing ME. Notably, larger values of ME tended to be positive—that is, these methods *over*-estimated the true underlying effect. Somewhat counter-intuitively, increasing *k* impaired, rather than improved, 95% CI coverage; this is because increasing *k* tightened confidence intervals around the upwardly-biased point estimates.

When δ = 0.50 or δ = 0.80, acceptable and good performance was observed most consistently for the random-effects method, trim-and-fill, and PEESE, although these methods only rarely showed good or even acceptable performance for coverage probability. Of these methods, trim-and-fill and PEESE performed the most consistently well, whereas the random-effects method showed decreasing performance in the face of increasing heterogeneity. QRPs decreased performance for PEESE, primarily at the highest levels of δ and heterogeneity, whereas in some cases increasing QRPs increased performance by the trim-and-fill and the random-effects method. Unlike when δ = 0 or δ = 0.20, estimation errors by PEESE were primarily in the direction of *under*-estimation, whereas the random-effects method tended to over-estimate the true effect.

As is clear from Figure 6, the 0.05 threshold for false positive rates was reached very rarely, and when it was, almost exclusively by the meta-regression methods in the absence of heterogeneity. Otherwise, all methods tended to falsely reject the null hypothesis at far higher than the nominal rate. Where δ > 0, the trim-and-fill and random-effects model almost always rejected the null hypothesis (statistical power rarely below 0.95), regardless of condition. For PEESE, PET-PEESE, and p-curve, power tended to be adequate (i.e., at least 0.80), improved with increasing values of *k*, and decreased with increasing values for heterogeneity. Of all the methods, PET showed the lowest statistical power, particularly with increasing QRPs and at lower values of δ.

**Strong publication bias (90%)**

As was the case when publication bias was set to 60%, a large amount of variation in performance appeared to be due to the true value of δ when publication bias was set to 100% (Figure 7). When δ = 0 or δ = 0.20, the meta-regression methods showed the least poor performance, particularly PET. The exact method that performed best was a function of both δ and the level of heterogeneity, and performance tended to improve slightly with increases in QRPs. As was the case when publication bias was set to 60%, increases in heterogeneity tended to result in positive values of ME—that is, over-estimation of the true underlying effect. Notably, at the highest level of heterogeneity when δ = 0, no method showed anything but poor performance.

When δ = 0.50 or δ = 0.80, the random-effects method, trim-and-fill, PEESE, and p-curve showed the most consistent performance. Coverage probability was nearly uniformly poor for these methods, whereas ME and MSE performance tended to decrease with increasing heterogeneity but increase with increasing QRPs. For p-curve and PEESE, ME tended to be related to under-estimation as opposed to over-estimation, whereas the opposite was true for the random-effects method and the trim-and-fill.

For hypothesis testing (Figure 8), all methods except for PET and p-curve showed statistical power *and* false positive rates of approximately 0.80 or more—that is, regardless of whether the true average underlying effect was zero or not, these methods tended to reject the null hypothesis in the vast majority of cases. PET and p-curve fared better in terms of false positive rates, but only in the absence of heterogeneity and seemingly at the expense of statistical power as δ increased above zero.

**Estimates of heterogeneity**

**[Do we want to get into this? Is there even output for this? Many of our applied techniques are blind to heterogeneity and could not attempt to estimate it.]**

**Discussion**

We inspected and compared the efficacy of meta-analytic adjustments for bias across thousands of simulated literatures, representing different true effect sizes, degrees of heterogeneity, degrees of publication bias, and degrees of questionable research practices. We assessed the results according to the bias and efficiency of each estimator across conditions.

Across conditions, the performance of the three-parameter selection model was the most promising. Its estimates were largely unbiased and efficient even in the context of publication bias and heterogeneity. The three-selection parameter model therefore provides, in many settings, a substantial improvement over naïve random-effects meta-analysis.

By contrast, each other adjustment seemed to suffer one or more shortcomings. Trim-and-fill did little to alleviate publication bias; small and null effects (0 ≤ δ ≤ 0.2) remained substantially overestimated after adjustment. PET-PEESE estimates were highly variable due to the conditional nature of the estimator and the extrapolation involved in meta-regression. Both p-curve and p-uniform are upwardly biased and suffer from high Type I error rates in the presence of heterogeneity. “Top N” methods inflict a loss of efficiency too large to be justified by the very modest improvement in bias they provide.

Additionally, 3PSM was the only technique to provide acceptable Type I error rates across scenarios. Trim and fill and Top N did not adjust enough for bias to recover δ = 0 after any amount of publication bias. P-curve and p-uniform are upwardly biased in the presence of heterogeneity, and their Type I error rates increase rapidly with increasing heterogeneity and *k*. [PET-PEESE I have no idea what’s going on.]

That said, 3PSM does have two weaknesses. Its first weakness is downward bias when the meta-analyzed studies are influenced by QRPs, such that we have implemented them. This weakness is unremarkable, as it is shared by p-curve and p-uniform. Despite this weakness, we are encouraged to find that QRPs cause downward, rather than upward, bias in 3PSM – this allows 3PSM estimates to maintain conservative Type I error rates. The second weakness of 3PSM is that it cannot be implemented when all studies are statistically significant. When publication bias is very strong and unpublished results are difficult to retrieve, this may represent a substantial limitation to what one can accomplish through meta-analytic adjustment.

Other limitations apply that are inherent to meta-analysis and not limitations of the specific method. Point estimates can be highly variable, particularly when *k* is small and heterogeneity is large. As 3PSM has to model parameters for each of the average effect, the heterogeneity, and the selection probability, it can require many studies to provide accurate and precise parameter estimates.

**Other notes**

**Overadjustment.** In the course of our simulations, we found that some methods in some circumstances tended to suffer from high rates of a peculiar sort of type I error. This peculiarity was estimation of the true effect size as being both statistically significant and of opposite sign than the naïve estimate. PET, 3PSM, and p-curve would often report significant negative effects, particularly in the context of strong heterogeneity and QRPs. It seems very unlikely that one should recover a negative true effect from a literature that argues a significant positive effect. Thus, we encourage researchers to interpret adjusted effect sizes of opposite sign as overadjustments driven by heterogeneity and QRPs; a more likely effect size is zero.

**The influence of QRPs.** Looking at the performance of naïve random-effects meta-analysis across settings, we make the surprising observation that, given a degree of selection for significance in publication and given a number of published studies *k,* the degree of QRPs used in publishing those *k* studies has only a small effect in increasing the bias of the estimate. This suggests to us that the primary cause of bias in meta-analysis may be the publication filter that favors statistically-significant results. When QRPs are common but every result is published, there is only a very slight upward bias – every significant result is still matched by several failures to replicate, and of the studies that are p-hacked, not every one finds statistical significance, and many yet publish null results. By contrast, when publication bias is strong, the bias is already inflicted: It does not matter whether the ten significant results come from twenty p-hacked studies or from two hundred honest studies.

Given that selection bias seems to have the stronger role in meta-analytic bias, we suspect that the effect of QRPs in meta-analysis is less so to bias the effect size as to increase the number of studies that appear to confirm the biased estimate. Consider a hypothetical example: Suppose we plan to perform two hundred experiments on a null effect, and journals will only publish the significant results. Suppose further that every experiment is a two-sample between-subjects design with twenty subjects per cell. To reach statistical significance, and hence, publication, a study must yield an effect size of *d* ≥ 0.64. If there is no p-hacking, we expect (on average) 10 false positives to be published; each reports *d* ≥ 0.64, and so the meta-analysis reports *d* ≥ 0.64. If there is p-hacking, Type I error is inflated, say, to 25%. In this case, fifty false positives are published; again, each reports *d* ≥ 0.64, so the meta-analysis similarly reports *d* ≥ 0.64. The degree of bias was predetermined by the sample size and the publication filter; p-hacking only makes it possible for more studies to reach that degree of bias. As one of our departments used to joke before awareness of p-hacking, “everything correlates at *r* = 0.3” – at least, when one is working from a sample of 40 observations and has filtered for statistical significance.

One might be tempted to continue the current norms of selective publishing and trust 3PSM to sort it all out. We do not encourage this. Selection filters harm the efficiency of 3PSM in two ways: first, publishing only the significant results reduces the analyzed studies *k,* and second, stricter selection filters introduce further sampling error into the parameter estimates. Indulging in p-hacking in primary research may also lead to bias in 3PSM estimates; an excess of Type I error in primary research may lead to an increased Type II error in meta-analysis.

**Limitations**

**QRPs.** We modeled some forms of QRPs and found that they had a small upward bias on naïve estimates and often a downward bias on adjusted estimates. Of course, QRPs are a heterogeneous batch of behaviors. Outlier exclusion, optional stopping, subgroup analysis, and outcome switching may each have their own effects on meta-analysis and adjustment. It is hard to say the degree to which adjustments are influenced by the kind of QRPs that happen in the real world. But this is a start.

**Selection filter.** We modeled a simple selection filter based on a single p < .05 threshold. Given the increase in appreciation for well-powered null results, it is possible that selection filters are less rigid for large sample sizes. Well-powered results immune to publication bias may be expected to improve the performance of PEESE and Top N, which emphasize the high-powered studies.

**Orthogonality of factors.** Here we have modeled effect size, heterogeneity, publication filters, and QRPs as independent processes. In the real world, they may not be independent. QRPs may be more common when they are necessary to reach statistical significance, such as when the true effect size is small or nil. Strict selection filters on a research area may encourage QRPs, and QRPs may give the impression that journals should expect statistically-significant results.

**Ways forward**

The most effective route to accurate meta-analytic results is to prevent bias directly in the primary literature. As can be seen from Figures 3 and 4, the standard random-effects method performs admirably in the absence of QRPs and publication bias, regardless of the true underlying effect or the level of heterogeneity.

Of course, for researchers interested in applying meta-analysis to psychology research as it currently is, this solution is of no value. Our results suggest that application of the three-parameter selection model is relatively unbiased, efficient, and appropriate for application in psychological research. However, its estimates are still subject to sampling error and can be misleading due to chance. Given the greater variance associated with 3PSM, it may be difficult to tell whether to place greater emphasis on a particular meta-analysis’ 3PSM result or its naïve random-effects estimate.

Researchers can take steps to improve the quality of meta-analytic conclusions regardless of the particular estimation strategy. Some steps are methodological. Conventional recommendations suggest limiting the influence of heterogeneity through application of sensible inclusion criteria. Publication bias can be mitigated by all relevant parties: meta-analysts, by searching for and retrieving unpublished research; journal editors, by publishing competent research regardless of the significance of its conclusions; authors, by abstaining from QRPs and selective publication. Estimates are more accurate and efficient when applied to comparable studies and when unpublished data is retrieved. Other steps regard the presentation of results. Share the data and encourage other researchers to double-check the coding of studies and the application of methods. Use funnel plots to visualize the data and identify outliers. Provide sensitivity analyses to diagnose outliers’ influence and the robustness of results.

Of course, the amount each party can accomplish on its own has its limits. Meta-analysis depends on there being a large literature and doing a lot of work to pick up unpublished data. Heterogeneity is often unavoidable and cannot be eliminated through moderator analysis or subgroups. Determining the comparability of studies (e.g., whether studies are comparable) is often a subjective and contentious decision. Inclusion criteria can’t prevent publication bias. Unpublished data is challenging, if not impossible to recover. If psychological science wants to take publication bias very seriously, perhaps it will be necessary to organize trial registries. [And what of QRPs?]. Furthermore, even in exact, direct replications of the same experiment, previous large scale multi-site collaborations found a considerable amount of heterogeneity (e.g., in the ManyLabs 1 study, Klein et al., 2014; for additional analyses, see McShane et al., 2016). Hence, even under optimal conditions heterogeneity should be expected.

Conclusions

As always, there are limits to what can be accomplished through the application of statistics. An ounce of bias prevention is worth a pound of selection modeling (paraphrasing <https://twitter.com/hildabast/status/787679389697904640>). If the field is serious about providing accurate, unbiased results with the greatest efficiency possible, authors, peer-reviewers, journals, and funders must commit to the complete and accurate report of all study data.

In the meantime, psychology must make the best decisions possible with the available evidence. To this end, we recommend the use of 3PSM, which appears to be unbiased and efficient in the context of publication selection and heterogeneity. Although it is downward-biased in the presence of QRPs, this seems preferable to an upward bias, as it allows conservative Type I error control. By contrast, p-curve and p-uniform suffer from upward bias under heterogeneity, PET has downward bias when the null is false, PEESE has upward bias when the null is false, and trim-and-fill and Top N fail to fully adjust for publication bias.

As ever, meta-analysis should not be seen as the final arbiter of the status of some phenomena, but rather as another piece of evidence to be considered. Additionally, we encourage meta-analysts to consider that there is more to a meta-analysis than the point estimate of the average effect size. It is useful to know the average effect size and whether it is statistically significant, but it is also helpful to discuss the number, quality, and comparability of available studies. Meta-analytic datasets should be archived and shared for further inspection and future innovations in bias-adjustment

[Prospective meta-analysis of pre-registered studies sees pretty compelling, right?]

**References**

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

Cooper, H., Hedges, L.V., & Valentine, J.C. (2009). The handbook of research synthesis and meta-analysis (2nd edition). New York, NY: Russell Sage Foundation.

Duval, S., and Tweedie, R. L. (2000). Trim and fill: a simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. Biometrics 56, 455–463. doi: 10.1111/j.0006-341X.2000.00455.x

Ferguson, C. J., and Heene, M. (2012). A vast graveyard of undead theories: publication bias and psychological science's aversion to the null. Perspect. Psychol. Sci. 7, 555–561. doi: 10.1177/1745691612459059

Fiedler, K., & Schwarz, N. (2015). Questionable Research Practices Revisited. Social Psychological and Personality Science, 1–18. http://doi.org/10.1177/1948550615612150

Franco, A., Malhotra, N., & Simonovits, G. (2015). Underreporting in psychology experiments: Evidence from a study registry. Social Psychological and Personality Science, 1948550615598377. doi:10.1177/1948550615598377

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. PloS one, 9(10), e109019.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Unpublished manuscript, retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p\_hacking.pdf

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. Psychol. Bull. 82, 1–20. doi: 10.1037/h0076157

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

Light, R. J., and Pillemer, D. B. (1984). Summing up: The Science of Reviewing Research. Cambridge, MA: Harvard University Press.

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, 0956797611417632.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-Curve and effect size correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666-681.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. Available at SSRN.

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

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(11), 1119-1129.

Rothstein, H., Sutton, A. J., & Borenstein, M. (2005). Publication bias and metaanalysis: Prevention, assessments and adjustments. Chichester: John Wiley & Sons.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. J Stat Softw, 36(3), 1-48.

Young, S. S., & Karr, A. (2011). Deming, data and observational studies.Significance, 8(3), 116-120.

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. Statistics in Medicine, 18(20), 2693–2708.

Stanley, T. D., & Doucouliagos, H. (2015). Neither fixed nor random: weighted least squares meta‐analysis. Statistics in medicine, 34(13), 2116-2127.

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., et al. (2009a). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. BMC Med. Res. Methodol. 9, 1–17. doi: 10.1186/1471-2288-9-2

Rücker, G., Carpenter, J. R., and Schwarzer, G. (2011a). Detecting and adjusting for small-study effects in meta-analysis. Biom. J. 52, 351–368. doi: 10.1002/bimj.201000151

Stanley, T. D., and Doucouliagos, H. (2013). Meta-regression approximations to reduce publication selection bias. Res. Synth. Methods 5, 60–78. doi: 10.1002/jrsm.1095

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014b). P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. Available online at SSRN:http://ssrn.com/abstract=2377290 or <http://dx.doi.org/10.2139/ssrn.2377290>

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2015). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*. doi:10.1037/met0000061

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*, 701–710. doi:10.1002/ejsp.2023

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904. doi:10.1007/s11192-011-0494-7

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, *2*, e124. doi:10.1371/journal.pmed.0020124

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. doi:10.1177/1745691616662243

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., R. B., Bahník, Š., Bernstein, M. J., Bocian, K., et al. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, *45*, 142–152. doi:10.1027/1864-9335/a000178

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, *3*, 109–117. doi:10.1214/ss/1177013012

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. doi:10.1037/met0000025

Bayarri, M. J., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, *72*, 90–103. doi:10.1016/j.jmp.2015.12.007

1. For a selection of conditions we also computed 10,000 simulations and compared the results to that of 1,000 simulations. These comparisons clearly demonstrated that 1,000 replications lead to stable estimates (see online appendix A, TODO OSF-Link; and also McShane et al., 2016). [↑](#footnote-ref-1)
2. Simonsohn et al. (2014) address this weakness by stressing a more nuanced, limited interpretation of the *p*-curve estimate: “It is the average effect size one expects to get if one were to rerun all studies included in the *p*-curve” (p. 667; xxxx). In practice, however, one generally hopes to describe the mean effect size of *all* studies. We apply *p*-curve in this fashion. [↑](#footnote-ref-2)
3. Confidence intervals could be computed using a bootstrap. But given that in many conditions only very few significant studies are entered, the bootstrap would have to draw from, say, four significant test statistics, which would not lead to numerically unstable estimates. [↑](#footnote-ref-3)
4. We used the default „P“ method from the *puniform* package for *R*, which relies on the Irwin-Hall distribution. [↑](#footnote-ref-4)
5. We employed the *estimate.onestep.selection.heterogeneous* function provided by McShane et al. (2016) with very general starting values for the optimizer: expected.d = 0.3, max.tau= 0.5, p.report = 0.99. [↑](#footnote-ref-5)