DEPARTMENT OF ECONOMICS AND FINANCE

SCHOOL OF BUSINESS AND ECONOMICS

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

Using Monte Carlo Experiments to Select Meta-Analytic Estimators

NOTE: This paper is a revision of University of Canterbury WP No. 2019/13

Sanghyun Hong
W. Robert Reed

WORKING PAPER

No. 10/2020

Department of Economics and Finance
School of Business
University of Canterbury
Private Bag 4800, Christchurch
New Zealand

WORKING PAPER No. 10/2020

Using Monte Carlo Experiments to Select Meta-Analytic Estimators

Sanghyun Hong[1]
W. Robert Reed[1†]

June 2020

Abstract: The purpose of this study is to show how Monte Carlo analysis of meta-analytic estimators can be used to select estimators for specific research situations. Our analysis conducts 1,620 individual experiments, where each experiment is defined by a unique combination of sample size, effect heterogeneity, effect size, publication selection mechanism, and other research characteristics. We compare eleven estimators commonly used in medicine, psychology, and the social sciences. These are evaluated on the basis of bias, mean squared error (MSE), and coverage rates. For our experimental design, we reproduce simulation environments from four recent studies: Stanley, Doucouliagos, & Ioannidis (2017), Alinaghi & Reed (2018), Bom & Rachinger (2019), and Carter et al. (2019a). We demonstrate that relative estimator performance differs across performance measures. An estimator that may be especially good with respect to MSE may perform relatively poorly with respect to coverage rates. We also show that sample size and effect heterogeneity are important determinants of relative estimator performance. We use these results to demonstrate how the observable characteristics of sample size and effect heterogeneity can guide the meta-analyst in choosing the estimators most appropriate for their research circumstances. All of the programming code and output files associated with this project are available at https://osf.io/pr4mb/.

Keywords: Meta-analysis, Estimator performance, Publication bias, Simulation design, Monte Carlo, Experiments

JEL Classifications: B41, C15, C18

Acknowledgments: We received especially helpful comments from several anonymous reviewers that resulted in substantial improvements over an earlier version of this manuscript. We also thank Isaiah Andrews, Marcel van Assen, Pedro Bom, Maximillian Kasy, and Tom Stanley for making their code available and for helpful discussions, and participants in the 2019 MAER-Net Colloquium for profitable comments and discussions. Finally, we gratefully acknowledge financial support from the Czech Science Foundation, Project 18-02513S.

[1] Department of Economics and Finance, University of Canterbury, NEW ZEALAND

†Corresponding author: Bob Reed, email: bob.reed@canterbury.ac.nz

1. INTRODUCTION

The purpose of this study is to show how simulation studies can be designed to recommend meta-analysis (MA) estimators for specific research situations. In order to be useful, any recommendation must be based on observable characteristics of the meta-analyst's sample of studies. We show that effect heterogeneity and number of studies in the meta-analyst's sample are two characteristics that can be used to guide the selection of best performing estimators. Other characteristics, such as size of the true effect, the degree of publication selection, and the severity of questionable research practices, while important determinants of estimator performance, are generally not observable, and thus cannot be used to select estimators.

In support of this demonstration, we undertake the largest, most extensive Monte Carlo analysis of MA performance to date. Our analysis conducts 1,620 individual experiments, where each experiment is defined by a specific combination of sample size, effect heterogeneity, effect size, publication selection mechanism, and other research characteristics. We compare eleven estimators commonly used in medicine, psychology, and the social sciences. We assess these estimators on the basis of bias, mean squared error (MSE), and coverage rates.

Rather than designing our own Monte Carlo experiments, we reproduce the experimental design from four previous studies: Stanley, Doucouliagos, & Ioannidis (2017), Alinaghi & Reed (2018), Bom & Rachinger (2019), and Carter et al. (2019a). We do this for two reasons. First, Monte Carlo experiments are by definition artificial representations of a complex reality. They involve a large number of subjective judgments. We wanted to select designs that had to some extent been approved by the peer review process. Second, we wanted to use multiple experimental designs to assess whether experimental design had a substantial influence on estimator performance.

Our study proceeds as follows. Section 2 describes the estimators that we study. Section 3 highlights the main characteristics of the different simulation environments used for our analysis. Section 4 defines the performance measures. Section 5 presents our results. Among other things, we demonstrate that relative estimator performance differs across performance measures. An estimator that may be especially good with respect to MSE may perform relatively poorly with respect to coverage rates. We also show that sample size and effect heterogeneity are important determinants of relative estimator performance. Section 6 follows these insights and gives an example of how the observable characteristics of sample size and effect heterogeneity can guide the selection of a "best" estimator for a given research situation. Section 7 concludes with a summary of our main results and suggestions for future research. All of the programming code and output files associated with this project are available at https://osf.io/pr4mb/. We note that our code borrows considerably from Carter et al. (2019b).

2. THE ESTIMATORS

As noted by Carter et al. (2019a, page 117), while many studies have analyzed the performance of meta-analyltic estimators, "…there is very little overlap among these studies in either the methods they have examined or the simulated conditions they have explored." TABLE 1 summarizes a selection of previous Monte Carlo studies and compares them in terms of the number of experiments and estimators studied. Our study analyses and compares the performance of eleven estimators. This compares favorably with previous studies both in terms of number of estimators and variety in the types of estimators. We chose our estimators because they either are widely used in the meta-analysis literature, or have recently appeared in prominent publications.

The context. The estimators are best described within a research context. The following example focuses on a linear regression model, but is easily extended to analyses involving Cohen's d and Log-Odds/logistic regression. Suppose a researcher is interested in synthesizing

the results of an empirical literature. The literature consists of studies that estimate the effect of *X* on *Y* using the following linear regression model,

(1) $Y_{it} = \alpha_i + \beta_i X_{it} + \sum_k \gamma_{ikt} Z_{ikt} + \epsilon_{it}, t = 1,2, \dots T_i,$

where *i* identifies a given regression having T_i observations. The true effect of *X* on *Y* in any given regression is given by β_i. β_i can differ across regressions for many reasons that are unobservable to the meta-analyst. The distribution of the population effect β_i across regressions is represented by $\beta_i \sim N(\mu, \tau^2)$, $\tau^2 \geq 0$.

Let $\hat{\beta}_i$ be the estimated effect from regression *i*. The meta-analyst collects a sample of estimates, $\hat{\beta}_i$, *=1,2,…,N,* and wants to estimate μ, the population mean effect of *X* on *Y.* They know that publication selection may distort their sample of estimates. They have the following estimators available to them: Trim-and-Fill, p-curve, p-uniform, Random Effects, Three-Parameter and Four-Parameter Selection Models, Andrews & Kasy's (2019) "symmetric selection" and "asymmetric selection" models, the Weighted Average of the Adequately Powered-WLS hybrid estimator, PET-PEESE, and Bom & Rachinger's (2019) Endogenous Kink estimator. Each of these is briefly described below.

Trim and Fill (TF). Trim and Fill (Duval & Tweedie, 2000) is a method that assumes that any asymmetry in the distribution of effect sizes and standard errors is due to publication selection. The method works by iteratively removing individual observations until symmetry in the distribution of effect sizes and standard errors is achieved. The removed observations are then added back into the sample, along with artificially generated effect/standard error observations that are the mirror images of the removed observations. This ensures that the reconstructed meta-analysis sample achieves symmetry. Our estimates are obtained using the *metafor* package in R.

p-curve (pC) / p-uniform (pU). The p-curve (Simonsohn, Nelson, & Simmons, 2014) and p-uniform (Van Assen, van Aert, & Wicherts, 2015) methods are conceptually identical

and similar in implementation. Both estimate the mean true effect from the sample of meta-analysis estimates that are statistically significant; i.e., have p-values less than 5%. Both assume that estimates with p-values less than 0.05 are equally likely to be published, and that the respective p-values are independently distributed. Both methods work from the starting point that the distribution of p-values (the "p-curve") will be uniformly distributed between 0 and 0.05 if the null hypothesis is true. Larger, positive effects produce a right skewness to the shape of the "p-curve".

Conceptually, both methods estimate the value of the true (unobserved) effect that would produce a "p-curve" closest to the observed "p-curve". Both define a loss function that measures the distance between the (transformed) expected and the observed p-curves and choose a mean true effect that minimizes that loss function. The two methods differ in the metric they use to measure distance. P-curve uses the Kolmogorov-Smirnov test statistic as a distance metric, while p-uniform's metric is based on the Irwin-Hall distribution. We follow standard practice and only include significant estimates that are same-signed (positive in our case). Our p-curve estimates are obtained from the programming code in Carter et al. (2019b). Our p-uniform estimates use method one in the *puniform* package in R.

Random Effects (RE). The random effects estimator is arguably the most commonly used meta-analytic estimator. It does not explicitly correct for publication selection other than giving greater weight to more precise estimates of β_i. It estimates the population mean effect μ assuming the following specification:

(2) $\hat{\beta}_i = \mu + \varepsilon_i$, *i = 1,2,...,N,*

where $\varepsilon_i \sim N(0, \sigma^2_{\hat{\beta}_i} + \tau^2)$, $\sigma^2_{\hat{\beta}_i}$ is the variance in $\hat{\beta}_i$ due to sampling error, and τ^2 is the variance of true effects across studies. $\sigma^2_{\hat{\beta}_i}$ is estimated by SE_i^2, where SE_i is the (estimated) standard error of the estimated effect, $\hat{\beta}_i$. A variety of procedures have been developed to estimate τ^2.

Our RE estimates are obtained using the R package *metafor*, where $\hat{\tau}^2$ is calculated using the restricted maximum likelihood method.

Three-Parameter and Four-Parameter Selection Models (3PSM and 4PSM): A variety of selection models have been proposed in the literature to correct for publication bias (Iyengar & Greenhouse, 1988; Vevea & Hedges, 1995; Vevea & Woods, 2005). A common model is the Three-Parameter Selection Model (3PSM). 3PSM assumes that standardized effect sizes ($\hat{\beta}_i/SE_i$) are distributed normally in the population. Publication selection induces differential probabilities of being published, with publication probabilities following a step function. The general method allows researchers to set the values of the steps. For our 3PSM analysis, we follow Carter et al. (2019a) in allocating estimates to two categories depending on whether the estimates are (i) correctly signed (positive) and statistically significant, $\left(\hat{\beta}_i/SE_i\right) \geq 1.96$; or (ii) not correctly signed and significant, $\left(\hat{\beta}_i/SE_i\right) < 1.96$. These have relative publication probabilities equal to 1 and p_1, respectively (see Panel A of FIGURE 1). The "Three-Parameters" correspond to the mean true effect (μ), the extent of effect heterogeneity (τ^2), and p_1.

We also consider a Four-Parameter Selection Model. Our 4PSM adds another category to the 3PSM model: positive and insignificant estimates. The respective categories then become (i) $\left(\hat{\beta}_i/SE_i\right) \geq 1.96$; (ii) $0 \leq \left(\hat{\beta}_i/SE_i\right) < 1.96$; and (ii) $\left(\hat{\beta}_i/SE_i\right) < 0$.[1] The associated relative publication probabilities are equal to 1, p_1, and p_2 (see Panel B of FIGURE 1); with μ, τ^2, p_1, and p_2 accounting for the Four-Parameters. We use R's *weightfunct* package to estimate 3PSM and 4PSM. When the relative probabilities of being published are equal to one (i.e., no publication selection), these models collapse to the RE model.

[1] Hedges and Vevea (1996) estimate a 5PSM with the following four categories: (i) $\left(\hat{\beta}_i/SE_i\right) \geq 1.64$, (ii) $\left(\hat{\beta}_i/SE_i\right) < 0$, (iii) $0 \leq \left(\hat{\beta}_i/SE_i\right) < 0.84$, and (iv) $0.84 \leq \left(\hat{\beta}_i/SE_i\right) < 1.64$.

AK1 and AK2. Similar to 3PSM and 4PSM are two new estimators from Andrews & Kasy (2019). Like 3PSM and 4PSM, these models categorize estimated effects into groups with different probabilities of being published. The AK1 model groups estimates into significant and insignificant estimates without respect to sign: (i) $\left|\hat{\beta}_i/SE_i\right| \geq 1.96$; and (ii) $\left|\hat{\beta}_i/SE_i\right| < 1.96$. Andrews & Kasy refer to this as the "symmetric selection" case (see Panel A of FIGURE 2). The relative probability that a significant estimate is published is fixed at 1, while estimates that are insignificant are published with probability p_1.

Andrews & Kasy (2019) propose another estimator that recognizes that the sign of the estimated effect may also affect selection. The AK2 estimator allocates estimates into four groups: (i) $\left(\hat{\beta}_i/SE_i\right) \geq 1.96$, (ii) $\left(\hat{\beta}_i/SE_i\right) < -1.96$, (iii) $-1.96 \leq \left(\hat{\beta}_i/SE_i\right) < 0$, and (iv) $0 \leq \left(\hat{\beta}_i/SE_i\right) < 1.96$. These have relative publication probabilities equal to 1, p_1, p_2, and p_3 (see Panel B of FIGURE 2). Andrews & Kasy call this the "asymmetric selection" case. Because the p-values produced by AK1 and AK2 are based on t-statistics, they require four and six observations, respectively, in order to obtain p-values for all the parameter estimates. This can be a problem for meta-analyses with very small samples, such as is common in medicine. We use the programming code that accompanies Andrews & Kasy (2019) to obtain our AK1 and AK2 estimates.

Weighted Average of the Adequately Powered-Weighted Least Squares hybrid estimator (WAAP). The Weighted Average of the Adequately Powered-Weighted Least Squares hybrid estimator was introduced in Stanley, Doucouliagos, & Ioannidis (2017). Conceptually, this estimator chooses a subset of the N estimates $\hat{\beta}_i$ that are "adequately powered", defined as coming from regression equations having a power of at least 80%. Weighted Least Squares (weights = $\frac{1}{SE_i^2}$) is used to estimate Equation (2) in order to obtain an initial estimate of μ.

To determine whether a particular estimate comes from an "adequately powered" regression equation, the WAAP estimator determines a threshold value, δ, for the effect standard error:

(3) $\delta = \frac{|\hat{\mu}|}{2.8}$,

where $\hat{\mu}$ is the WLS estimate of μ in Equation (2) based on the full sample of *N* estimated effects. Note that this initial estimate of μ does not correct for publication bias. WAAP then selects all the $\hat{\beta}_i's$ for which $SE_i < \delta$. Let $M \leq N$ of the $\hat{\beta}_i's$ satisfy this criterion. It then uses WLS to re-estimate Equation (2) using only the *M* estimates (the "adequately powered" estimates) to obtain a revised estimate of μ. A problem can arise when there too few effect estimates that are adequately powered. If there are fewer than two adequately powered effect estimates, the WAAP estimator uses the WLS estimate from the full sample of *N* estimated effects.

PET-PEESE (PP). PET-PEESE stands for Precision Effect Test and Precision Effect Estimate with Standard Error (Stanley & Doucouliagos, 2012). The PP estimator proceeds in two steps. The first step estimates a publication-corrected variant of Equation (2) using WLS:

(4.a) $\hat{\beta}_i = \mu + \rho \cdot SE_i + \varepsilon_i$, *i = 1,2,...,N.*

with weights equal to $\frac{1}{SE_i^2}$. It then tests whether $\mu = 0$. If it fails to reject this hypothesis, then PP takes $\hat{\mu}$ as an estimate of μ. If it rejects $\mu = 0$, it then estimates

(4.b) $\hat{\beta}_i = \mu + \rho \cdot SE_i^2 + \varepsilon_i$, *i = 1,2,...,N.*

The estimate of μ from (4.b) then becomes the updated PP estimate of μ. Following Stanley (2017)'s recommendation, we use one-tailed test when testing $\mu = 0$.

Endogenous Kink (EK). Bom & Rachinger (2019) recently proposed a modification to the PET-PEESE model. The modification concerns a nonlinearity between the size of the bias due to publication selection and the standard error. When μ is nonzero there is no publication

selection when *SE* is very small because all or virtually all estimates are statistically significant. As *SE* increases, the degree of publication selection increases. This induces a non-linearity in the relationship between bias and standard error. This nonlinearity is the reason why Stanley & Doucouliagos (2012) propose including SE^2 in Equation (4.b).

As an alternative, Bom & Rachinger (2019) propose the following kinked regression specification:

(5) $\hat{\beta}_i = \mu + \rho \cdot [SE_i - a] I_{SE \geq a} + \varepsilon_i$, *i = 1,2,...,N.*

where $I_{SE \geq a}$ is a dummy variable that takes the value 1 whenever *SE* is larger than a cutoff point *a.* This induces a kink at $SE = a$. To determine a, Bom & Rachinger (2019) follow a two-step procedure. They first estimate μ as if one was implementing the first stage of the PET-PEESE procedure.

Assuming the estimated effect is positive, they then calculate the lower bound of a 95% confidence interval around $\hat{\mu}$ where the standard error is derived from a RE model (to accommodate effect heterogeneity): $\hat{\mu} - 1.96 \cdot \sqrt{SE_i^2 + \hat{\tau}^2}$. The cutoff value *a* is the value of *SE* that satisfies the equality $\hat{\mu} - 1.96 \cdot \sqrt{SE_i^2 + \hat{\tau}^2} = 1.96 \cdot SE_i$. Below a, most estimates of μ are likely to be statistically significant and thus unaffected by publication selection. Beyond a, publication selection is likely to become an increasing problem, causing the bias to be linearly related to *SE.* To estimate the EK model, we use programming code provided by Bom and Rachinger.

3. THE SIMULATION ENVIRONMENTS

To assess the eleven estimators above, we reproduce the simulation designs from four recently published studies: Stanley, Doucouliagos, & Ioannidis (2017), Alingahi & Reed (2018), Bom & Rachinger (2019), and Carter et al. (2019a). We chose to work with multiple simulation environments in light of Carter et al. (2019a, page 117)'s assessment of previous research:

> 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.

Working with multiple simulation environments allows us to determine the sensitivity of our results to alternative experimental designs.

Our choice of simulation environments was made to ensure that we covered scenarios of interest to multiple disciplines. Stanley, Doucouliagos, and Ioannidis (2017) was published in *Statistics in Medicine.* Carter et al. (2019a) was published in *Advances in Methods and Practices in Psychological Science.* Alinaghi & Reed (2019) and Bom & Rachinger (2019) were recently published in *Research Synthesis Methods.* Each of the simulation designs are briefly described below. More extensive discussions can be found in the original articles.

Stanley, Doucoucliagos, & Ioannidis (2017). SD&I consider two scenarios where researchers are interested in determining the effect of a given treatment, $treat = \{0,1\}$. In the "Log Odds Ratio" scenario, primary studies track the effect of a treatment on a binary indicator of "success". Individual observations are simulated such that the probability of "success" (Y=1) is 10% for the control group, and (10% + a fixed effect + a mean zero, random component) for the treatment group. Effect heterogeneity is regulated by the variance of the random component, σ_h^2.

Primary studies estimate a logistic regression to determine the effect of the treatment on Prob(Y=1). The parameter of interest is the coefficient on $treat$, α_1. Each study produces a single estimated effect. Variation in the standard error of the estimated effects across studies is generated by allowing the primary studies to have different numbers of observations. The

mean effect of treatment across all studies, α_1, equals 0.0, 0.30, or 0.54, depending on the experiment. Sample sizes for the simulated meta-analyses vary across experiments and are pre-determined to consist of 5, 10, 20, 40, or 80 estimated effects. In the absence of publication selection, a regression of the estimated treatment effects on a constant should produce an unbiased estimate of α_1 in any given meta-analysis sample.

Publication selection consists of two regimes: no publication selection, or 50% publication selection. Under 50% publication selection, estimates are sequentially evaluated for inclusion in the meta-analyst's sample. Each estimate has a 50% chance of being "selected". If it avoids selection, the estimate is "published" without consideration to its sign and statistical significance. If selected, the estimate is "published" if it is positive and significant. If not, new estimates are generated until a positive and significant estimate is found. This continues until the meta-analyst's sample attains its pre-determined size (see APPENDIX 1, Panel A).

In the second scenario, "Cohen's *d*", primary studies estimate the effect of a treatment, but this time the dependent variable is continuous. The difference in outcomes between the treatment and control groups is equal to a fixed effect, α_1, plus a random component that differs across studies. Effect heterogeneity is introduced through this random component, which is regulated by the parameter σ_h^2.

Each primary study calculates Cohen's *d*, which is the standardized difference in the mean outcome values across the two groups. The mean value of *d* across studies is set equal to either 0 or 0.5, depending on the experiment. Differences in the standard errors of *d* are generated by allowing the simulated primary studies to have different sample sizes. In the absence of publication selection, a regression of the estimated treatment effects on a constant will produce an unbiased estimate of the population mean of *d*. Sample sizes for the simulated meta-analyses are pre-determined to consist of 5, 10, 20, 40, or 80 estimated effects, depending on the experiment. The Cohen's *d* experiments include the no publication selection and 50%

publication selection scenarios used for the Log OR scenario, plus one more: 75/100% publication selection. Under 75/100% publication selection, positive and statistically significant estimates are selected with probability 75%, but 100% of the estimates are restricted to be positive (see APPENDIX 1, Panel B).

Alinaghi & Reed (2018). A&R study univariate regression models where a variable *X* affects a continuous variable *Y*. The parameter of interest is the coefficient on *X*. In the "Random Effects" data environment, each study produces one estimate and the population effect differs across studies. The coefficient on *X* equals a fixed component, α_1, plus a random component that is fixed within a study but varies across studies. The overall mean effect of *X* on *Y* is given by α_1. To estimate α_1, the meta-analyst regresses the study specific estimates on a constant. In the absence of publication selection, the resulting estimate will be unbiased.

A distinctive feature of A&R's experiments is that they fix the size of the sample of estimated effects before publication selection, rather than after. The size of the meta-analyst's sample is thus determined endogenously, and is affected by the size of the effect. For example, very large population effects will be subject to relatively little publication selection as most estimates will satisfy the selection criteria, whether it be statistical significance or correct sign.

Another distinctive feature of A&R's experiments is that they separate statistical significance from the sign of the estimated effect as criteria for selection. Other studies commonly combine these two, assuming a mechanism that selects estimates that are both positive and statistically significant. A&R's experiments accommodate the fact that these two criteria have different, sometimes conflicting, consequences for estimator performance. All significant/correctly-signed estimates are "published", while insignificant/wrong-signed estimates only have a 10% chance of getting published.

A&R design their simulations to be representative of meta-analyses in economics and business. These typically have samples of several hundred estimates and substantial effect

heterogeneity. In addition to the "Random Effects" data environment described above, A&R also construct a "Panel Random Effects" data environment, where each study has 10 estimates. This models the fact that the overwhelming share of meta-analyses in economics and business have multiple estimates per study. Effect estimates and standard errors are simulated to be more similar within studies than across studies. Publication selection targets the study rather than individual estimates. To be included in the meta-analyst's sample, a study must have at least 7 out of the 10 estimates be significant/correctly signed.

Bom & Rachinger (2019). B&R consider univariate regression environments where researchers are interested in estimating the effect of a variable X_1 on a dependent variable Y, represented by the parameter α_1. Variation in the standard errors of estimated effects is accomplished by allowing sample sizes to differ across primary studies. Effect heterogeneity is introduced via an omitted variable (X_2) that is correlated with X_1. The coefficient on the omitted variable, α_2, is randomly distributed across studies with mean zero and variance σ_h^2. Individual estimates of α_1 will be biased for nonzero values of α_2. In the population of all studies, the omitted variable bias averages out. However, publication selection induces a bias in the meta-analyst's sample when selection depends on the sign and significance of $\hat{\alpha}_1$.

The experiments are designed to produce 5, 10, 20, 40, or 80 "studies" for a given simulated meta-analysis, with each study consisting of one estimated effect. In the absence of publication selection, the regression on a constant produces an unbiased estimate of α_1, where α_1 equals either 0 or 1 depending on the experiment. Publication selection consists of four regimes: no publication selection, 25%, 50%, and 75% publication selection. The publication selection algorithm is modelled after SD&I's 50% publication selection algorithm (see Panel A of APPENDIX 1).

Carter et al. (2019a). In the simulation environment of CSG&H (for Carter, Schönbrodt, Gervais, and Hilgard), primary studies estimate the effect of a treatment using Cohen's *d* as

their measure of effect. The difference in outcomes for the treatment and control groups is equal to a fixed effect, α_1, plus a random component that differs across studies. Effect heterogeneity is introduced through this random component, which is regulated by the parameter σ_h^2. The mean value of *d* takes on four values depending on the experiment: 0, 0.2, 0.5, and 0.8. Differences in the standard errors of *d* for a given experiment are generated by allowing the simulated primary studies to have different sample sizes.

CSG&H introduce two types of distortions in the research environment. They employ a publication selection algorithm in which the probability of estimates being "published" depends nonlinearly on both the sign of the estimated effect and its p-value. They construct three different publication selection regimes which they call "No Publication Bias", "Medium Publication Bias", and "Strong Publication Bias". These are obtained by altering the parameters of the publication selection algorithm. They also simulate four different types of "questionable research practices" (QRPs): (a) optional removal of outliers, (b) optional selection between two dependent variables, (c) optional use of moderators, and (d) optional stopping. Finally, CSG&H also construct experiments in which the simulated meta-analysis samples take on four different sizes: 10, 30, 60, and 100.

TABLE 2 reports the number of experiments for each of the four simulation environments, categorized by number of estimates included in the meta-analyst's sample ("Sample Size") and the extent of measured effect heterogeneity ("I^2"). We calculate I^2 as:

$$(6) \qquad I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2},$$

where $\hat{\tau}^2$ is the estimate of effect heterogeneity using the restricted maximum likelihood method, and

$$(7) \qquad \hat{\sigma}^2 = \frac{\sum w_i(N-1)}{(\sum w_i)^2 - \sum w_i^2},$$

$w_i = 1/SE_i^2$, and N is the number of estimates in the meta-analyst's sample. I^2 takes values between 0 and 100%. I^2 is often interpreted as a measure of the share of effect size variance that is due to heterogeneity in true effects in the population. However, Augusteijn et al. (2019) demonstrate, that it is affected by publication selection. The effect of publication selection can be large, and can either increase or decrease the value of I^2. Our simulations calculate I^2 post-publication selection. Whether that vitiates the usefulness of I^2 in the selection of estimators is an empirical question.

In order to induce greater overlap in the simulation environments, we added simulations to the SD&I (2017), B&R (2019), and CSG&H (2019a) experimental designs that allow for larger sample sizes. These are yellow-highlighted in the table. This resulted in a total of 1,620 experiments, where an experiment is defined as a unique set of parameters determining (i) effect size, (ii) effect heterogeneity, (iii) publication selection, (iv) sample size, and (v) (for CSG&H, 2019a) questionable research practices. This compares favorably with previous studies (see TABLE 1). Details about the experiments are reported in APPENDIX 2.

4. THE PERFORMANCE MEASURES

We evaluate estimators on three performance measures: (i) Bias, (ii) Mean Squared Error (MSE), and (iii) 95% Coverage Rates. With respect to bias, the average bias for any given experiment *k* is calculated by

$$Bias_k = \left(\frac{1}{R_k}\right)\sum_{i=1}^{R_k}(Estimated\ Effect_{ki} - True\ Effect_k),$$

where $R_k$is the total number of iterations for that experiment (typically 3,000). Note that $Bias_k$ can be positive or negative. When aggregating over experiments to obtain a summary measure of performance, we calculate the average of absolute values, $|Bias| = \left(\frac{1}{R}\right)\sum_{k=1}^{R}|Bias_k|$, , where *R* is the total number of experiments included in the evaluation. "Best estimator" with respect to bias is defined as the estimator with the smallest value of average $|Bias|$.

MSE for a given experiment k is calculated by

$MSE_k = \left(\frac{1}{R_k}\right) \sum_{i=1}^{R_k} (Estimated\ Effect_{ki} - True\ Effect_k)^2$.

When used as a summary measure of performance, it is calculated by $MSE = \left(\frac{1}{R}\right) \sum_{k=1}^{R} MSE_k$. "Best estimator" with respect to MSE is defined as the estimator with the smallest value of MSE.

$95\%\ Coverage\ Rate_k$ calculates the percentage of iterations for which the estimated effect's 95% confidence interval contains the true effect for a given experiment k. The corresponding summary measure averages the individual $95\%\ Coverage\ Rate_k$ values over the respective set of experiments. "Best estimator" with respect to coverage rates is defined as the estimator that produces an average coverage rate closest to 95%.

5. RESULTS

Relative performance differs across criteria. TABLE 3 ranks the performance of the estimators for all 1,620 experiments. Results are separated by performance measure. The purpose of this table is not to demonstrate overall superiority for any given estimator. As we shall show below, aggregating results across simulation environments is perilous. The main purpose of this table is to note that estimators that dominate on one criterion may perform relatively poorly on another.

For example, on the dimension of bias, Bom & Rachinger (2019)'s Endogenous Kink (EK) estimator produces the lowest overall, mean absolute bias ("|Bias|"). However, it is dominated by Stanley, Doucouliagos, and Ioannidis (2017)'s WAAP estimator when it comes to mean squared error ("MSE"); and Andrews & Kasy (2019)'s "asymmetric selection" estimator (AK2) with respect to 95% coverage rates. The table color-codes the three estimators that perform best on the respective criteria to facilitate comparison. While relative positions in the table are subject to randomness, the differences in most cases are substantial.

The other purpose of TABLE 3 is to highlight the poor performance of all the estimators when it comes to coverage rates. While AK2 may be the "best" performing estimator on this dimension, its mean coverage rate of 81.4% is well below the expected 95%. Most of the estimators have coverage rates below 70%. This should cause researchers to question the reliability of any hypothesis testing about effect sizes that is performed in meta-analyses that use these estimators. Because of the poor coverage rate of all the estimators with respect to coverage rate, our subsequent analysis ignores this dimension and focuses on bias and MSE.

Relative performance differs across simulation environments. We next focus on the sensitivity of results to simulation environment. TABLE 4 collects the results from all 1,620 experiments and breaks them out according to each of the four simulation environments. Panel A reports average results for bias, and Panel B reports results for MSE. While average results can be misleading because they can conceal much diversity, they can be still be useful for identifying general performance. In both panels, we are looking for consistency in relative performance across simulation environments.

In Panel A, simulations for three of the four simulation environments lead to the conclusion that the AK2 estimator is best, producing the smallest bias. However, in the CSG&H simulations, the AK2 estimator ranks 9th of eleven. The 3PSM estimator ranks second in the SD&I's simulations, but 8th in A&R's, 5th in B&R's, and 5th in CSG&H's simulations. These inconsistencies are not unusual. In Panel B, the AK2 estimator ranks 1st in the SD&I and A&R simulations with respect to smallest MSE. However, it ranks 6th in B&R's simulations, and 9th in CSG&H's. Other examples are easily identified in the results.

On the basis of TABLE 4 we conclude that the simulation environment one uses to assess performance makes a difference. An estimator that performs well in one study employing one type of simulation design may not perform well in another simulation

environment. As a result, any analysis of estimator performance must inevitably make a decision about which simulation environment best represents a given research situation.

This situation is worrisome. As quoted from Carter et al. (2019a) above, "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." The situation may not be quite this bleak. There may be indicators both in the meta-analyst's sample and in the type of question being analyzed that can provide assistance in selecting the most appropriate simulation environment. We give an example below of how this might work. In the meantime, our subsequent analyses will avoid combining results from different simulation environments.

Two determinants of relative performance are sample size and effect heterogeneity. It is well-known that estimator performance generally declines as effect heterogeneity increases and improves as the meta-analyst's sample size gets larger (Moreno et al., 2009; Stanley, 2017). Less well-known is that relative estimator performance is also affected by these factors. In this section we demonstrate both phenomena. We use the results from the CSG&H simulations to estimate the following regressions for each of the eleven estimators (j):

(8.a) $$Bias_{ij} = \beta_0 + \beta_{SampleSize} \cdot (SampleSize)_{ij} + \beta_{I-squared} \cdot (I^2)_{ij} + \varepsilon_{ij} \ ,$$

and

(8.b) $$MSE_{ij} = \beta_0 + \beta_{SampleSize} \cdot (SampleSize)_{ij} + \beta_{I-squared} \cdot (I^2)_{ij} + \varepsilon_{ij} \ .$$

Regressions were estimated using OLS with bootstrapped t-statistics to obtain p-values. Each regression used the Bias/MSE results for a given estimator j. The respective samples were constructed from the individual results of the 756 experiments in the CSG&H simulations (see Panel D of TABLE 2).

TABLE 5 presents the results. They provide strong evidence that Bias and MSE performance decline as effect heterogeneity (I^2) increases. With only one exception, the coefficient on the I^2 term is positive and significant in both the Bias and MSE regressions for each of the eleven estimators. The exception is the coefficient for I^2 in the MSE regression for the p-curve estimator (pC). Sample size is also strongly associated with MSE performance. Sample size is negatively and significantly associated with MSE for each of the eleven estimators. The evidence for sample size affecting bias is not as strong. Still, nine of the eleven estimated coefficients are negative, with five of eleven negative and significant at the 5-percent level.

While TABLE 5 documents changes in absolute estimator performance, TABLE 6 presents evidence of changes in relative performance. Once again we use the CSG&H simulation results and focus on bias and MSE. We divide the 756 CSG&H experiments into 21 separate cells depending on sample size (10, 30, 60, 100, 200, 400, 800) and effect heterogeneity ($I^2 \leq 0.25$, $0.25 < I^2 \leq 0.75$, $0.75 < I^2$). Panel D of TABLE 2 reports the number of experiments for each sample size/ I^2 cell.

For both Bias and MSE, we identify the top two estimators in the cell for smallest sample size (10) and effect heterogeneity (low I^2). For Bias, these are the AK1 and 4PSM estimators. For MSE, they are AK1 and 3PSM. We then track the relative position of these estimators as sample size and effect heterogeneity increases. The respective estimators are color-coded to facilitate tracking across cells.

TABLE 6 clearly reveals that there is substantial movement in the relative rankings of the estimators as sample size and effect heterogeneity change. In some cases, the change in relative ranking is dramatic. When sample size = 10, the 4PSM estimator ranks 2nd and 1st on Bias, respectively, for low and moderate effect heterogeneity. It falls to 9th when effect

heterogeneity is high. In other cases, relative performance is relatively stable: Across all sample sizes, AK1 is either ranked 1st or 2nd in terms of lowest MSE.

The table demonstrates two things. It underscores a point made previously that no estimator dominates in all research settings. However, it also suggests that there may be circumstances where one estimator is generally preferred. For example, if a researcher is interested in estimator efficiency and works in an area where effect heterogeneity is expected to be high, *and* if the researcher is convinced that the CSG&H simulation environment captures the key elements of their research situation, then TABLE 6 suggests that AK1 may be the best estimator for their analysis. However, the TABLE 6 results are based on average performance within a given {sample size, effect heterogeneity} cell. As demonstrated previously, averages can conceal much variation. The next section illustrates how further investigation can lead to a more definitive conclusion regarding "best" estimator.

6. AN EXAMPLE OF HOW SIMULATION EXPERIMENTS CAN GUIDE THE SELECTION OF A "BEST" ESTIMATOR

Previous sections demonstrated that there is no superior estimator for all research situations. "Best" is conditional on performance measure, and depends on observable characteristics of the meta-analyst's sample such as sample size and effect heterogeneity. It also can depend on unobservable characteristics such as the type of publication selection (statistical significance, correct sign, both), the extent of publication selection, and other factors such as assorted questionable research practices (QRPs). By conditioning on observables and investigating performance over unobservables, one can study the relative performance of estimators and use the results to guide estimator selection for use in a given research situation. This section demonstrates how this can be done.

Suppose a meta-analyst is studying the extant empirical literature on a given "effect", measured by Cohen's *d.* They collect a sample of 100 estimates, and initial analysis indicates a high degree of effect heterogeneity ($I^2 > 0.75$). While they are unsure whether publication

selection is a problem, if it does exist, they believe selection would depend on both correct sign and statistical significance. Looking over the alternatives, it is their experienced judgment that the CSG&H simulation environment best captures the salient aspects of their research situation. However, they do not have strong priors about the size of the effect, the severity of publication selection, nor the extent of QRPs. While they would like to have an estimator that minimized bias and produced accurate coverage rates, their main priority is choosing an estimator that is efficient. We show how simulation results can be used to guide that selection.

TABLE 7 reports the individual experimental results for sample size = 100/ High I^2. There are a total of 30 experimental results (cf. TABLE 2), covering a wide range of effect sizes {0, 0.2, 0.5, 0.8}, severities of publication selection {No, Medium, Strong}, and QRP behaviors {None, Medium, High} (see APPENDIX 2). We suppose the meta-analyst is interested in not just average MSE performance, but also the variation of MSE values across situations. Since they do not know which of the respective experiments best represents their research situation, they want to avoid an estimator that occasionally produces a bad result, even if it does well on average.

The top part of the table reports the individual MSE experimental results. We yellow-highlight the minimum MSE value in each experiment. The bottom part of the table reports the overall average value across all 30 experiments, along with the minimum and maximum MSE values. Of the eleven estimators, all but two of them (WAAP and PP) are "best" in at least one experiment. This again highlights the fact that no estimator is best in all research situations.

Given that the researcher doesn't know which simulated situation best represents their actual research situation, they first consider the estimator with the lowest overall average MSE. That is the AK1 estimator. It has an overall average value of 0.020. The next best estimator is the WAAP, with an overall average of 0.027. AK1 also takes on a relatively narrow range of values across the 30 experiments. Its minimum value is 0.001, and its maximum value is 0.081.

This compares favorably with most of the other estimators, but not all. For example, Bom & Rachinger's EK estimator, while producing a slightly larger overall average value of MSE (0.028), takes on a narrower set of values (minimum = 0.008, maximum = 0.052). The WAAP and PP estimators have similar characteristics.

With respect to AK1, it is worth noting that simulations will tend to be biased towards selection models, because selection models have been designed to capture the very kinds of behaviors built into selection algorithms. This is not necessarily a bad thing. However, to the extent that actual publication selection behavior differs from simulated selection behavior, results may overstate the performance of selection models in real world datasets.

The researcher's choice comes down to a trade-off between mean and dispersion, a choice that is complicated by the fact that randomness in the simulation process cautions against attaching too much significance to small numerical differences. We propose one possible solution, with the researcher choosing the AK1 estimator as best (yellow-highlighted), while also choosing one or two other estimators (WAAP, PP, EK; highlighted in blue) for robustness checking.

7. CONCLUSION

The subject of meta-analysis (MA) estimator performance has received much attention in the literature (Alinaghi & Reed, 2018; Bom & Rachinger, 2019; Carter et al., 2019a; Hedges & Vevea, 1996; McShane, Böckenholt, & Hansen, 2016; Moreno et al., 2009; Rücker, Carpenter, & Schwarzer, 2011; Simonsohn, Nelson, & Simmons, 2014; Stanley. 2017; Stanley & Doucouliagos, 2014; Stanley, Doucouliagos, & Ioannidis, 2017; van Aert, Wicherts, & van Assen, 2016; van Assen, van Aert, & Wicherts, 2014). A goal of many of these studies has been to find a "best" estimator. However, there is an increasing awareness that no single estimator is "best" in all circumstances (Carter et al., 2019a). This study demonstrates how

Monte Carlo experiments can be used to guide researchers in choosing the most appropriate estimator for their specific research situation.

We demonstrate that sample size and effect heterogeneity exert a substantial influence on the relative performance of MA estimators. Accordingly, researchers can use that information to match the sample size and effect heterogeneity that characterize their specific research situation with corresponding results from Monte Carlo experiments. There is no guarantee that even within a given sample size/effect heterogeneity category that one estimator will be "best". However, we demonstrate that possibility by using results on MSE performance for the Carter et al. (2019a) simulation environment.

Our results highlight the importance of simulation design for the evaluation of MA estimators. Experimental results can differ markedly depending on the simulation environment used to study estimators. Relatively little attention has been directed towards justifying the use of one simulation design versus others. While it is true that many important elements are unobservable, such as the true mean effect size, the type and severity of publication selection, and the presence and extent of questionable research practices, there are observable features that can be assessed. For example, Alinaghi & Reed (2018) examine the characteristics of their post-selection meta-analysis samples such as the distribution of t-statistics, the share of significant estimates, and the estimate of effect heterogeneity (I^2), and claim that their simulated meta-analysis samples look "realistic." Surely more could be done along these lines.

Relatedly, more needs to be understood about which aspects of simulation designs are important for estimator performance. For example, what is it about the simulation environment of Carter et al. (2019a) that causes the AK2 estimator to perform relatively poorly compared to its performance in other simulation environments (see TABLE 4)? The investigation of these and other related questions would be greatly facilitated if researchers made their programming

code available for others to reproduce and extend their experiments. Towards that end, all of our programming code is posted online at https://osf.io/pr4mb/.

A final contribution of our study is that all of our experimental results are also available for inspection at https://osf.io/pr4mb/. TABLE 7 presented the results of 30 Monte Carlo experiments from the Carter et al. (2019a) simulation environment for sample sizes of 100 and high effect heterogeneity. The online results allow researchers to explore other scenarios that may be more relevant for their particular research situations.

HIGHLIGHTS

- Despite much previous research, meta-analysts do not have much guidance when it comes to selecting a “best” estimator
- This study shows how Monte Carlo experiments can be used to select the “best” estimator for a given research situation
- We compare eleven estimators commonly used in medicine, psychology, and the social sciences
- The estimators are evaluated on three performance measures: bias, mean squared error (MSE), and coverage rates
- We conduct 1,620 individual experiments, where an experiment is defined by a unique combination of sample size, effect heterogeneity, effect size, publication selection mechanism, and other research characteristics
- Estimators that are relatively good on one performance measure may perform relatively poorly on another
- Sample size and effect heterogeneity are important determinants of relative estimator performance
- We demonstrate how the observable characteristics of sample size and effect heterogeneity can guide the meta-analyst to select the most appropriate estimator for their research circumstances

REFERENCES

Alinaghi, N. & Reed, W.R. (2018). Meta-analysis and publication bias: How well does the FAT-PET-PEESE procedure work? *Research Synthesis Methods*, 9(2), 285-311.

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8), 2766-94.

Augusteijn, H. E., van Aert, R., & van Assen, M. A. (2019). The effect of publication bias on the Q test and assessment of heterogeneity. *Psychological methods*, 24(1), 116.

Bom, P. R., & Rachinger, H. (2019). A kinked meta-regression model for publication bias correction. *Research Synthesis Methods*, 10, 497–514.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019a). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science,* 2(2), 115-144.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019b). Source code to accompany Carter et al. (2019a). Retrieved from osf.io/rf3ys.

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

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

Harbord, R. M., Egger, M., & Sterne, J. A. (2006). A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine*, 25(20), 3443-3457.

Hedges, L. V., & Vevea, J. L. (1996). Estimating Effect Size Under Publication Bias: Small Sample Properties and Robustness of a Random Effects Selection Model. *Journal of Educational and Behavioral Statistics*, 21(4), 299-332.

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4(3), 245-253.

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

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, 11(5), 730-749.

Moreno, S. G., Sutton, A. J., Ades, A. 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(1), 2.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *JAMA*, 295(6), 676-680.

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 (2015-30): 1—40. http://dx.doi.org/10.5018/economics-ejournal.ja.2015-30

Rücker, G., Schwarzer, G., Carpenter, J. R., Binder, H., & Schumacher, M. (2011). Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics*, 12(1), 122-142.

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.

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, 8(5), 581-591.

Stanley, T.D., & Doucouliagos, H. (2012). *Meta-regression analysis in economics and business*. Routledge: Oxford.

Stanley, T.D., Doucouliagos, H., & Ioannidis, J. (2017). Finding the power to reduce publication bias. *Statistics in Medicine*, 36 (10), 1580–1598.

Stanley, T. D., Jarrell, S. B., & Doucouliagos, H. (2010). Could it be better to discard 90% of the data? A statistical paradox. *The American Statistician*, 64(1), 70-77.

Van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological methods*, 20(3), 293.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60(3), 419-435.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: sensitivity analysis using a priori weight functions. *Psychological methods*, 10(4), 428.

TABLE 1
Summary of Selected Monte Carlo Studies of Estimator Performance: Number of Experiments and Estimators Studied

Study	*Experiments*	*Estimators*
Stanley, Doucouliagos, & Ioannidis (2017)	180	RE, WLS, WAAP, PP
Alinaghi & Reed (2018)	74	WLS-FE, WLS-RE, PP
Bom & Rachinger (2019)	215	FE, RE, WAAP, PP, EK
Carter et al. (2019a)	432	TF, pC, pU, RE, 3PSM WAAP, PP
Hedges & Vevea (1996)	176	5PSM
McShane, Bockenhold, & Hansen (2016)	125	pC, pU, 3PSM
Moreno et al. (2009)	240	TF(FE-FE), TF(FE-RE), TF(RE-RE), FE, RE, FE-se, RE-se, D-se, FE-var, RE-var, D-var, Harbord, Peters, and Harbord-C
Reed (2015)	36	OLS, PET, PEESE, FE, WLS, RE
Rucker, Carpenter, & Schwarzer (2011)	36	TF, CSM, RE, LMA
Simonsohn, Nelson, & Simmons (2014)	30	TF, pC, FE
Stanley (2017)	120	WLS, FE, PP
Stanley & Doucouliagos (2014)	60	FE, RE, Top10, PEESE, PP, WLS-se, WLS-Quadratic, WLS-Cubic
van Aert, Wicherts, & van Assen (2016)	25	pC, pU, FE, RE
Van Assen, van Aert, & Wicherts (2014)	36	FE, TF, pU, TES
Our study	*1620*	*TF, pC, pU, RE, 3PSM, 4PSM, AK1, AK2, WAAP, PP, EK*

Estimators:
- 3PSM/4PSM/5PSM = Three-Parameter, Four-Parameter, and Five Parameter Selection Models
- AK1 = Andrews & Kasy (2019)'s "symmetric selection" model
- AK2 = Andrews & Kasy (2019)'s "asymmetric selection" selection
- CSM = Copas selection model (Copas, 1999)
- EK = Bom & Rachinger (2019)'s Endogenous Kink estimator
- FE = Fixed Effects

- FE-se, RE-se, and WLS-se/D-se/PET = Estimates the following model using FE, RE, and WLS: $\widehat{effect}_{\mathrm{i}} = \alpha + \beta se\left(\widehat{effect}\right) + \epsilon_{\mathrm{i}}$
- FE-var, RE-var, and PEESE/D-var/ = Estimates the following model using FE, RE, and WLS. $\widehat{effect}_{\mathrm{i}} = \alpha + \beta \mathrm{se}^2\left(\widehat{effect}\right) + \epsilon_{\mathrm{i}}$
- Harbord/Harbord-C = Harbord, Egger, & Sterne (2006)'s "Regression test for small-study effects" and variant
- LMA = Limit meta-analysis (Rucker et al., 2011).
- OLS = OLS regression of estimated effects on a constant.
- pC = p-curve
- pU = p-uniform
- Peters = Peters et al. (2006)'s "Regression test for funnel asymmetry"
- PP = PET-PEESE (Stanley and Doucouliagos, 2012)
- RE = Random Effects
- TES = Test for excess significance (Ioannidis & Trikalinos, 2007)
- TF/TF(RE-RE) = Trim and Fill with RE used for both the "trim" and "fill" components
- TF(FE-FE)/TF(FE-RE) = Trim and Fill with variants depending on whether FE or RE is used for the "trim" and "fill" components, respectively
- Top10 = Estimator which uses only the most precise 10% of estimates (Stanley et al., 2010)
- WLS/WLS-FE = Weighted Least Squares with weights $\left(\frac{1}{SE_i^2}\right)$
- WLS-RE = Weighted Least Squares with weights $\left(\frac{1}{SE_i^2+\tau^2}\right)$
- WLS- Quadratic= Estimates the following model using WLS: $\widehat{effect}_{\mathrm{i}} = \alpha + \beta_1 se\left(\widehat{effect}\right) + \beta_2 \mathrm{se}^2\left(\widehat{effect}\right) + \epsilon_{\mathrm{i}}$
- WLS-Cubic = Estimates the following model using WLS: $\widehat{effect}_{\mathrm{i}} = \alpha + \beta_1 se\left(\widehat{effect}\right) + \beta_2 \mathrm{se}^2\left(\widehat{effect}\right) + \beta_3 \mathrm{se}^3\left(\widehat{effect}\right) + \epsilon_{\mathrm{i}}$
- WAAP = Stanley, Doucouliagos, and Ioannidis (2017)'s Weighted Average of the Adequately Powered-WLS-FE hybrid estimator.

TABLE 2
Number of Experiments by Sample Size and Extent of Effect Heterogeneity

A. Stanley, Doucouliagos, & Ioannidis (2017)

Sample Size	Low I^2 $I^2 \leq 0.25$	Moderate I^2 $0.25 < I^2 \leq 0.75$	High I^2 $0.75 < I^2$	Total
{5,10}	30	27	15	72
20	15	10	11	36
40	15	10	11	36
80	13	12	11	36
{100, 200, 400, 800}	51	49	44	144
Total	124	108	92	324

B. Alinaghi & Reed (2018)

Sample Size	Low I^2 $I^2 \leq 0.25$	Moderate I^2 $0.25 < I^2 \leq 0.75$	High I^2 $0.75 < I^2$	Total
$0 < SS \leq 100$	0	0	0	0
$100 < SS \leq 500$	0	0	13	13
$500 < SS$	0	1	22	23
Total	0	1	35	36

C. Bom & Rachinger (2019)

Sample Size	Low I^2 $I^2 \leq 0.25$	Moderate I^2 $0.25 < I^2 \leq 0.75$	High I^2 $0.75 < I^2$	Total
{5, 10}	20	27	65	112
20	5	18	33	56
40	5	17	34	56
80	5	17	34	56
{100, 200, 400, 800}	20	68	136	224
Total	55	147	302	504

D. Carter et al. (2019a)

Sample Size	Low I^2 $I^2 \leq 0.25$	Moderate I^2 $0.25 < I^2 \leq 0.75$	High I^2 $0.75 < I^2$	Total
10	33	68	7	108
30	29	57	22	108
60	28	54	26	108
100	28	50	30	108
{200, 400, 800}	81	147	96	324
Total	199	376	181	756

NOTE: The table lists the number of experiments for each {sample size, effect heterogeneity} category, by simulation environment. An experiment is defined as a unique set of parameters determining (i) effect size, (ii) effect heterogeneity, (iii) publication selection, (iv) sample size, and (v) (for Carter et al., 2019) questionable research practices (see APPENDIX 2). Each experiment consists of 3000 simulated meta-analyses. I^2 measures the share of effect size variance that is due to heterogeneity in true effects. It is based on $\hat{\tau}^2$, which we, following Carter et al. (2019a), estimate using restricted maximum likelihood (REML) (see Equation 6 in the text and the associated discussion).

TABLE 3
Comparison of Estimator Performance: All Experiments

Performance Criterion					
|Bias|		**MSE**		**95% Coverage Rate**	
EK	0.076	WAAP	0.075	AK2	0.814
PP	0.081	PP	0.106	3PSM	0.761
AK2	0.083	EK	0.107	4PSM	0.708
4PSM	0.090	TF	0.110	EK	0.669
3PSM	0.101	AK1	0.120	PP	0.668
WAAP	0.109	AK2	0.136	WAAP	0.650
AK1	0.132	3PSM	0.140	AK1	0.633
TF	0.140	pU	0.160	TF	0.560
RE	0.216	4PSM	0.163	RE	0.453
pU	0.229	RE	0.195	pU	0.380
pC:	0.333	pC:	0.608	pC	NA

NOTE: The values in the table represent the average values of the respective performance measures across all 1,620 experiments. The three "best" performing estimators on the dimensions of bias, MSE, and coverage rates (EK, WAAP, and AK2) are color-coded to faciliate comparison across performance measures.

Estimators:
- 3PSM/4PSM = Three-Parameter/Four-Parameter Selection Models
- AK1 = Andrews & Kasy (2019)'s "symmetric selection" model
- AK2 = Andrews & Kasy (2019)'s "asymmetric selection" selection
- EK = Bom & Rachinger (2019)'s Endogenous Kink estimator
- pC = p-curve
- pU = p-uniform
- PP = PET-PEESE (Stanley and Doucouliagos, 2012)
- RE = Random Effects
- TF = Trim and Fill
- WAAP = Stanley, Doucouliagos, and Ioannidis (2017)'s Weighted Average of the Adequately Powered-WLS hybrid estimator

TABLE 4
Comparison of Estimator Performance across Simulation Environments

A. |Bias|

SD&I (2017)		*A&R (2018)*		*B&R (2019)*		*CSG&H (2019a)*	
AK2	0.031	AK2	0.200	AK2	0.071	PP	0.058
3PSM	0.036	EK	0.213	EK	0.089	WAAP	0.062
4PSM	0.040	PP	0.256	4PSM	0.099	AK1	0.064
PP	0.050	WAAP	0.263	PP	0.124	EK	0.071
EK	0.053	TF	0.284	3PSM	0.147	3PSM	0.080
AK1	0.060	4PSM	0.298	WAAP	0.187	TF	0.091
WAAP	0.083	AK1	0.390	TF	0.238	4PSM	0.095
TF	0.088	3PSM	0.468	AK1	0.262	pU	0.105
RE	0.107	RE	0.550	RE	0.361	AK2	0.107
pU	0.146	pC	1.530	pU	0.373	pC	0.114
pC	0.420	pU	1.556	pC	0.521	RE	0.150

B. MSE

SD&I (2017)		*A&R (2018)*		*B&R (2019)*		*CSG&H (2019a)*	
AK2	0.013	AK2	0.229	WAAP	0.171	AK1	0.011
3PSM	0.013	TF	0.244	pU	0.184	WAAP	0.016
AK1	0.016	4PSM	0.318	EK	0.250	3PSM	0.021
WAAP	0.024	AK1	0.363	PP	0.262	TF	0.021
TF	0.024	WAAP	0.423	TF	0.289	PP	0.021
PP	0.024	PP	0.456	AK2	0.324	EK	0.025
EK	0.026	3PSM	0.468	AK1	0.333	pU	0.025
RE	0.033	RE	0.484	4PSM	0.366	4PSM	0.028
pU	0.049	EK	0.567	3PSM	0.376	AK2	0.036
4PSM	0.144	pC	3.518	RE	0.502	RE	0.045
pC	1.209	pU	3.626	pC	0.836	pC	0.060

NOTE: The two panels rank the performance of the eleven estimators on the basis of their Bias and MSE performance, disaggregated by simulation environment. Estimators are ranked from "best" (least Bias, smallest MSE) to worst. Values in the tables are the average values for the respective performance measures and simulation environments.

TABLE 5
Sample Size and Effect Heterogeneity as Determinants of
Absolute Estimator Performance: CSG&H (2019a) Simulation Environment

Estimator	*Bias*		*MSE*	
	$\beta_{SampleSize}$	$\beta_{I-squared}$	$\beta_{SampleSize}$	$\beta_{I-squared}$
AK1	-0.0143* (0.0074)	0.1147*** (0.0068)	-0.0101*** (0.0018)	0.0240*** (0.0018)
4PSM	0.0112 (0.0116)	0.2214*** (0.0098)	-0.0160*** (0.0045)	0.0812*** (0.0040)
3PSM	0.0029 (0.0101)	0.1624*** (0.0088)	-0.0156*** (0.0034)	0.0536*** (0.0030)
WAAP	-0.0366*** (0.0091)	0.1163*** (0.0078)	-0.0235*** (0.0027)	0.0344*** (0.0024)
TF	-0.0150 (0.0121)	0.1555*** (0.0093)	-0.0121*** (0.0042)	0.0413*** (0.0033)
AK2	-0.0355** (0.0168)	0.1883*** (0.0122)	-0.0458*** (0.0099)	0.0676*** (0.0075)
PP	-0.0206*** (0.0069)	0.0868*** (0.0060)	-0.0443*** (0.0040)	0.0468*** (0.0037)
RE	-0.0222 (0.0178)	0.2180*** (0.0151)	-0.0139** (0.0083)	0.0927*** (0.0071)
EK	-0.0286*** (0.0058)	0.0125*** (0.0058)	-0.055*** (0.0041)	0.0399*** (0.0039)
pU	-0.0180 (0.0124)	0.1352*** (0.0121)	-0.0182*** (0.0050)	0.0575*** (0.0053)
pC	-0.0403*** (0.0151)	0.1360*** (0.0150)	-0.1140*** (0.0302)	-0.0025 (0.0318)

NOTE: The table reports the results of estimating Equations (8.a) and (8.b) in the text. Regressions were estimated using OLS with bootstrapped *t*-statistics to obtain *p*-values. Each regression used the Bias/MSE results for a given estimator *j*. The respective samples were constructed from the individual results of the 756 experiments in the Carter et al. (2019a) simulations. Bootstrap standard errors are reported in parentheses. When estimating the model we use SampleSize/1000. This transformation increases the size of $\beta_{SampleSize}$ by a factor of 1000, but leaves economic and statistical significance unchanged.

TABLE 6
The Relationship Between Relative Estimator Performance, Sample Size, and I^2 :
CSG&H (2019a) Simulation Environment

A. Sample Size = 10

Bias						*MSE*					
Low I^2		*Moderate I^2*		*High I^2*		*Low I^2*		*Moderate I^2*		*High I^2*	
AK1	0.028	4PSM	0.071	AK1	0.097	AK1	0.006	AK1	0.027	AK1	0.027
4PSM	0.033	3PSM	0.074	WAAP	0.104	3PSM	0.007	pU	0.042	TF	0.045
3PSM	0.035	PP	0.087	TF	0.110	4PSM	0.008	TF	0.043	WAAP	0.057
WAAP	0.040	AK1	0.088	AK2	0.119	TF	0.010	3PSM	0.043	3PSM	0.069
TF	0.042	EK	0.098	3PSM	0.146	WAAP	0.010	WAAP	0.046	RE	0.075
AK2	0.047	pU	0.107	EK	0.153	AK2	0.010	4PSM	0.049	AK2	0.078
PP	0.063	WAAP	0.112	PP	0.160	RE	0.018	RE	0.068	4PSM	0.093
RE	0.082	TF	0.127	4PSM	0.177	PP	0.021	PP	0.081	pU	0.114
pU	0.090	pC	0.147	RE	0.179	pU	0.023	EK	0.092	pC	0.164
EK	0.101	AK2	0.160	pU	0.253	EK	0.030	AK2	0.102	PP	0.209
pC	0.150	RE	0.188	pC	0.270	pC	0.278	pC	0.203	EK	0.220

B. Sample Size = 30

Bias						*MSE*					
Low I^2		*Moderate I^2*		*High I^2*		*Low I^2*		*Moderate I^2*		*High I^2*	
WAAP	0.012	PP	0.048	EK	0.094	WAAP	0.002	AK1	0.011	AK1	0.026
AK1	0.019	AK1	0.068	PP	0.106	AK1	0.002	pU	0.015	TF	0.041
TF	0.020	EK	0.071	AK1	0.115	TF	0.002	3PSM	0.020	WAAP	0.045
3PSM	0.026	3PSM	0.074	WAAP	0.116	3PSM	0.003	PP	0.020	3PSM	0.053
4PSM	0.026	pU	0.076	TF	0.144	4PSM	0.003	WAAP	0.020	PP	0.071
AK2	0.028	WAAP	0.078	3PSM	0.145	PP	0.004	4PSM	0.024	EK	0.077
PP	0.029	4PSM	0.079	4PSM	0.190	AK2	0.004	TF	0.024	pU	0.077
RE	0.049	pC	0.083	RE	0.202	RE	0.005	EK	0.030	4PSM	0.078
EK	0.073	TF	0.102	AK2	0.217	EK	0.011	AK2	0.036	pC	0.081
pU	0.081	AK2	0.113	pU	0.218	pU	0.014	pC	0.049	RE	0.084
pC	0.094	RE	0.181	pC	0.224	pC	0.077	RE	0.052	AK2	0.096

C. Sample Size = 60

Bias						*MSE*					
Low I^2		*Moderate I^2*		*High I^2*		*Low I^2*		*Moderate I^2*		*High I^2*	
WAAP	0.010	PP	0.050	EK	0.081	WAAP	0.001	AK1	0.009	AK1	0.021
TF	0.016	EK	0.065	PP	0.089	TF	0.001	pU	0.012	WAAP	0.033
AK1	0.017	AK1	0.066	WAAP	0.104	AK1	0.001	PP	0.012	TF	0.034
3PSM	0.022	WAAP	0.066	AK1	0.107	3PSM	0.002	WAAP	0.014	PP	0.040
4PSM	0.023	pU	0.073	3PSM	0.137	PP	0.002	EK	0.018	3PSM	0.041
PP	0.024	pC	0.073	TF	0.138	4PSM	0.002	3PSM	0.018	EK	0.044
AK2	0.026	3PSM	0.084	AK2	0.163	AK2	0.003	pC	0.018	AK2	0.050
RE	0.042	4PSM	0.090	4PSM	0.189	RE	0.004	4PSM	0.022	pU	0.065
EK	0.063	TF	0.102	RE	0.201	EK	0.007	TF	0.022	4PSM	0.065
pU	0.074	AK2	0.118	pU	0.205	pU	0.010	AK2	0.035	pC	0.068
pC	0.081	RE	0.180	pC	0.213	pC	0.049	RE	0.051	RE	0.075

D. Sample Size = 100

Bias						*MSE*					
Low I^2		*Moderate I^2*		*High I^2*		*Low I^2*		*Moderate I^2*		*High I^2*	
WAAP	0.009	PP	0.049	EK	0.073	WAAP	0.001	AK1	0.007	AK1	0.020
TF	0.015	WAAP	0.055	PP	0.086	TF	0.001	PP	0.009	WAAP	0.027
AK1	0.017	AK1	0.061	WAAP	0.104	AK1	0.001	pC	0.009	PP	0.028
AK2	0.021	EK	0.064	AK1	0.110	PP	0.001	pU	0.009	EK	0.028
3PSM	0.021	pC	0.066	3PSM	0.131	3PSM	0.001	WAAP	0.010	3PSM	0.034
PP	0.022	pU	0.068	AK2	0.148	4PSM	0.002	EK	0.013	TF	0.035
4PSM	0.022	3PSM	0.089	TF	0.149	AK2	0.002	3PSM	0.018	AK2	0.041
RE	0.041	TF	0.094	4PSM	0.179	RE	0.003	TF	0.019	4PSM	0.056
EK	0.060	4PSM	0.097	pU	0.196	EK	0.006	4PSM	0.021	pU	0.058
pU	0.071	AK2	0.108	pC	0.204	pU	0.009	AK2	0.026	pC	0.062
pC	0.074	RE	0.168	RE	0.218	pC	0.030	RE	0.046	RE	0.079

E. Sample Size = 200

Bias						*MSE*					
Low I^2		*Moderate I^2*		*High I^2*		*Low I^2*		*Moderate I^2*		*High I^2*	
WAAP	0.008	PP	0.048	EK	0.072	TF	0.001	AK1	0.006	EK	0.018
TF	0.013	WAAP	0.052	PP	0.089	WAAP	0.001	PP	0.007	AK1	0.019
AK1	0.016	AK1	0.060	WAAP	0.097	AK1	0.001	WAAP	0.008	PP	0.021
AK2	0.020	EK	0.063	AK1	0.109	PP	0.001	pC	0.008	WAAP	0.022
3PSM	0.020	pC	0.067	3PSM	0.132	3PSM	0.001	pU	0.009	3PSM	0.033
4PSM	0.021	pU	0.068	AK2	0.144	4PSM	0.001	EK	0.009	TF	0.034
PP	0.021	3PSM	0.091	TF	0.151	AK2	0.001	3PSM	0.017	AK2	0.036
RE	0.036	TF	0.095	4PSM	0.185	RE	0.002	TF	0.019	4PSM	0.055
EK	0.057	4PSM	0.100	pU	0.196	EK	0.005	4PSM	0.021	pU	0.056
pC	0.063	AK2	0.121	pC	0.207	pC	0.007	AK2	0.028	pC	0.061
pU	0.067	RE	0.167	RE	0.218	pU	0.007	RE	0.045	RE	0.078

F. Sample Size = 400

Bias						*MSE*					
Low I^2		*Moderate I^2*		*High I^2*		*Low I^2*		*Moderate I^2*		*High I^2*	
WAAP	0.008	PP	0.046	EK	0.070	TF	0.000	PP	0.005	EK	0.013
TF	0.013	WAAP	0.048	PP	0.091	WAAP	0.000	AK1	0.006	AK1	0.018
AK1	0.016	AK1	0.059	WAAP	0.097	AK1	0.001	WAAP	0.006	PP	0.018
3PSM	0.020	EK	0.061	AK1	0.107	PP	0.001	EK	0.007	WAAP	0.020
4PSM	0.020	pC	0.064	3PSM	0.139	3PSM	0.001	pC	0.007	3PSM	0.033
PP	0.021	pU	0.066	TF	0.150	4PSM	0.001	pU	0.008	TF	0.033
AK2	0.021	3PSM	0.087	AK2	0.158	AK2	0.001	3PSM	0.015	AK2	0.039
RE	0.036	4PSM	0.093	pU	0.187	RE	0.002	4PSM	0.017	pU	0.052
EK	0.056	TF	0.093	4PSM	0.193	EK	0.004	TF	0.018	4PSM	0.055
pC	0.061	AK2	0.115	pC	0.200	pC	0.006	AK2	0.026	pC	0.057
pU	0.065	RE	0.161	RE	0.222	pU	0.007	RE	0.042	RE	0.078

G. Sample Size = 800

Bias						*MSE*					
Low I^2		*Moderate I^2*		*High I^2*		*Low I^2*		*Moderate I^2*		*High I^2*	
WAAP	0.007	PP	0.046	EK	0.070	WAAP	0.000	PP	0.005	EK	0.010
TF	0.013	WAAP	0.047	PP	0.093	TF	0.000	EK	0.006	PP	0.017
AK1	0.015	AK1	0.058	WAAP	0.097	AK1	0.001	WAAP	0.006	AK1	0.017
4PSM	0.019	EK	0.060	AK1	0.107	PP	0.001	AK1	0.006	WAAP	0.018
3PSM	0.020	pC	0.064	3PSM	0.140	3PSM	0.001	pC	0.007	3PSM	0.032
PP	0.020	pU	0.066	TF	0.150	4PSM	0.001	pU	0.007	TF	0.033
AK2	0.021	3PSM	0.087	AK2	0.162	AK2	0.001	3PSM	0.015	AK2	0.040
RE	0.036	4PSM	0.093	pU	0.187	RE	0.002	4PSM	0.017	pU	0.052
EK	0.055	TF	0.094	4PSM	0.195	EK	0.004	TF	0.018	4PSM	0.056
pC	0.060	AK2	0.101	pC	0.201	pC	0.006	AK2	0.020	pC	0.058
pU	0.064	RE	0.161	RE	0.222	pU	0.006	RE	0.042	RE	0.078

NOTE: The panels above rank the performance of the eleven estimators on the basis of their Bias and MSE performance, disaggregated by {sample size, effect heterogeneity} categories. Estimators are ranked from "best" (least Bias, smallest MSE) to worst. Values in the tables are the average values for the respective performance measures and {sample size, effect heterogeneity} categories. For both Bias and MSE, the top two estimators in the cell for smallest sample size (10) and effect heterogeneity (low I^2) are identified by color-coding. For Bias, these are the AK1 and 4PSM estimators. For MSE, they are AK1 and 3PSM. The relative position of these estimators are then tracked as sample size and effect heterogeneity increases.

TABLE 7
Comparison of MSE Performance: Sample Size = 100, High I^2, CSG&H (2019a) Simulation Environment

{Effect Size, I^2, QRP, Publication Selection}	*Estimators*										
	TF	*pC*	*pU*	*RE*	*3PSM*	*4PSM*	*AK1*	*AK2*	*WAAP*	*PP*	*EK*
{0, 0.822, None, No}	0.005	0.207	0.203	0.002	0.004	0.008	0.002	0.012	0.007	0.019	0.025
{0.2, 0.821, None, No}	0.006	0.110	0.106	0.002	0.004	0.008	0.002	0.012	0.017	0.019	0.025
{0.5, 0.818, None, No}	0.007	0.032	0.029	0.002	0.004	0.008	0.003	0.011	0.014	0.012	0.027
{0.8, 0.810, None, No}	0.010	0.004	0.003	0.002	0.005	0.008	0.004	0.009	0.010	0.013	0.029
{0, 0.864, Med, No}	0.004	0.101	0.096	0.003	0.020	0.028	0.001	0.025	0.005	0.030	0.034
{0.2, 0.856, Med, No}	0.006	0.046	0.040	0.003	0.032	0.037	0.003	0.034	0.016	0.035	0.039
{0.5, 0.835, Med, No}	0.012	0.008	0.006	0.003	0.048	0.066	0.012	0.079	0.015	0.027	0.043
{0.8, 0.805, Med, No}	0.019	0.003	0.004	0.003	0.051	0.079	0.021	0.083	0.010	0.016	0.039
{0, 0.879, High, No}	0.003	0.070	0.065	0.004	0.040	0.050	0.001	0.043	0.005	0.042	0.045
{0.2, 0.869, High, No}	0.006	0.029	0.024	0.005	0.063	0.065	0.004	0.055	0.015	0.050	0.052
{0.5, 0.837, High, No}	0.010	0.005	0.003	0.006	0.096	0.109	0.018	0.120	0.013	0.034	0.052
{0.8, 0.803, High, No}	0.020	0.005	0.007	0.004	0.109	0.142	0.031	0.142	0.011	0.019	0.047
{0, 0.769, None, Med}	0.022	0.056	0.055	0.053	0.009	0.020	0.021	0.009	0.020	0.014	0.010
{0, 0.763, Med, Med}	0.047	0.009	0.009	0.100	0.006	0.018	0.014	0.009	0.020	0.011	0.008
{0, 0.757, High, Med}	0.061	0.003	0.003	0.125	0.010	0.012	0.013	0.005	0.021	0.013	0.010
{0, 0.920, None, Med}	0.031	0.201	0.197	0.103	0.006	0.086	0.041	0.023	0.040	0.030	0.027
{0.2, 0.859, None, Med}	0.034	0.107	0.103	0.106	0.009	0.063	0.027	0.022	0.033	0.034	0.022
{0.5, 0.785, None, Med}	0.012	0.030	0.028	0.058	0.007	0.021	0.011	0.018	0.016	0.012	0.017
{0.8, 0.774, None, Med}	0.003	0.003	0.003	0.023	0.004	0.007	0.003	0.012	0.007	0.008	0.026
{0, 0.908, Med, Med}	0.050	0.100	0.090	0.136	0.072	0.151	0.029	0.048	0.040	0.027	0.024
{0.2, 0.829, Med, Med}	0.036	0.045	0.038	0.119	0.060	0.146	0.009	0.060	0.030	0.030	0.021

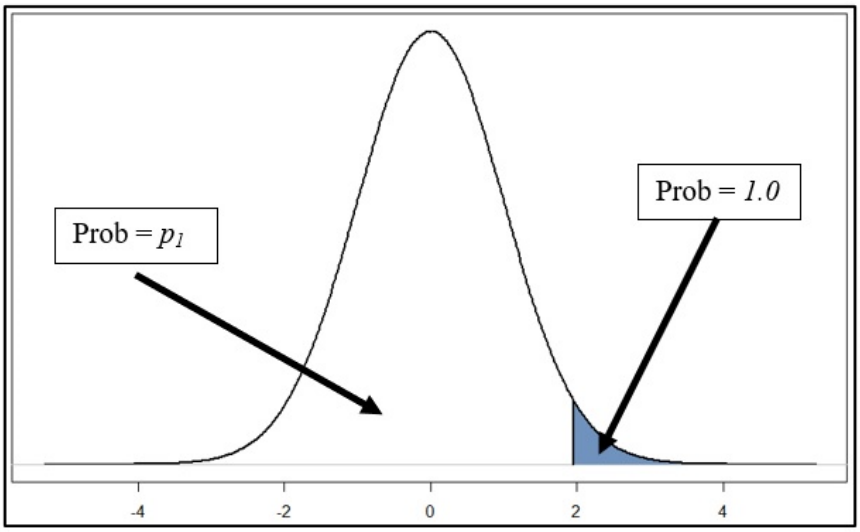
{Effect Size, I^2, QRP, Publication Selection}	*Estimators*										
	TF	***pC***	***pU***	***RE***	***3PSM***	***4PSM***	***AK1***	***AK2***	***WAAP***	***PP***	***EK***
{0, 0.901, High, Med}	0.060	0.069	0.060	0.153	0.108	0.136	0.026	0.042	0.042	0.030	0.027
{0.2, 0.816, High, Med}	0.038	0.029	0.022	0.124	0.111	0.145	0.006	0.067	0.030	0.032	0.022
{0, 0.755, None, Strong}	0.071	0.056	0.056	0.133	0.010	0.009	0.036	---*	0.033	0.030	0.011
{0, 0.895, None, Strong}	0.116	0.202	0.196	0.255	0.019	0.101	0.081	---*	0.087	0.074	0.043
{0.2, 0.807, None, Strong}	0.080	0.108	0.104	0.185	0.018	0.043	0.050	---*	0.064	0.053	0.023
{0.5, 0.759, None, Strong}	0.019	0.030	0.028	0.083	0.009	0.013	0.014	---*	0.023	0.015	0.014
{0.8, 0.768, None, Strong}	0.002	0.003	0.003	0.031	0.005	0.006	0.003	---*	0.008	0.008	0.027
{0, 0.843, Med, Strong}	0.130	0.101	0.090	0.271	0.051	0.053	0.056	---*	0.081	0.060	0.034
{0, 0.823, High, Strong}	0.135	0.071	0.060	0.277	0.041	0.041	0.051	---*	0.080	0.057	0.031
Average MSE =	0.035	0.062	0.058	0.079	0.034	0.056	0.020	0.041	0.027	0.028	0.028
(Smallest, Largest) =	(0.002, 0.135)	(0.003, 0.207)	(0.003, 0.203)	(0.002, 0.277)	(0.004, 0.111)	(0.006, 0.151)	(0.001, 0.081)	(0.005, 0.142)	(0.005, 0.087)	(0.008, 0.074)	(0.008, 0.052)

* Indicates that all estimates failed to converge for that experiment.

NOTE: This table reports estimator MSE performance results for the 30 experiments included within the {sample size = 100, high I^2} category of the CSG&H (2019a) simulations. The estimators are described in Section 2 of the text. The first column gives details about the individual experiment (cf. the bottom panel in APPENDIX 2). Each cell represents results for a single experiment consisting of 3,000 simulated meta-analyses. Each simulated meta-analysis produces a single estimate of the mean population effect. The numbers in the table are the averaged mean squared error (MSE) value for the 3,000 simulated meta-analyses for that estimator and experiment. The last two rows of each panel report the overall average MSE, followed by the smallest and largest (average) MSE values over the 30 experiments. Yellow-highlighted cells in the upper panel of the table identify the smallest (average) MSE for each experiment. The yellow-highlighted cell in the bottom panel of the table identifies the estimator (AK1) with the lowest overall, averaged MSE value. The blue-highlighted cells identify estimators that are close to AK1 in terms of overall performance.

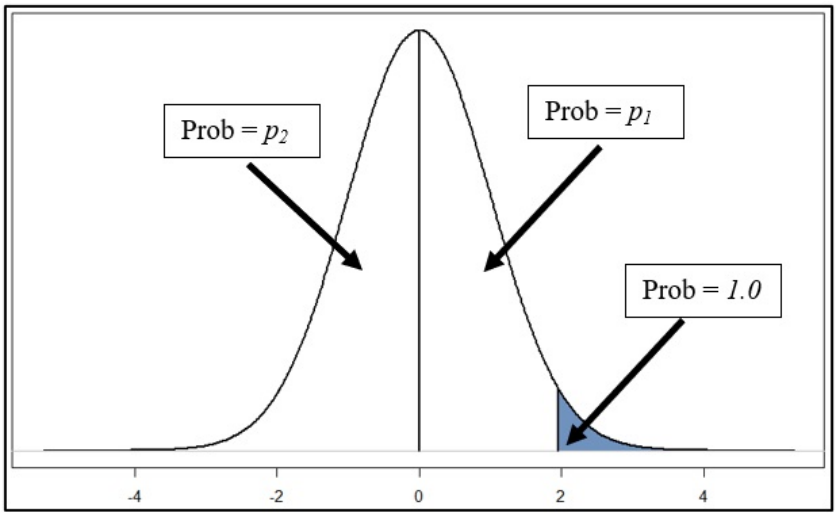
FIGURE 1
Illustration of 3PSM and 4PSM

A. 3PSM (Positive and Significant)



$$\text{Relative probability} = \begin{cases} 1, & if\ \hat{\beta}_i/SE_i \geq 1.96 \\ p_1, & if\ \ \hat{\beta}_i/SE_i < 1.96 \end{cases}$$

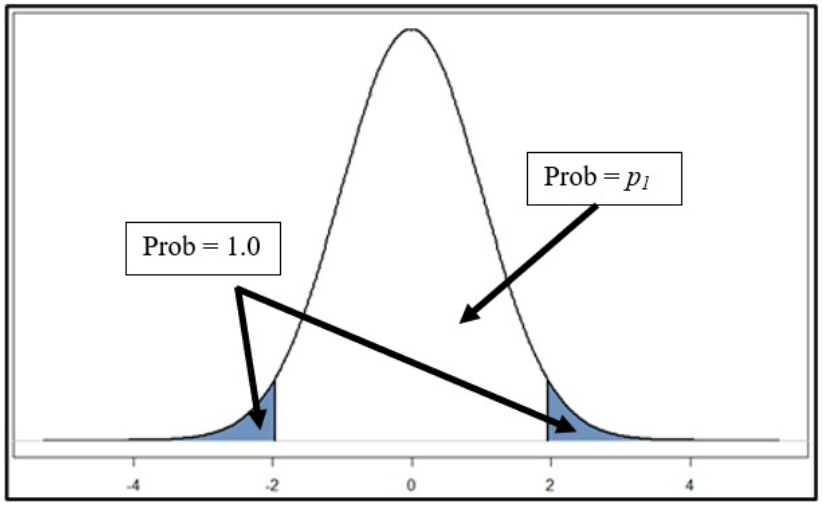
B. 4PSM (Positive/Insignificant and Positive/Significant)



$$\text{Relative probability} = \begin{cases} 1, & if\left(\hat{\beta}_i/SE_i\right) \geq 1.96 \\ p_1, & if\ 0 \leq \left(\hat{\beta}_i/SE_i\right) < 1.96 \\ p_2, & if\ \left(\hat{\beta}_i/SE_i\right) < 0 \end{cases}$$

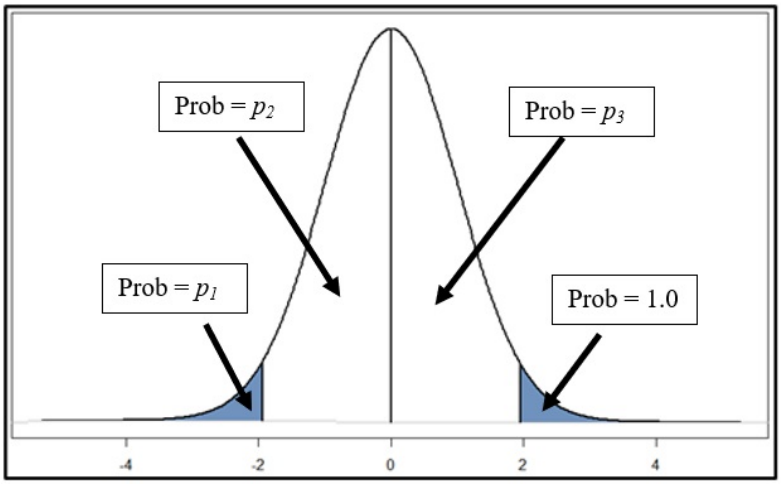
FIGURE 2
Illustration of AK1 and AK2

A. AK1 (Symmetric Selection)



$$\text{Relative probability} = \begin{cases} 1, & if\left|\hat{\beta}_i/SE_i\right| \geq 1.96 \\ p_1, & if\ \left|\hat{\beta}_i/SE_i\right| < 1.96 \end{cases}$$

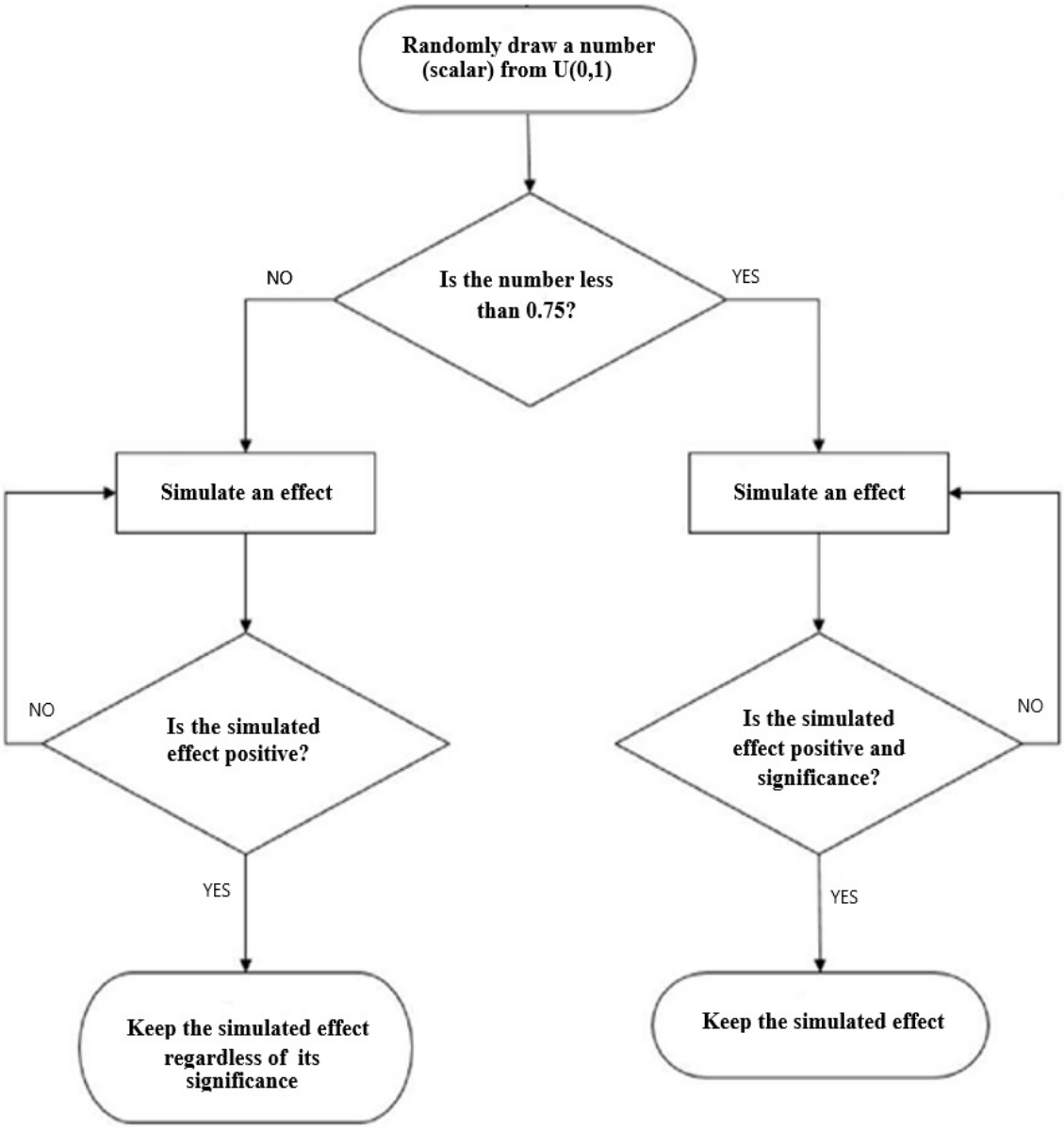
B. AK2 (Asymmetric Selection)



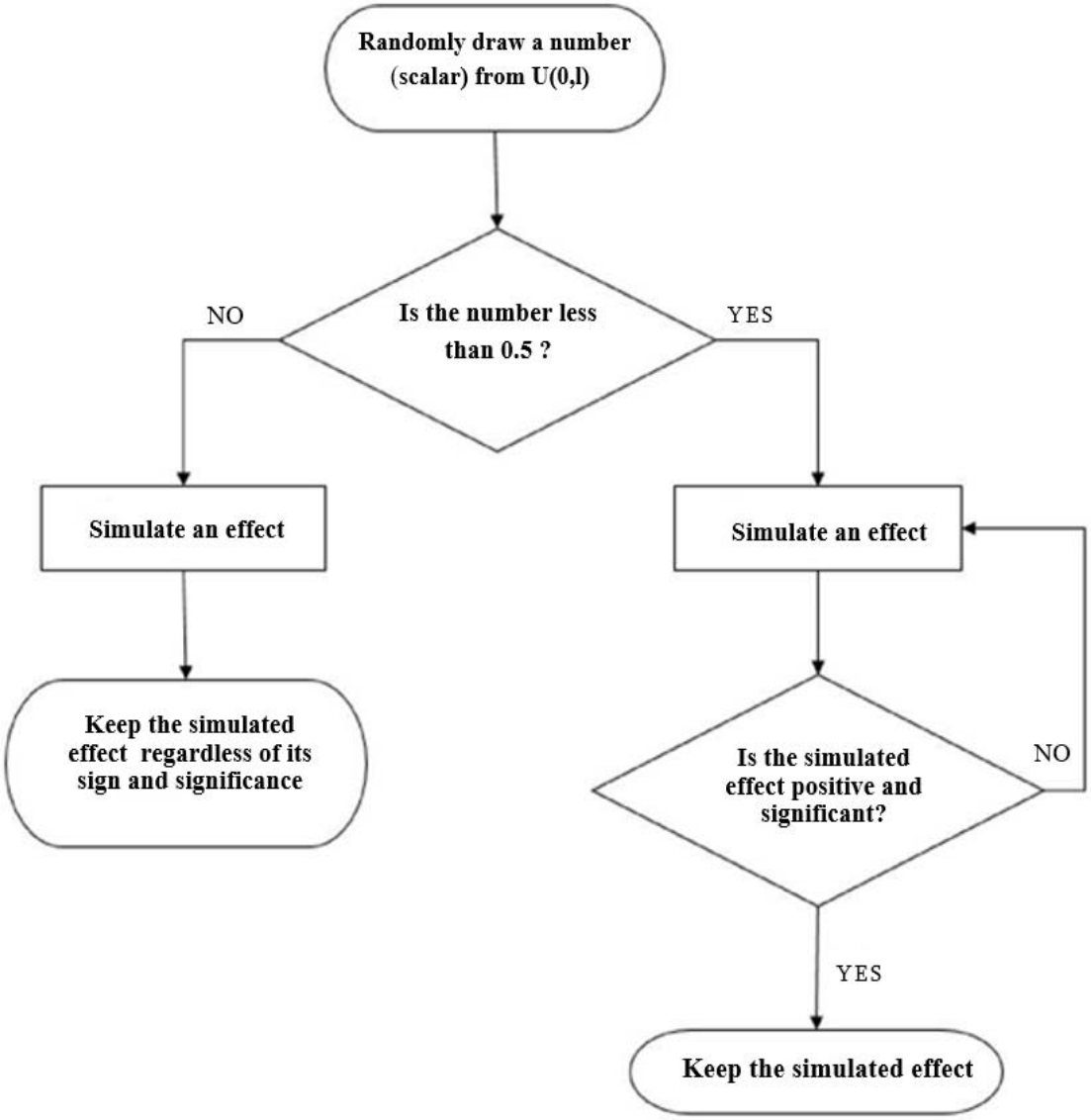
$$\text{Relative probability} = \begin{cases} p_1, & if\left(\hat{\beta}_i/SE_i\right) < -1.96 \\ p_2, & if\ -1.96 \leq \left(\hat{\beta}_i/SE_i\right) < 0 \\ p_3, & if\ 0 \leq \left(\hat{\beta}_i/SE_i\right) < 1.96 \\ 1, & if\left(\hat{\beta}_i/SE_i\right) \geq 1.96 \end{cases}$$

APPENDIX 1
Different Publication Selection Procedures in SD&I (2017)

A. "50% Selective Reporting"



B. "75/100% Selective Reporting"



APPENDIX 2
Description of Experiments

STUDY: Stanley, Doucouliagos, & Ioannidis (2017). Finding the power to reduce publication bias. *Statistics in Medicine*.
TABLES: I, II, III, IV, V, VI

	Log Odds Ratio	***Cohen's d***
Effect Size	{0, 0.3, 0.54}	{0, 0.5}
Heterogeneity	$\sigma_h = \{0, 0.006\}$*	{0, 6.25, 12.5, 25, 50}
Publication Selection	{0%, 50%}*	{0%, 50%, 75/100%}
Sample Size	{5, 10, 20, 40, 80, 100, 200, 400, 800)	{5, 10, 20, 40, 80, 100, 200, 400, 800)
# of Experiments	3 x 2 x 9 = 54	2 x 5 x 3 x 9 = 270

STUDY: Alinaghi & Reed (2018). Meta-analysis and publication bias: How well does the FAT-PET-PEESE procedure work? *Research Synthesis Methods*.
TABLE: 6

	Regression (Random Effects)	***Regression (Panel Random Effects)***
Effect Size	{0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4}	{0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4}
Heterogeneity	Endogenous: $I^2 > 75\%$	Endogenous: $I^2 > 75\%$
Publication Selection	{Significance, Correct Sign}	{Significance, Correct Sign}
Sample Size	Endogenous: > 100	Endogenous: > 100
# of Experiments	9 x 2 = 18	9 x 2 = 18

STUDY: Bom and Rachinger (2019). A kinked meta-regression model for publication bias correction. *Research Synthesis Methods.*
TABLES: 1, 2, 3; FIGURES: 4, 5, 6, 7

	Regression
Effect Size	{0, 1}
Heterogeneity	{0, 0.125, 0.25, 0.5, 1, 2, 4}
Publication Selection	{0%, 25%, 50%, 75%}
Sample Size	{5, 10, 20, 40, 80, 100, 200, 400, 800}
# of Experiments	2 x 7 x 4 x 9 = 504

STUDY: Carter et al. (2019a). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science.*

TABLES: Reports stored online at https://osf.io/rf3ys

	Cohen's d
Effect Size	{0, 0.2, 0.5, 0.8}
Heterogeneity	$\{\sigma_h = 0, 0.2, 0.4\}$
Publication Selection	{"No", "Medium", Strong"}
Questionable Research Practice	{"None", "Medium", High"}
Sample Size	{10, 30, 60, 100, 200, 400, 800)
# of Experiments	4 x 3 x 3 x 3 x 7 = 756

*When publication selection is 0%, Stanley, Doucouliagos, and Ioannidis (2017) only allow $\sigma_h = 0.006$; and when publication selection is 50%, they only allow $\sigma_h = 0.006$.