aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2016, Vol. 11(5) 730–749

DOI: 10.1177/1745691616662243
pps.sagepub.com
SAGE

# Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes

**Blakeley B. McShane[1], Ulf Böckenholt[1], and Karsten T. Hansen[2]**
[1]Kellogg School of Management, Northwestern University and,
[2]Rady School of Management, University of California, San Diego

**Abstract**
We review and evaluate selection methods, a prominent class of techniques first proposed by Hedges (1984) that assess and adjust for publication bias in meta-analysis, via an extensive simulation study. Our simulation covers both restrictive settings as well as more realistic settings and proceeds across multiple metrics that assess different aspects of model performance. This evaluation is timely in light of two recently proposed approaches, the so-called *p*-curve and *p*-uniform approaches, that can be viewed as alternative implementations of the original Hedges selection method approach. We find that the *p*-curve and *p*-uniform approaches perform reasonably well but not as well as the original Hedges approach in the restrictive setting for which all three were designed. We also find they perform poorly in more realistic settings, whereas variants of the Hedges approach perform well. We conclude by urging caution in the application of selection methods: Given the idealistic model assumptions underlying selection methods and the sensitivity of population average effect size estimates to them, we advocate that selection methods should be used less for obtaining a single estimate that purports to adjust for publication bias ex post and more for sensitivity analysis—that is, exploring the range of estimates that result from assuming different forms of and severity of publication bias.

**Keywords**
meta-analysis, effect size, selection methods, *p*-curve, *p*-uniform

Meta-analysis is a well-established statistical technique that synthesizes two or more studies of a common phenomenon. Insofar as the studies measure the common phenomenon with some degree of error, a meta-analysis, which pools the results from the studies via a weighted average, will yield a measurement that is on average more accurate than that of any individual study. Thus, one purpose of meta-analysis is to estimate the average effect size in a set of studies.

In some cases, one may wish to move beyond using meta-analysis to estimate the average effect size in a set of studies to using it to estimate the average effect size in some larger population. This poses at least two difficulties that have analogues in single study research. First, just as the single study researcher seeking to generalize an effect must define the relevant population of individuals, the meta-analyst must define the relevant population of studies. Second, insofar as the set of individuals examined in a single study or the set of studies examined in a meta-analysis is not representative of this population, estimates will be biased.

Publication bias is the term given to concerns over the representativeness of any given study or set of studies. These concerns date back centuries (Boyle, 1661/1965; Dickersin, 2005; Editors, 1909; Ferriar & Simmons, 1792; Hall, 1959; Lane & Dunlap, 1978; Sterling, 1959) and relate not only to issues surrounding the size, direction, and statistical significance of study results but also to issues surrounding the availability and accessibility of studies, including cost, language, and familiarity. Thus, publication bias encompasses any and all biases in the

**Corresponding Author:**
Blakeley B. McShane, Kellogg School of Management, Northwestern University, Donald P. Jacobs Center, 2001 Sheridan Rd., Evanston, IL 60208
E-mail: b-mcshane@kellogg.northwestern.edu

set of studies available to the meta-analyst and to researchers more broadly.

Although publication bias is typically considered a problem only for meta-analysis, it is of course just as much of a problem for single study research: A single study drawn from a set of biased studies is precisely as biased as a meta-analysis of the set of studies. However, meta-analysis is advantageous relative to single study research in this regard because, unlike single study research, it can be used to assess and adjust for publication bias. Indeed, as reviewed in the invaluable tome by Rothstein, Sutton, and Borenstein (2005), vast effort has been expended developing various techniques to do exactly this.

A prominent class of these techniques are so-called selection methods. First proposed by Hedges (1984), selection methods assess and adjust for publication biases relating to the size, direction, and statistical significance of study results.

In this article, we review and evaluate selection methods via an extensive simulation study that covers both restrictive settings, involving rigid publication (or selection) rules and homogeneous effect sizes, as well as more realistic (though still rather idealistic) settings, involving more flexible publication rules and heterogeneous effect sizes. Our evaluation proceeds across multiple metrics that assess different aspects of model performance such as estimation accuracy and confidence-interval coverage. This type of evaluation is new to the selection methods literature and is critical: Proper model evaluation inherently requires the assessment of model performance across a variety of settings and metrics because models may perform well in some respects but poorly in others. In addition, our evaluation is timely in light of two recently proposed approaches, the so-called *p*-curve and *p*-uniform approaches (Simonsohn, Nelson, & Simmons, 2014; van Assen, van Aert, & Wicherts, 2015), which can be viewed as alternative implementations of the original Hedges (1984) selection method approach that employ different estimation strategies.

We find that the *p*-curve and *p*-uniform approaches perform reasonably well but not as well as the original Hedges (1984) approach in the restrictive setting for which all three were designed; this is a direct consequence of the alternative estimation strategies they employ. We also find that the *p*-curve and *p*-uniform approaches are sensitive to deviations from their model assumptions and thus perform poorly in more realistic settings; in contrast, the Hedges (1984) approach generalizes easily to—and thus variants of it perform well in—these more realistic settings.

In the remainder of this article, we first provide a brief review of selection methods. We then note that it is typical in behavioral research that studies with results that are not statistically significant are published and that effect sizes are heterogeneous; this is critical because falsely assuming otherwise, as the original Hedges (1984) approach and the *p*-curve and *p*-uniform approaches do, results in a loss of efficiency (i.e., noisier estimates of the population average effect size) and bias, respectively. We next present our simulation study. Finally, we conclude by urging caution in the application of selection methods: Given the idealistic model assumptions underlying selection methods and the sensitivity of population average effect size estimates to them, we advocate that selection methods should be used less for obtaining a single estimate that purports to adjust for publication bias ex post and more for sensitivity analysis—that is, exploring the range of estimates that result from assuming different forms of and severity of publication bias. We summarize our key points and recommendations in Table 1.

## Selection Methods

In this section, we introduce selection methods and discuss several of their benefits. We then provide a review of various selection method approaches. In particular, we first discuss the early contributions of Hedges (1984) and Iyengar and Greenhouse (1988). We then discuss generalizations of these early methods. Finally, we discuss the *p*-curve and *p*-uniform approaches, which can be viewed as alternative implementations of the original Hedges (1984) selection method approach that employ different estimation strategies. We summarize several selection method approaches in Table 2.

### *Introduction to selection methods*

Selection methods are a prominent class of techniques that assess and adjust for publication biases relating to the size, direction, and statistical significance of study results. The statistical model underlying selection methods consists of two components, a data model and a selection model.

The data model describes how the data are generated in the absence of any publication bias, and it is generally chosen to be equivalent to the data models typically employed in behavioral research. For example, when interest centers on the difference between two independent means, the data model typically specifies, as is common in behavioral research, that individual-level observations are normally distributed with common but unknown variance.

The selection model describes the publication process, and it can take a wide variety of forms. For example, it might specify that (a) only studies with results that are statistically significant are published, (b) only studies with results that are statistically significant and

**Table 1.** Key Points and Recommendations

1. Publication bias distorts meta-analytic estimates of both the population average effect size and the degree of heterogeneity. Estimates of the former are typically biased upward, thus giving the false impression of large effect sizes, whereas estimates of the latter are typically biased downward, thus giving the false impression of homogeneity.
2. Selection methods are a prominent class of techniques that assess and adjust for publication bias in meta-analysis. They were first proposed by Hedges (1984) and have been an active research area ever since.
3. Two recent proposals, the so-called *p*-curve and *p*-uniform approaches (Simonsohn, Nelson, & Simmons, 2014; van Assen, van Aert, & Wicherts, 2015), can be viewed as alternative implementations of the original Hedges (1984) selection method approach that employ different estimation strategies.
4. The Hedges (1984), *p*-curve, and *p*-uniform approaches are all one-parameter approaches that assume (a) that only studies with results that are statistically significant are published and (b) that effect sizes are homogeneous across studies. When these assumptions hold, the *p*-curve and *p*-uniform approaches perform reasonably well but, as a result of the alternative estimation strategies they employ, not as well as the original Hedges (1984) approach.
5. Falsely assuming that assumptions (a) and (b) hold results in a loss of efficiency (i.e., noisier estimates of the population average effect size) and bias, respectively. Consequently, when one or both assumptions fail to hold, the *p*-curve and *p*-uniform approaches perform poorly, whereas variants of the Hedges (1984) approach perform well.
6. Both assumptions are nearly always false in behavioral research. Consequently, a simple three-parameter variant of the Hedges (1984) approach that relaxes them should be the minimal model considered in applied work. More advanced selection methods may also be considered.
7. Idealistic model assumptions underlie even the most advanced selection methods, and population average effect size estimates can be highly sensitive to these assumptions. Consequently, we advocate that selection methods should be used less for obtaining a single estimate that purports to adjust for publication bias ex post and more for sensitivity analysis—that is, exploring the range of estimates that result from assuming different forms of and severity of publication bias (see Vevea & Woods, 2005, and Hedges & Vevea, 2005).

directionally consistent are published, or (c) studies with results that are not statistically significant (or directionally consistent) are relatively less likely to be published than studies with results that are statistically significant (and directionally consistent).

Because the data model and selection model are explicitly specified by selection methods, they possess a number of principal advantages:

1. They allow identifiability—that is, whether it is possible even in principle to estimate the underlying model parameters—to be assessed.
2. They allow estimation to proceed easily via the maximum likelihood estimation strategy, which has strong theoretical properties (Bickel & Doksum, 2007; Casella & Berger, 2002), yields standard errors and confidence intervals, and allows for hypothesis tests of model parameters.
3. They provide a framework that can in principle accommodate almost any meta-analytic setting, including, for example, heterogeneous effect sizes, study-level moderators, and other features in the data model as well as highly flexible forms of publication bias in the selection model.
4. They can be used to test for, evaluate the extent of, examine the sensitivity to, and adjust for publication bias as specified by the selection model.

Given these benefits, selection methods have been an active research area since they were originally proposed by Hedges (1984; for an overview, see Hedges & Vevea, 2005; Chapter 13 of Schmidt & Hunter, 2014; and Jin, Zhou, & He, 2015).

## *Early selection methods*

The original selection method of Hedges (1984) assumes that (a) effect sizes are homogenous across studies and effect size estimates are normally distributed with unknown variance (i.e., so that individual study *t* statistics are modeled as noncentral *t* distributed; this data model arises from, inter alia, the canonical case in which a study follows a two-condition between-subjects design, interest centers on the difference between the means of the two conditions, and the individual-level observations are normally distributed with common but unknown variance) and (b) only studies with results that are statistically significant are published. These assumptions for the data model and selection model, respectively, imply a simple one-parameter likelihood function (i.e., one parameter for the data model and zero parameters for the selection model). However, the simplicity is deceptive: Although, strictly speaking, the applicability of the Hedges (1984) approach is limited, it nonetheless contains the two ingredients of a selection method (i.e., a data model and a selection model) and shows how to combine them. It thus provides a framework that is easy to build upon and that can accommodate (at least conceptually) almost any meta-analytic setting.

**Table 2.** Summary of Several Selection Methods

| Article(s) | Data model | Selection model |
|---|---|---|
| Hedges (1984) | Effect sizes are modeled as homogeneous across studies. Effect size estimates are modeled as normally distributed with unknown variance (i.e., so that individual study $t$ statistics are modeled as noncentral $t$ distributed). | Only studies with results that are statistically significant are published. |
| Iyengar and Greenhouse (1988) | As in Hedges (1984). In the discussion and rejoinder, the data model was conceptually expanded to accommodate heterogeneous effect sizes as in Hedges (1992). | Studies with results that both are and are not statistically significant are published but with different relative likelihoods. The relative likelihood is modeled via one of two simple one-parameter functions. |
| Dear and Begg (1992); Hedges (1992) | Effect sizes are modeled as heterogeneous across studies via a normal distribution with common mean and common variance. Effect size estimates are modeled as normally distributed with known variance. | Studies with results that both are and are not statistically significant are published but with different relative likelihoods. The relative likelihood is modeled via complex multiparameter functions. |
| Vevea and Hedges (1995) | Effect sizes are modeled as heterogeneous across studies via a normal distribution with mean that is a linear function of study-level moderators and common variance. Effect size estimates are modeled as normally distributed with known variance. | As in Hedges (1992). |
| Copas (1999); Copas and Li (1997); Copas and Shi (2001) | As in Hedges (1992). | Studies with results that both are and are not statistically significant are published but with different relative likelihoods. The relative likelihood is modeled via a linear function that depends on the estimate of the effect size and its standard error. |
| Simonsohn et al. (2014) | As in Hedges (1984). | As in Hedges (1984). |
| van Assen et al. (2015) | As in Hedges (1984). | As in Hedges (1984). |

Note: The estimation strategy employed by all but the last two articles is maximum likelihood. Simonsohn, Nelson, and Simmons (2014) and van Assen, van Aert, and Wicherts (2015) employed a distance-based estimation strategy based on the Kolmogorov–Smirnov statistic and, in the most recent implementation (van Aert, Wicherts, & van Assen, 2016), the Irwin-Hall distribution, respectively.

Iyengar and Greenhouse (1988) generalized the Hedges (1984) approach to allow for a less rigid selection model for the publication process that accommodates the publication of studies with results that both are and are not statistically significant. In particular, Iyengar and Greenhouse (1988) introduced a weight function approach that models the likelihood that a study with results that are not statistically significant is published relative to the likelihood that a study with results that are statistically significant is published. This relative likelihood is estimated from the data rather than fixed at zero, as is implicit in the Hedges (1984) approach.

More specifically, Iyengar and Greenhouse (1988) considered two relatively simple weight functions for the selection model: a one-parameter power function that implies the relative likelihood that a study with results that are not statistically significant is published increases as those results approach statistical significance and a one-parameter step function that implies that this relative likelihood is constant. A principal benefit of the Iyengar and Greenhouse (1988) approach is that both of the weight functions they considered accommodate the setting in which all studies are published (i.e., the no publication bias setting), the setting in which only studies with results that are statistically significant are published (i.e., the Hedges, 1984, setting), and settings that fall between these two extremes. Both weight functions also result in a two-parameter likelihood function (i.e., one parameter for the data model and one parameter for the selection model).

In the discussion and rejoinder surrounding Iyengar and Greenhouse (1988), the data model was conceptually expanded to accommodate effect sizes that are heterogeneous across studies, thus resulting in three-parameter likelihood models (i.e., two parameters for the data model and one parameter for the selection model). In this article, we limit ourselves to comparing more recent approaches to these very simple, early selection methods (adopting

for simplicity the Iyengar & Greenhouse, 1988, one-parameter step function as the weight function for the selection model). Thus, the models we consider have no more than three parameters: an effect size parameter, which gives the population average effect size; a heterogeneity parameter, which gives the degree of heterogeneity in the effect sizes; and a weight parameter, which gives the likelihood that a study with results that are not statistically significant is published relative to a study with results that are statistically significant. These parameters allow for the estimation and testing of (a) effect sizes in a manner that adjusts for publication bias; (b) the degree of heterogeneity in a manner that adjusts for publication bias; and (c) the degree of publication bias. Further, given the parsimony of the models (i.e., three or fewer parameters), they can be well estimated with comparably little data (i.e., a relatively small number of studies).

As a technical point, we note that Hedges (1984) and Iyengar and Greenhouse (1988) assumed two-sided rather than one-sided selection (i.e., the selection model assumed that only studies with results that are statistically significant—rather than statistically significant and directionally consistent—are published). As the difference between one-sided and two-sided selection is trivial conceptually and requires only the most minor of modifications to the likelihood function, we ascribe both the one-sided and two-sided versions to the respective authors. However, for consistency with Simonsohn et al. (2014) and van Assen et al. (2015), we evaluate the one-sided version.

### *Generalized selection methods*

Selection methods can easily be extended to accommodate very general data models and selection models. For example, although the data models employed by Hedges (1984) and Iyengar and Greenhouse (1988) were relatively simple, they have been extended to accommodate heterogeneous effect sizes, study-level moderators, and other features (Hedges, 1992; Hedges & Vevea, 2005; Vevea & Hedges, 1995).

Similarly, the selection model can be made very general by building on the weight function approach of Iyengar and Greenhouse (1988; Dear & Begg, 1992; Hedges, 1992; Hedges & Vevea, 2005; Vevea & Hedges, 1995). In these models, the weight function specifies the relative likelihood that a study is published as a function of its one-sided $p$ value. When the direction of the results is relevant (or, conversely, is not relevant) for selection, the weight function is asymmetric (or, conversely, is symmetric) about $p = .5$. We note the use of the one-sided $p$ value in the selection model does not imply or require that the original studies conducted one-sided tests; the one-sided $p$ value is used solely because it preserves information about not only the statistical significance of the results but also their direction.

Although the weight functions implicit in the selection model of Hedges (1984) and employed in the selection model of Iyengar and Greenhouse (1988) were relatively simple, they have been extended to accommodate highly flexible forms of publication bias (Dear & Begg, 1992; Hedges, 1992; Vevea & Hedges, 1995). In particular, Dear and Begg (1992), Hedges (1992), and Vevea and Hedges (1995) used complex multiparameter weight functions that can in principle approximate any functional form. A potential advantage of this approach is that the complexity of the weight function can be tuned to the amount of data available: When there are relatively few (or, conversely, many) studies available, less (or, conversely, more) complex weight functions with fewer (or, conversely, more) parameters can be used. Nonetheless, these approaches can sometimes pose difficulties for estimation and interpretation (Hedges & Vevea, 2005); consequently, one recommendation to which we return in the Discussion section is to use these approaches for sensitivity analysis (Copas, 2013; Vevea & Woods, 2005) rather than estimation.

We also note there is an alternative class of weight function approaches that depend on the estimate of the effect size and its standard error rather than on the $p$ value (Copas, 1999; Copas & Li, 1997; Copas & Shi, 2001). In principle, these methods are more flexible because the weight function depends on the estimate of the effect size and its standard error separately, rather than through their ratio as in the $p$-value approach. In practice, it is not always possible to estimate all of the parameters of these models simultaneously; further, even when it is, the likelihood suggests that there is little information about the key parameters. Consequently, these methods are mostly used for sensitivity analysis (Copas, 1999; Copas & Shi, 2001).

Finally, we note that additional work has considered parametric forms for the weight function, Bayesian estimation strategies, and other issues (for an overview, see Hedges & Vevea, 2005; Chapter 13 of Schmidt & Hunter, 2014; and Jin et al., 2015).

### p-*methods*

Two approaches that are nearly identical to one another and that are based on assumptions identical to those of the original Hedges (1984) approach have recently been proposed (Simonsohn et al., 2014; van Assen et al., 2015). These approaches, the so-called $p$-curve and $p$-uniform approaches, respectively, return to the simple data model and selection model of Hedges (1984) but forgo the maximum likelihood estimation strategy in favor of alternative estimation strategies. In particular, because the distribution of the $p$ value under the null hypothesis that the effect size is equal to the true effect size is uniform, they estimate the effect size as that which minimizes the

distance between the observed distribution of *p* values (conditional on their being statistically significant, as per the assumed selection model) and the uniform distribution. The two approaches differ only in their distance metric: The *p*-curve approach uses the Kolmogorov–Smirnov statistic as the distance metric, whereas the most recent implementation of the *p*-uniform approach (van Aert, Wicherts, & van Assen, 2016, this issue) uses a moment estimator based on the Irwin–Hall distribution. Thus, rather than being viewed as new methods per se, they can be viewed as alternative implementations of the original Hedges (1984) selection method approach that employ different estimation strategies; thereby, they affirm the enduring value of this venerable approach.

Although alternative estimation strategies not based on the likelihood function are common in statistics, they are typically used in order to circumvent the model specification assumptions required by likelihood-based estimation strategies (e.g., individual-level observations are normally distributed) or to provide robustness to violations of them. However, given a set of model specification assumptions, the maximum likelihood estimation strategy is highly principled: It has a number of desirable theoretical properties, such as yielding asymptotically minimum variance unbiased estimators with normal sampling distributions and likelihood functions that can be used to test hypotheses about model parameters (Bickel & Doksum, 2007; Casella & Berger, 2002). Given that the *p*-curve and the *p*-uniform approaches require model specification assumptions (i.e., in order to compute the distribution of the *p* value under the null) and indeed make the very same assumptions as Hedges (1984), we find the benefits of using alternative estimation strategies unclear as they lack these desirable theoretical properties.

We also believe that there are a number of other disadvantages associated with these alternative estimation strategies in this setting. First, as the long literature discussed in the prior two subsections demonstrates, it is conceptually and practically easy to generalize the data model and the selection model (e.g., to accommodate heterogeneous effect sizes, study-level moderators, and the publication of studies with results that are not statistically significant) when employing a likelihood-based estimation strategy; although this is theoretically possible under these alternative estimation strategies, doing so does not seem nearly as conceptually or practically straightforward. Second, as noted, the maximum likelihood estimation strategy naturally yields likelihood values that can be compared across model variants as well as (asymptotic) standard errors and thus confidence intervals; on the other hand, the *p*-uniform approach yields only a confidence interval, and the *p*-curve approach yields neither a standard error nor a confidence interval (although the bootstrap can in theory be used to obtain standard errors and confidence intervals for the *p*-curve approach—or for that matter any other approach—this is complicated in practice because [a] the bootstrap is numerically intensive, [b] there are often few observations available for bootstrapping, [c] the bootstrap distribution can be highly non-unimodal, and [d] numerical instability issues with the lower boundary of the *p*-curve estimation strategy can cause a large fraction of bootstrap iterations to estimate improperly at the lower boundary; for an illustration, see a histogram of the bootstrap distribution of the "choice is bad" results presented in Simonsohn et al., 2014).

## Modeling Considerations

We have posited that a major advantage of the selection method approach is that it provides a framework that can in principle accommodate almost any meta-analytic setting. Thus, although the original Hedges (1984) method assumed that only studies with results that are statistically significant are published and that effect sizes are homogeneous across studies, these assumptions can be easily relaxed, as demonstrated by Iyengar and Greenhouse (1988), Hedges (1992), and much subsequent literature (see Hedges & Vevea, 2005). In this section, we discuss why it is critical to relax these two assumptions: Both are nearly always false in behavioral research, and falsely assuming otherwise—as not only the original Hedges (1984) approach but also the *p*-curve and *p*-uniform approaches do—results in a loss of efficiency (i.e., noisier estimates of the population average effect size) and bias, respectively.

Although studies with results that are statistically significant are overrepresented in the published literature relative to those with results that are not statistically significant (Fanelli, 2010, 2012; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995), it is clearly not the case that there are no instances of the latter in the published literature. Methods such as the Hedges (1984) approach and the *p*-curve and *p*-uniform approaches, which falsely assume that only studies with results that are statistically significant are published, must necessarily exclude studies with results that are not statistically significant in estimation; this results in a loss of efficiency relative to methods that can incorporate these studies in estimation by assuming more general selection models. This loss of efficiency can be dramatic in practice even when studies with results that are not statistically significant have a small relative likelihood of publication.

In addition, although heterogeneity has long been regarded as important across sets of studies that consist

of general (i.e., systematic or conceptual) replications, there is mounting evidence of and growing appreciation for heterogeneity across sets of studies that consist entirely of close replications (i.e., studies that use identical or very similar materials). For example, consider the Many Labs Replication Project, which provides 16 estimates of 13 classic and contemporary behavioral research effects from 36 independent samples totaling 6,344 subjects (Klein et al., 2014). Each of the 36 laboratories involved in the Many Labs project used identical materials, and these materials were administered through a web browser in order to minimize heterogeneity. Nonetheless, random effects meta-analyses of these studies conducted by the Many Labs authors yielded nonzero estimates of heterogeneity for all 14 of the effects they found to be non-null. Further, 40% of the total variability on average across these effects was due to heterogeneity (see the $I^2$ statistics reported in Table 3 of Klein et al., 2014).

In addition, among the 6,344 Many Labs subjects were 1,000 recruited via Amazon's Mechanical Turk. The study materials were administered to these 1,000 subjects over 7 unique days, beginning on August 29, 2013, and ending on September 11, 2013 (i.e., 7 consecutive days excluding Fridays, weekends, and the Labor Day holiday). Restricting attention to only these subjects and treating each unique day as a separate sample yields seven extremely close replications of each effect. Again, however, despite the extreme degree of closeness, heterogeneity is nontrivial: Random effects meta-analyses yield nonzero estimates of heterogeneity for nine of the 14 non-null effects, and across these, 21% of the total variability on average was due to heterogeneity.

Given the degree of heterogeneity present in the Many Labs studies (for which the only difference among the studies was the location of the laboratory) and in the Mechanical Turk subsample of these studies (for which the only difference among the studies was the day on which the study materials were administered), it seems reasonable to conclude that some degree of heterogeneity is the norm in behavioral research (see also Gelman, 2015; McShane & Böckenholt, 2014; McShane & Gal, 2016).

When there is publication bias stemming from the statistical significance of study results and heterogeneity is nonzero (i.e., nearly always), methods such as the Hedges (1984) approach and the *p*-curve and *p*-uniform approaches that falsely assume homogeneity will, by Jensen's inequality (Jensen, 1906), produce upwardly biased estimates of the population average effect size (for a nontechnical discussion of Jensen's inequality, see McShane & Böckenholt, 2016). This bias can be dramatic in practice, thus resulting in poor estimates. In addition, approaches that falsely assume homogeneity ignore an important source of variation and thus produce standard errors that are too small and confidence intervals that are too narrow, thereby reflecting an overly optimistic level of certainty.

Further complicating matters is that estimates of heterogeneity from standard meta-analytic methods are generally downwardly biased when there is publication bias, thus giving the false impression of homogeneity. Consequently, it is not only potentially deleterious to rely on methods such as the Hedges (1984) approach and the *p*-curve and *p*-uniform approaches in this setting but also difficult, if not impossible, to determine when one faces such a situation without using methods that account for both heterogeneity and publication bias; nonetheless, as we have suggested in the above paragraphs, this setting is the norm in behavioral research.

In subsequent sections, we focus on generalizability issues surrounding these two factors: the publication of studies with results that are not statistically significant and heterogeneity. However, we note that there are many other important factors to consider accounting for, such as study-level moderators, studies with multiple dependent effects of interest, and dependence among studies in both the data model and the selection model; we lay these aside for the moment and return to them in the Discussion section.

## Simulation Evaluation

In this section, we discuss our simulation design and evaluation metrics. Our simulation covers both restrictive settings, involving rigid publication (or selection) rules and homogeneous effect sizes, as well as more realistic (though still rather idealistic) settings, involving more flexible publication rules and heterogeneous effect sizes. Our evaluation proceeds across multiple metrics that assess different aspects of model performance such as estimation accuracy and confidence-interval coverage. To briefly preview our results, the *p*-curve and *p*-uniform approaches perform worse—and, in the most realistic setting, much worse—than the maximum likelihood estimation approaches of Hedges (1984) and Iyengar and Greenhouse (1988).

### *Simulation design and evaluation metrics*

We assume that each simulated study follows a two-condition between-subjects design, that interest centers on the difference between the means of the two conditions, and that the individual-level observations are normally distributed with common but unknown variance. We further assume that the sample size per condition in a given study is equal across conditions and is uniformly

distributed between 25 and 100, that the true effect size in a given study as measured on the standardized Cohen's *d* scale is normally distributed with mean $\delta$ and standard deviation $\tau$, and that studies are "published" in a biased manner. In particular, we assume that studies are always published if the results are directionally consistent with $\delta$ and a two-sided *t* test is significant at the $\alpha = .05$ level (or, equivalently, if a one-sided *t* test is significant at the $\alpha = .025$ level) and that they are published with probability *q* otherwise; in practice, this assumption does not require that all studies with results that are statistically significant are published but simply that their likelihood of publication relative to those with results that are not statistically significant is 1:*q*.

We vary the number of published studies included in the meta-analysis from 20 to 100 in increments of 20 and the population average effect size $\delta$ from .1 to .9 in increments of .2, and we explore values of $\tau$ and *q* that continuously build upon each other. In particular, we first consider the case in which (a) only studies with results that are statistically significant and directionally consistent are published and (b) effect sizes are homogeneous across studies (Simulation 1: $q = 0$, $\tau = 0$). We then relax the assumption that only studies with results that are statistically significant are published (Simulation 2: $q > 0$, $\tau = 0$). Finally, we relax the assumption that effect sizes are homogeneous (Simulation 3: $q > 0$, $\tau > 0$). Because these simulations build upon one another in a continuous fashion, they are better viewed not as multiple distinct simulations but as one single simulation with multiple aspects designed to demonstrate the value of a framework for model generalization.

In our evaluation, we focus on the estimation of $\delta$ by evaluating estimates across five metrics:

1. Bias: the difference between the average estimate of $\delta$ and $\delta$. All else being equal, zero bias is desirable.
2. RMSE: The root mean square error of the estimate of $\delta$. This metric measures the accuracy of the estimate; roughly speaking, the absolute difference between the estimate of $\delta$ and $\delta$ will be less than or equal to one RMSE about two-thirds of the time and two RMSEs about 95% of the time. Thus, a smaller RMSE is more desirable.
3. Log(*SE*/*SD*): the average natural logarithm of the estimated standard error of the estimate of $\delta$ divided by the standard deviation of the estimates of $\delta$. This metric measures the accuracy of the estimated standard errors; values above (or, conversely, below) zero give, roughly speaking, the percentage by which they are too large (or, conversely, small), and thus zero log(*SE*/*SD*) is desirable.
4. Coverage percentage: the percentage of the estimated 95% confidence intervals that cover $\delta$; values above (or, conversely, below) 95% imply intervals that are too wide (or, conversely, narrow), and thus 95% is desirable.
5. Coverage width: the average width of the estimated 95% confidence intervals; conditional on having an accurate 95% coverage percentage, smaller widths are more desirable.

Metrics were averaged over 1,000 repetitions of each simulation setting.

We note that proper model evaluation inherently requires the assessment of model performance across a variety of settings and metrics because models may perform well in some respects but poorly in others. However, for the evaluation to be sensible, it is critical that the settings and metrics are chosen reasonably. Thus, we provide some justification for our choice of settings and metrics. Our principal simulation parameters, the relative likelihood that a study with results that are not statistically significant is published *q* and the degree of heterogeneity $\tau$, were chosen based on theoretical considerations that allow us to make predictions about the performance of the various methods. As discussed in the Modeling Considerations section, when *q* and $\tau$ are greater than zero, methods such as the Hedges (1984) approach and the *p*-curve and *p*-uniform approaches that falsely assume both to be zero will be inefficient and biased, respectively. The actual parameter value settings of *q* and $\tau$ we use were, as discussed in the subsections below and as is desirable, chosen to span the range typical in behavioral research.

Settings for our other key parameters, such as the sample size per condition in a given study, the number of published studies, and $\delta$, were also chosen to span the range typical in behavioral research. We note that, qualitatively speaking, our results are not sensitive to the exact distribution assumed for the sample size per condition, provided that the sample size per condition is not unreasonably low (and thus outside the range typical in behavioral research). We further note that we did not examine the case of $\delta = 0$, given that we find it unreasonable both generally (Cohen, 1994; Tukey, 1991) and especially in the case of one-sided selection examined here, because being directionally consistent with $\delta$ is meaningless when $\delta = 0$. We finally note that our results should not be particularly sensitive to other settings, such as the equality of the sample size across conditions and the size of the test $\alpha$, provided that they are chosen reasonably.

Our metrics were chosen to assess estimates of $\delta$ as well as estimates of uncertainty (i.e., standard errors and confidence intervals) from multiple perspectives.

Although these metrics are comprehensive, they are by no means exhaustive, and others are certainly possible. Nonetheless, we believe that they assess the aspects of model performance most germane to readers. As noted, metrics were averaged over 1,000 repetitions of each simulation setting. This was a sufficiently large number of repetitions to ensure that simulation error was effectively negligible: Simulation error was less than 0.01 (often substantially less so) for each simulation setting for each model for each metric and was always much smaller than any important differences among models we discuss. Finally, estimation convergence issues were also negligible: Estimates nearly always converged for all 1,000 repetitions of each simulation setting for each model and failed to converge less than 0.05% of the time in the worst simulation setting/model pair.

For the reader who wishes to focus on a single metric, we suggest focusing on RMSE, given that it evaluates models on the criterion that ultimately matters most—namely, estimation accuracy. We urge such readers to also consider coverage percentage, because it evaluates interval estimation accuracy and—for the reader interested in such things—has implications for null hypothesis significance testing.

We also note that simply comparing an estimate to the true value, either on a single simulation iteration or on average across many simulation iterations, is equivalent to focusing solely on bias, and bias is a particularly poor metric for evaluating models with respect to estimation accuracy (as well as more generally) because biased estimators can and often do perform exceedingly well in both theory and practice. Although it may seem counterintuitive that an estimator that is systematically wrong on average (i.e., biased) would be preferable to one that is correct on average (i.e., unbiased), a biased but more stable (i.e., lower variance) estimator is very often preferable to an unbiased but less stable one. This holds because a slightly biased but very stable estimator will typically yield an estimate that is close to the true value even if tends to be systematically wrong on average, whereas an unbiased but very unstable estimator will typically yield an estimate that is far from the true value even if tends to be systematically correct on average.

Finally, we note that in our discussion of our simulation results, we generally focus on simulation settings with small and medium effect sizes $\delta$ because this is the most interesting setting for selection methods; in contrast, when each study has very high power, whether because of large effect sizes $\delta$ or large sample sizes, publication bias is much less problematic (i.e., because each study is very likely to have results that are statistically significant). Results for all simulation settings can be found in our Supplemental Material available online.

## *Simulation 1*

We begin by investigating the setting of Hedges (1984) in which (a) only studies with results that are statistically significant and directionally consistent are published and (b) effect sizes are homogeneous across studies (i.e., $q = 0$ and $\tau = 0$, respectively). Because the principled maximum likelihood estimation strategy employed by the Hedges (1984) approach yields an asymptotically minimum variance unbiased estimator and this is a one-parameter setting, we have strong reason to believe it will outperform the alternative estimation strategies employed by the *p*-curve and *p*-uniform approaches, which were also designed for this setting. Thus, the purpose of this simulation is to quantify the degree to which it outperforms them and investigate whether finite sample issues cause it ever to not do so.

We present our results in Figure 1 (for more details, see our Supplemental Material). In terms of bias, all methods perform similarly: Each has a small negative bias that decreases as the effect size $\delta$ and the number of studies increase. This bias is negligible in all simulation settings and for all methods. Specifically, it is asymptotically zero for the Hedges (1984) and *p*-curve approaches (because they correctly assume a noncentral *t* distribution as the sampling distribution of the individual study *t* statistics) and is asymptotically trivial for the *p*-uniform approach (because it employs an accurate though slightly biased normal approximation to the noncentral *t* distribution).

In terms of RMSE, all methods perform reasonably well, but the maximum likelihood selection method approach of Hedges (1984) outperforms the *p*-uniform approach, which in turn outperforms the *p*-curve approach. In particular, the *p*-curve approach performs 6% to 15% worse and the *p*-uniform approach performs 4% to 9% worse than the Hedges (1984) approach, depending on the simulation setting; further, the *p*-curve approach performs 10% worse and the *p*-uniform performs approach 7% worse on average across all simulation settings. The relative degree to which the Hedges (1984) approach outperforms the other two approaches decreases as the effect size $\delta$ increases but does not appear to substantially depend on the number of studies.

Given that only the Hedges (1984) approach produces a standard error, it is the only method that can be evaluated in term of $\log(SE/SD)$. As can be seen, the estimated standard error is quite on target: It is always within 5% of the standard deviation of the estimates of $\delta$ and is generally much closer. It appears to improve as the effect size $\delta$ and the number of studies increase.

Both the *p*-uniform and Hedges (1984) approaches produce 95% confidence intervals that are properly calibrated (i.e., they have coverage percentage equal to

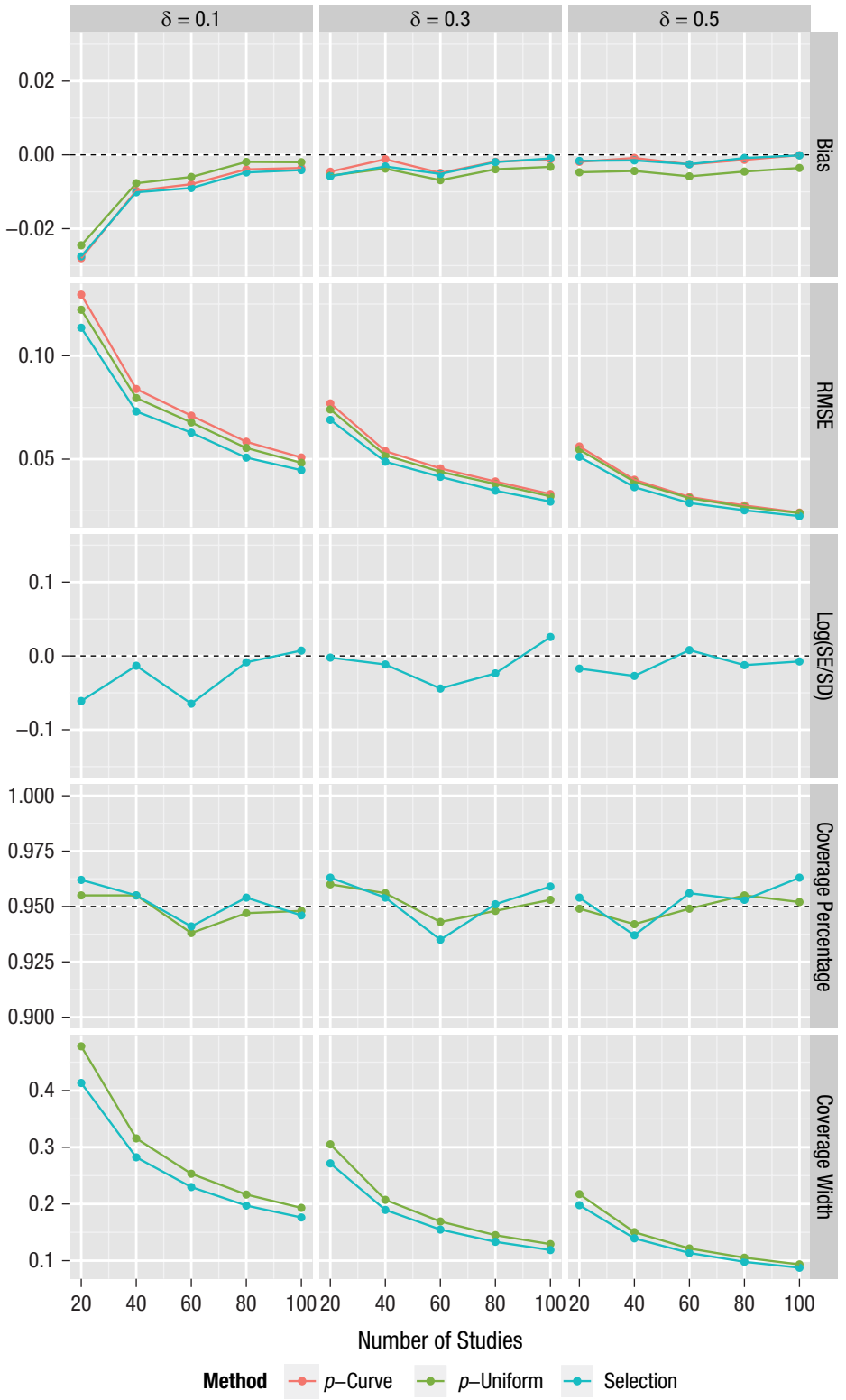


**Fig. 1.** Simulation 1 results. The figure plots the bias, root mean square error (RMSE), log(*SE*/*SD*), coverage percentage, and coverage width of the three methods for three values of δ as a function of the number of studies. The *p*-curve and *p*-uniform approaches perform modestly worse than the maximum likelihood selection method approach of Hedges (1984) because of the alternative estimation strategies they employ. For more details, see our Supplemental Material.

95%). However, the intervals produced by the Hedges (1984) approach are superior to those produced by the *p*-uniform approach because they maintain 95% coverage while being narrower (i.e., they have smaller coverage width). Indeed, they tend to be 5% to 16% narrower depending on the simulation setting and 10% narrower on average across all simulation settings.

In sum, all three methods perform well in this setting for which they were designed. However, the Hedges (1984) approach performs best, modestly though non-trivially outperforming the other two approaches.

Before proceeding, we note that when the selection model rigidly assumes that only studies with results that are both statistically significant and directionally consistent are published (as here), estimation difficulties can arise. Though they did not arise in this simulation, for completeness, we provide a discussion of these difficulties in Appendix A.

## *Simulation 2*

We now relax the assumption that only studies with results that are statistically significant are published; in particular, we assume that the likelihood that a study with results that are not both statistically significant and directionally consistent is published is respectively one-tenth or one-fourth that of a study with results that are statistically significant and directionally consistent (i.e., $q = .10$ or .25, respectively). However, we still assume that effect sizes are homogeneous across studies (i.e., $\tau = 0$). This is the setting of the method of Iyengar and Greenhouse (1988).

We note that the value of $q$ chosen does not mean that studies with results that are not statistically significant make up $q \times 100\%$ of published studies; rather, it means that studies with results that are not statistically significant are $q \times 100\%$ as likely to be published as those with results that are statistically significant. The percentage of studies with results that are not statistically significant among the published studies will vary based on the effect size $\delta$ and can in principle be quite small even when $q$ is large (e.g., if $\delta$ is large). We believe the values of $q$ chosen here span the range typical in behavioral research.

We also note that the data model assumed by all three approaches is correct in this setting. However, unlike the Iyengar and Greenhouse (1988) approach, the *p*-curve and *p*-uniform approaches employ a rigid selection model that assumes that only studies with results that are both statistically significant and directionally consistent are published; this selection model is clearly incorrect in this setting, thereby causing these approaches to ignore data, which in turn causes them to be inefficient (i.e., to yield noisier estimates of the population average effect size). Given this, the Iyengar and Greenhouse (1988) approach should outperform the *p*-curve and *p*-uniform approaches in this simulation. Therefore, the purpose of this simulation is to quantify the degree to which it outperforms them and to investigate whether finite sample issues cause it ever to not do so. We note that the degree to which it outperforms them should be larger than in Simulation 1, in which efficiency issues were not a concern and all three approaches were correctly specified; we also note that it will decrease as the relative likelihood of publication $q$ decreases and the effect size $\delta$ increases, because the former implies more rigid selection and the latter implies selection is much less problematic.

We present our results in Figure 2. We note that *y*-axis limits have been set so that, in some cases, methods that perform particularly poorly in a given simulation setting are partially excluded from the plot; for more details, see Appendix B and our Supplemental Material. The *p*-curve and *p*-uniform approaches exhibit a negative bias that is especially pronounced when both the effect size $\delta$ and the number of studies are small. Further, they yield inaccurate estimates: The RMSE of these approaches is much larger than that of the Iyengar and Greenhouse (1988) approach, and this is particularly the case when the effect size $\delta$ is small, when the relative likelihood of publication $q$ is moderate, and when the number of studies is small (see Appendix B). The inaccurate estimates are a direct consequence of the loss of efficiency that results from ignoring data.

The Iyengar and Greenhouse (1988) approach produces reasonable standard error estimates and, along with the *p*-uniform approach, calibrated confidence intervals; nonetheless, the Iyengar and Greenhouse (1988) approach is able to achieve calibration with narrower intervals (much narrower intervals when the effect size $\delta$ is small).

In sum, the loss of efficiency associated with the *p*-curve and *p*-uniform approaches results in poor performance in this setting. Indeed, the performance of these methods is acceptable relative to that of the Iyengar and Greenhouse (1988) approach only when the relative likelihood of publication $q$ is small (i.e., much below .10, so that selection is rigid) or the effect size $\delta$ is relatively large (i.e., so that selection is much less problematic).

## *Simulation 3*

We now relax the assumption that effect sizes are homogeneous; in particular, we assume that the true effect size in a given study is normally distributed with mean $\delta$ and standard deviation $\tau = .20$. This is the setting of the method of Iyengar and Greenhouse (1988) expanded to accommodate heterogeneity as discussed in the comments to and rejoinder of that article.

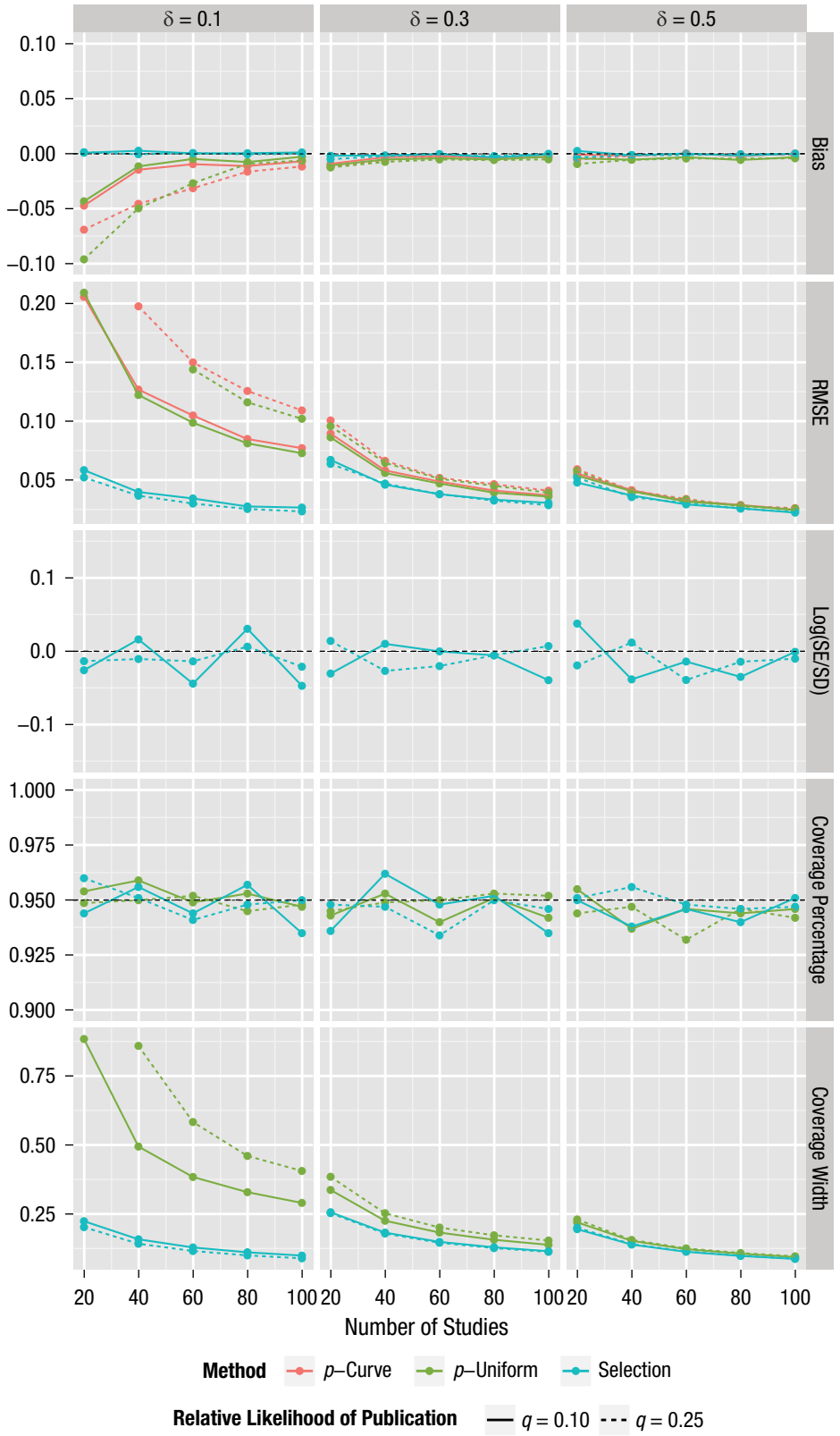


**Fig. 2.** Simulation 2 results. The figure plots the bias, root-mean-square error (RMSE), log(*SE*/*SD*), coverage percentage, and coverage width of the three methods for three values of δ as a function of the number of studies and the relative likelihood of publication *q*. The *p*-curve and *p*-uniform approaches perform considerably worse than the maximum likelihood selection method approach of Iyengar and Greenhouse (1988) because of loss of efficiency. For more details, see our Supplemental Material.

We note that setting $\tau$ to .20 means that roughly half the total variability on average across the simulated studies is due to heterogeneity (i.e., $I^2 \approx 50\%$). Thus, according to the taxonomy of Pigott (2012), this is a medium amount of heterogeneity (she defines low heterogeneity as $I^2 = 25\%$, medium heterogeneity as $I^2 = 50\%$, and high heterogeneity as $I^2 = 75\%$).

Before proceeding to our results, we note that approaches that falsely assume homogenous effects will here, by Jensen's inequality (Jensen, 1906), produce upwardly biased estimates of the population average effect size. They will also produce downwardly biased estimates of standard errors and confidence intervals. Thus, one purpose of this simulation is to quantify the degree of these biases.

We present our results in Figure 3. We note that, as for the results of Simulation 2, *y*-axis limits have been set so that, in some cases, methods that perform particularly poorly in a given simulation setting are partially excluded from the plot; for more details, see Appendix B and our Supplemental Material. The *p*-curve and *p*-uniform approaches exhibit a positive bias that is especially pronounced when the effect size $\delta$ is small. Further, they yield inaccurate estimates: The RMSE of these approaches is much larger than that of the maximum likelihood selection method approach.

The maximum likelihood selection method approach produces standard error estimates that are too small, and this effect is particularly pronounced when the effect size $\delta$ is large and the number of studies is small. This results in confidence intervals that have coverage somewhat below the nominal coverage percentage. On the other hand, the coverage percentage of the intervals produced by the *p*-uniform approach is extremely poor (see Appendix B).

Although we have focused on estimation of $\delta$ in this and prior simulations, we briefly wish to show that selection also has implications for estimation of heterogeneity. Rather than perform a full evaluation, as estimation of heterogeneity is not our focus, we simply present the average estimate of $\tau$ in a given simulation setting in Figure 4 for the selection method as well as a standard meta-analytic method (for more details, see our Supplemental Material). As can be seen, the standard meta-analytic approach tends to underestimate heterogeneity on average when there is selection; however, this result is not uniform, and indeed it sometimes demonstrates an upward bias. On the other hand, the maximum likelihood selection method approach does reasonably well with sufficient data.

In sum, the *p*-curve and *p*-uniform approaches perform poorly in this setting. On the other hand, the maximum likelihood selection method approach produces reasonable estimates but standard errors that are too small (and thus confidence intervals that are too narrow) in some settings. As it turns out, one-sided selection combined with heterogeneity can make estimation particularly difficult (see Appendix A). Further, Figure 4 demonstrates that estimates from standard meta-analytic approaches cannot be relied upon to assess heterogeneity when there is selection.

We note that these results are not idiosyncratic to the value of $\tau$ presented. Results for $\tau = .10$ (i.e., $I^2 \approx 25\%$, or low heterogeneity) and $\tau = .40$ (i.e., $I^2 \approx 75\%$, or high heterogeneity) were qualitatively similar, although the performance of the *p*-curve and uniform approaches declined as heterogeneity increased (for more details, see our Supplemental Material).

### *Summary*

The *p*-curve and *p*-uniform approaches perform reasonably well in the setting for which they were designed, namely, that of Simulation 1. However, as a result of the alternative estimation strategies they employ, they do not perform as well as the original Hedges (1984) approach. Further, as shown and quantified in Simulations 2 and 3, they are sensitive to deviations from their model assumptions: When selection is slightly less rigid (i.e., studies with results that are not statistically significant are published with some small probability) or effect sizes are heterogeneous, these methods perform poorly (less rigid selection causes them to be inefficient because they ignore data, whereas heterogeneity implies, by Jensen's inequality, that they are upwardly biased). In contrast, the Hedges (1984) approach generalizes easily to accommodate these features, and thus variants of it perform well in these more realistic settings.

Given that studies with results that are not statistically significant are indeed published and that heterogeneity is the norm in behavioral research, we should indeed be circumspect about the application of one-parameter approaches such as the *p*-curve and *p*-uniform approaches, as well as the original Hedges (1984) approach—particularly since standard meta-analytic approaches cannot be relied on to assess heterogeneity when there is selection. At the very least, the simple three-parameter variant of the Hedges (1984) approach used in Simulation 3 (i.e., the method of Iyengar & Greenhouse, 1988, but accounting for heterogeneity) should be the minimal model considered in applied work.

## Discussion

Although selection method approaches to assess and adjust for publication biases in meta-analysis have a long history that dates back over 30 years, they have yet to be subject to an extensive simulation study that covers multiple continuously related settings and that evaluates

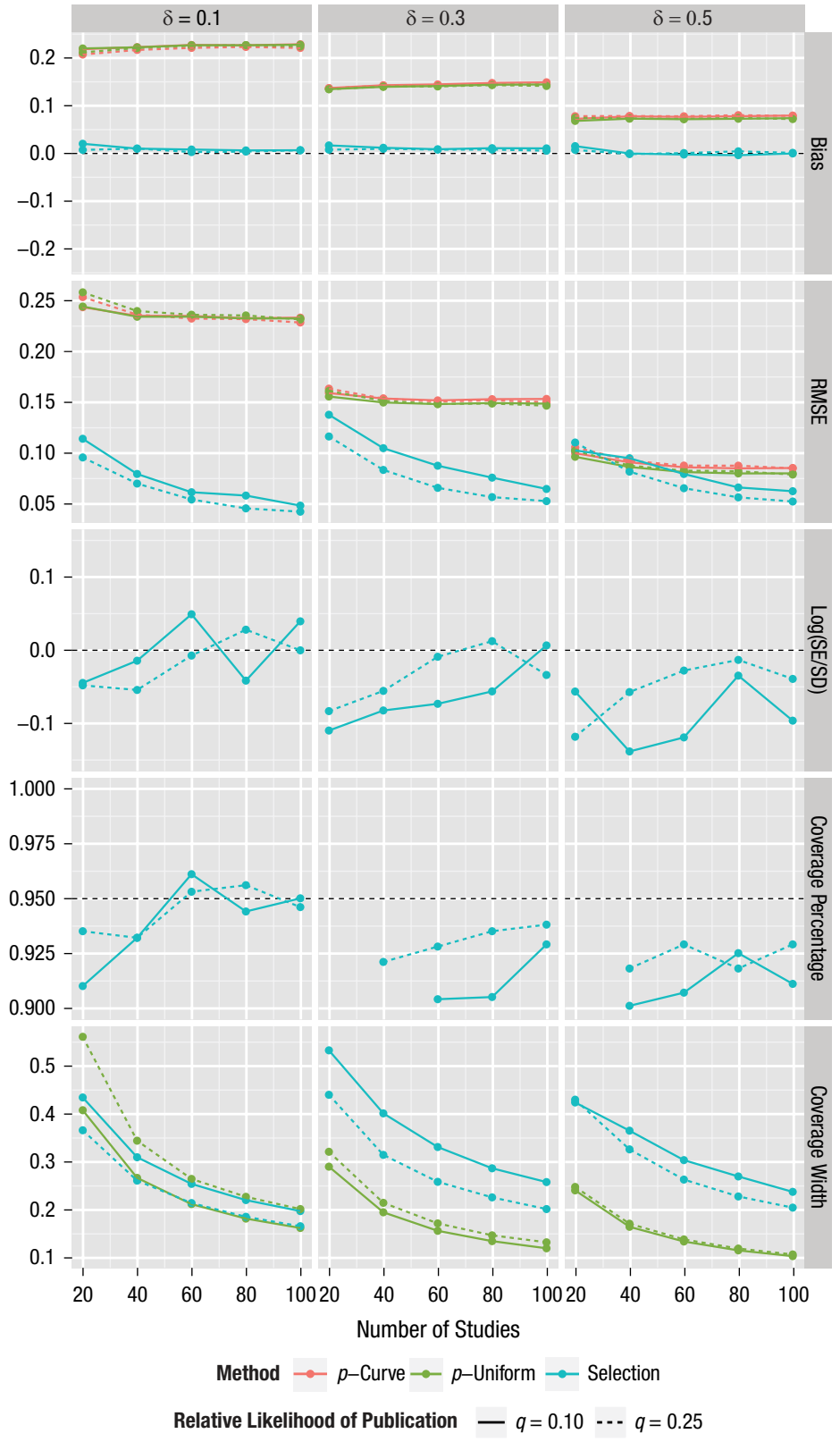


**Fig. 3.** Simulation 3 results. The figure plots the bias, root-mean-square error (RMSE), log(*SE/SD*), coverage percentage, and coverage width of the three methods for three values of δ as a function of the number of studies and the relative likelihood of publication *q*. The *p*-curve and *p*-uniform approaches perform considerably worse than the maximum likelihood selection method approach because of heterogeneity. For more details, see our Supplemental Material.

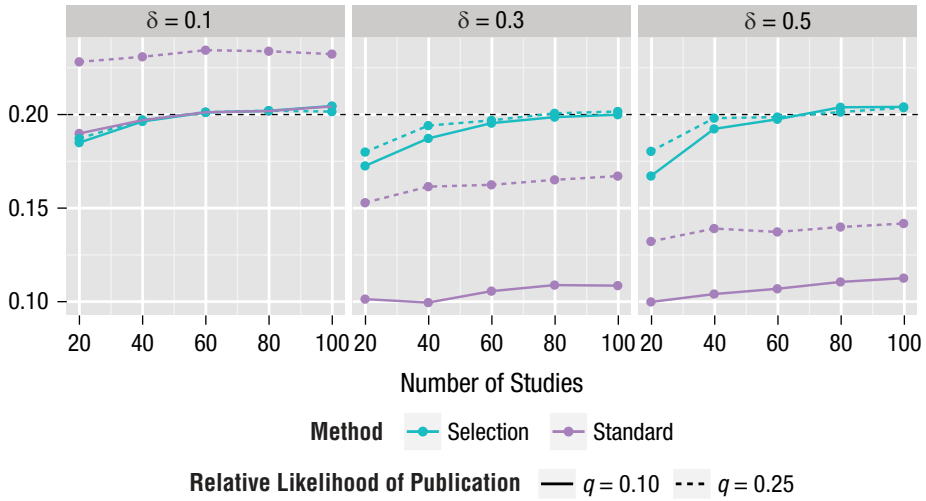


**Fig. 4.** Simulation 3 estimate of heterogeneity. The figure plots the average estimate of heterogeneity $\tau$ of both methods for three values of $\delta$ as a function of the number of studies and the relative likelihood of publication $q$. The standard meta-analytic approach tends to underestimate heterogeneity on average when there is selection, whereas the maximum likelihood selection method approach does reasonably well with sufficient data. For more details, see our Supplemental Material.

model performance across multiple dimensions. Such a comparative analysis, which can identify the strengths and weaknesses of various approaches and can provide statistical benchmarks against which new approaches can be assessed, is long overdue. Further, its benefits are clear: We find that the selection methods originated by Hedges (1984) and Iyengar and Greenhouse (1988) and extended over the past three decades remain state-of-the-art in adjusting for publication bias. Although the recently proposed *p*-curve and *p*-uniform approaches have increased awareness about the consequences of publication bias in meta-analysis, they fail to improve upon, and indeed are inferior to, methods proposed decades ago.

Despite the strong performance of selection methods in our simulations, we note that the assumptions underlying these methods and simulations are idealistic—even in the most realistic setting of Simulation 3. Further, even the more general selection methods discussed in the Generalized Selection Methods subsection above rely on assumptions that are likely to be quite idealistic in practice.

Real life data models and selection models are far more complicated, sequential, and iterative and involve not only authors but also editors and reviewers. For example, studies seldom have a single effect of interest; typically, studies have multiple effects of interest (e.g., a simple effect and an interaction effect in 2 × 2 studies), these multiple effects are dependent, and selection is likely to be based on the size, direction, and statistical significance of these multiple dependent effects jointly. In addition, studies are seldom independent. At minimum, the multiple studies of a common phenomenon that appear in a typical behavioral research article are likely to be dependent; more realistically, there are likely to be more complex dependencies among, for example, sets of studies conducted by the same authors across multiple articles, sets of studies using the same or similar materials, and sets of studies administered to the same pool of subjects. Further, there are likely multiple sequential selection mechanisms that operate on the studies that appear in a single article because, for example, the standards of the field may require that the first study reported in an article show a convincing simple effect, the second moderation, the third mediation, and the fourth application to a new domain. Finally, researchers engage in questionable research practices and make unintended errors in their single study analyses; these impact what studies are reported, what is reported from those studies, and how it is reported.

In theory, more general selection methods can be designed to account for all of these issues. However, the population average effect size estimates produced by selection methods can be highly sensitive to the data model and the selection model assumed (particularly the latter), and more realistic data models and selection models typically cannot be well estimated without a large amount of data. Moreover, even if such more general selection methods were practically tractable, they would still fail to account for issues of selection resulting from

the availability and accessibility of studies; such selection is (a) likely to be as important or more important than selection resulting from the size, direction, and statistical significance of study results and (b) far more difficult to model (even conceptually).

Thus, we are not optimistic that much more can be done under current data collection conditions. As R. A. Fisher noted in his 1938 presidential address to the first Indian Statistical Congress, "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of" (Fisher, 1938, p. 17). In the context of meta-analysis, the ex post application of selection methods is reasonable if and only if the data model and the selection model assumed are reasonably accurately specified and there is sufficient data; otherwise, selection methods simply cannot be relied upon or expected to provide accurate estimates.

Consequently, given the idealistic model assumptions underlying selection methods and the sensitivity of population average effect size estimates to them, we advocate that selection methods should be used less for obtaining a single estimate that purports to adjust for publication bias ex post and more for sensitivity analysis—that is, exploring the range of estimates that result from assuming different forms of and severity of publication bias. In particular, one can apply a variety of selection models that assume different forms of and severity of selection and then examine the variation in the resulting estimates. If the estimates are relatively stable regardless of the selection model assumed, this suggests that publication bias is unlikely to drive the unadjusted estimate. On the other hand, if the estimates vary considerably depending on the selection model assumed, this suggests that publication bias may well drive the unadjusted estimate.

We also advocate that the simple three-parameter variant of the approaches of Hedges (1984) and Iyengar and Greenhouse (1988) used in Simulation 3, which allows for the publication of studies with results that are not statistically significant and accounts for heterogeneity, should be the minimal model considered for sensitivity analysis in applied work, given that both of these are the norm in behavioral research. One-parameter approaches such as the *p*-curve and *p*-uniform approaches as well as the original Hedges (1984) approach, which allow for neither, are simply too unrealistic in behavioral research.

To facilitate sensitivity analyses based on the simple three-parameter model and more general selection methods, we note that Vevea and Woods (2005) is an excellent reference on how to use selection methods to conduct a sensitivity analysis to assess publication bias and further note that Hedges and Vevea (2005) provides additional examples (see also Copas, 1999; Copas & Shi, 2001; and Copas, 2013). We also note that the simple three-parameter model and generalizations of it have been implemented in the "weightr" package (Coburn & Vevea, 2016) for R (R Core Team, 2012) and on an easy-to-use website available at https://vevealab.shinyapps.io/WeightFunctionModel/. Conveniently, the default model currently implemented in the package and on the website is the simple three-parameter model (it specifies the *p*-value cutpoint in terms of a one-sided test, and thus the default cutpoint of .05 corresponds to a two-sided test at $\alpha = .10$; one should change the cutpoint to .025 if one wishes it to correspond to a two-sided test at $\alpha = .05$). Finally, we note that we provide our code as well as a simple example in our Supplemental Material.

Although we believe that the primary application of selection methods should not be obtaining a single estimate that purports to adjust for publication bias ex post, we note that there is an alternative perspective. This perspective agrees that even the more general selection methods discussed in the Generalized Selection Methods subsection rely on assumptions that are likely to be quite idealistic in practice. However, it emphasizes that they nonetheless capture an important aspect of the selection process. Thus, it concludes that, particularly in meta-analyses with large numbers of studies, an imperfectly adjusted estimate is better than an unadjusted estimate. We disagree with none of this in principle but argue that the adjustments provided by the more general methods are likely still too imperfect to be relied upon for estimation because they fail to account for the complexity of real life data models and selection models discussed above as well as selection resulting from the availability and accessibility of studies.

In closing, we discuss three final issues. First, although we have focused on selection methods, we note that selection methods are by no means the only class of techniques that have been proposed to assess and adjust for publication bias in meta-analysis. Other techniques include the funnel plot (Sterne, Becker, & Egger, 2005), nonparametric and regression-based tests (Begg & Mazumdar, 1994; Egger, Smith, Schneider, & Minder, 1997; Sterne & Egger, 2005), Rosenthal's fail-safe *N* method and its many variants (Becker, 2005; Rosenthal, 1979), the trim-and-fill method (Duval, 2005; Duval & Tweedie, 2000a, 2000b), and the Precision-Effect Test–Precision-Effect Estimate with Standard Error (PET-PEESE) method (Stanley & Doucouliagos, 2014). In contrast to selection methods, these alternative techniques do not posit a data model and a selection model—or any other statistical model—and thus lack many of the advantages of selection methods; indeed, the lack of a statistical model has been a point of criticism for these alternative techniques (see, for example, Becker, 2005). In addition, these alternative techniques have yet to be subject to an extensive evaluation. Without such an evaluation, it is not possible to know in what settings and on what dimensions these techniques perform well versus

poorly or how their performance compares to that of alternative techniques such as selection methods. Unfortunately, designing such an evaluation for these techniques is complicated by the fact that they do not posit an underlying statistical model. Further, our caution that the ex post application of selection methods is reasonable if and only if the data model and the selection model assumed are reasonably accurately specified and there is sufficient data would also seem to apply, mutatis mutandis, for these alternative techniques.

Second, given our belief that meta-analytic estimates of population average effect sizes are likely to be inaccurate because of selection issues, one may wonder whether we ever advocate conducting a meta-analysis. Our answer is a resounding "yes" because meta-analysis has much to offer beyond estimation of the population average effect size. It is useful for cataloguing the various study designs, dependent variables, moderators, and other methods factors used in studies in a given domain. In addition, it can—at least for the set of studies examined—quantify (a) the average effect size, (b) the degree of heterogeneity induced by differences in unaccounted for (and potentially unknown) method factors, and (c) the association between study results and accounted for, known method factors such as study-level moderators. Importantly, single study analyses can provide none of these benefits, so meta-analysis, however flawed, is the only option.

Third, we note that the quality of an estimate ultimately depends on the quality of the data used to produce that estimate as well as the purpose of the estimate. Thus, as noted, unadjusted meta-analytic estimates are still reliable when one does not seek to generalize to some larger population. They are also reliable if, for some reason, selection is independent of effect sizes and effect size estimates. Finally, they are reliable when all data are reported, as in the Many Labs project (Klein et al., 2014). Fortunately, we expect that practices like those employed by the Many Labs authors are becoming much more typical in behavioral research. Consequently, we believe that an ex ante multipronged preventive approach that includes training in statistical reasoning, preregistration of studies, use of more powerful study designs, and open data will prove superior to—both for meta-analysis and more broadly—any ex post statistical approach that attempts to adjust for potential biases after it may be too late.

## Appendix A

### *Estimation Issues*

In this section, we discuss difficulties with estimating selection methods when the selection model rigidly assumes that only studies with results that are both statistically significant and directionally consistent are published. We note that these difficulties do not apply when this extremely rigid one-sided selection model is relaxed (for instance, if only studies with results that are statistically significant—but not directionally consistent—are published or if studies with results that are not statistically significant are published with some probability). Consequently, these difficulties are unlikely to be encountered in practice.

Returning to the one-parameter setting of Hedges (1984) that assumes rigid one-sided selection and homogenous effect sizes, Hedges and Vevea (2005) pointed out that the likelihood function is poorly behaved when one observes a single data point that is small: It is relatively flat for small positive and negative values of $\delta$. Further, when this data point is so small that it is just barely statistically significant, the likelihood increases as $\delta$ decreases and is unbounded. Clearly, this is pathological.

We note that this issue is not in theory limited to only a single data point. When the combined effect size from multiple data points is small, the likelihood can be relatively flat for small positive and negative values of $\delta$ or can increase monotonically as $\delta$ decreases. The only solution to this problem is to increase precision, either by increasing the sample size per study or by increasing the total number of studies. However, we note that this is not likely to be a problem in practice because (a) a small number of studies will generally provide enough precision to make the likelihood function well behaved and (b) this extremely rigid one-sided selection model with homogeneous effects rarely if ever holds in behavioral research and thus should not be considered.

We also note that this problem holds even if a normal distribution is assumed for the observed effects rather than a noncentral *t* distribution and that it is a theoretical issue distinct from any numerical issues involved in computing tail probabilities. We further note that, like the maximum likelihood estimation strategy employed by the Hedges (1984) approach, the estimation strategies employed by the *p*-curve and *p*-uniform approaches also perform pathologically in this setting (van Aert et al., 2016), presumably as a result of an implicit dependence on the likelihood.

Extending to the two-parameter setting that assumes rigid one-sided selection and heterogeneous effect sizes (i.e., the setting of Hedges, 1984, but accounting for heterogeneity), the problem becomes even more intractable. In particular, there is a near ridge in the likelihood function along $\delta$ and $\tau$, which results in pathological behavior favoring large negative values of $\delta$ and large positive values of $\tau$. This problem actually follows from the problem associated with the one-parameter setting. In particular, for fixed $\tau$, the one-dimensional profile of the two-parameter likelihood viewed as a function of $\delta$ is akin to the one-parameter likelihood function (and is indeed identical to it for $\tau = 0$); however, the

pathological behavior of the one-parameter likelihood function occurs much more commonly when $\tau$ is set large as compared to when $\tau$ is set to zero, thus compounding matters. Again, the solution to this problem is to increase precision. Also again, this is not likely to be a problem in practice because the extremely rigid one-sided selection model rarely if ever holds in behavioral research.

We note that penalized likelihood estimation strategies and Bayesian estimation strategies with informative priors improve performance in this setting by keeping effect size estimates bounded. Indeed, informative priors are particularly appropriate here because the effect size is measured on the standardized Cohen's *d* scale. Consequently, these estimation strategies should be considered in future research when the rigid one-sided selection model applies.

## Appendix B

**Table: Data Points Excluded From Figures 2 and 3**

| Simulation | Metric | Method | $\delta$ | $q$ | Number of studies | Value |
|---|---|---|---|---|---|---|
| 2 | RMSE | *p*-curve | .1 | .25 | 20 | 0.277 |
| 2 | RMSE | *p*-uniform | .1 | .25 | 20 | 0.443 |
| 2 | RMSE | *p*-uniform | .1 | .25 | 40 | 0.275 |
| 2 | Coverage width | *p*-uniform | .1 | .25 | 20 | 1.848 |
| 3 | Coverage percentage | *p*-uniform | .1 | .10 | 20 | 0.472 |
| 3 | Coverage percentage | *p*-uniform | .1 | .10 | 40 | 0.196 |
| 3 | Coverage percentage | *p*-uniform | .1 | .10 | 60 | 0.070 |
| 3 | Coverage percentage | *p*-uniform | .1 | .10 | 80 | 0.033 |
| 3 | Coverage percentage | *p*-uniform | .1 | .10 | 100 | 0.009 |
| 3 | Coverage percentage | *p*-uniform | .1 | .25 | 20 | 0.576 |
| 3 | Coverage percentage | *p*-uniform | .1 | .25 | 40 | 0.376 |
| 3 | Coverage percentage | *p*-uniform | .1 | .25 | 60 | 0.185 |
| 3 | Coverage percentage | *p*-uniform | .1 | .25 | 80 | 0.104 |
| 3 | Coverage percentage | *p*-uniform | .1 | .25 | 100 | 0.060 |
| 3 | Coverage percentage | *p*-uniform | .3 | .10 | 20 | 0.553 |
| 3 | Coverage percentage | Selection | .3 | .10 | 20 | 0.866 |
| 3 | Coverage percentage | *p*-uniform | .3 | .10 | 40 | 0.258 |
| 3 | Coverage percentage | Selection | .3 | .10 | 40 | 0.874 |
| 3 | Coverage percentage | *p*-uniform | .3 | .10 | 60 | 0.114 |
| 3 | Coverage percentage | *p*-uniform | .3 | .10 | 80 | 0.042 |
| 3 | Coverage percentage | *p*-uniform | .3 | .10 | 100 | 0.012 |
| 3 | Coverage percentage | *p*-uniform | .3 | .25 | 20 | 0.593 |
| 3 | Coverage percentage | Selection | .3 | .25 | 20 | 0.896 |
| 3 | Coverage percentage | *p*-uniform | .3 | .25 | 40 | 0.330 |
| 3 | Coverage percentage | *p*-uniform | .3 | .25 | 60 | 0.177 |
| 3 | Coverage percentage | *p*-uniform | .3 | .25 | 80 | 0.073 |
| 3 | Coverage percentage | *p*-uniform | .3 | .25 | 100 | 0.043 |
| 3 | Coverage percentage | *p*-uniform | .5 | .10 | 20 | 0.748 |
| 3 | Coverage percentage | Selection | .5 | .10 | 20 | 0.887 |
| 3 | Coverage percentage | *p*-uniform | .5 | .10 | 40 | 0.564 |
| 3 | Coverage percentage | *p*-uniform | .5 | .10 | 60 | 0.455 |
| 3 | Coverage percentage | *p*-uniform | .5 | .10 | 80 | 0.338 |
| 3 | Coverage percentage | *p*-uniform | .5 | .10 | 100 | 0.241 |
| 3 | Coverage percentage | *p*-uniform | .5 | .25 | 20 | 0.737 |
| 3 | Coverage percentage | Selection | .5 | .25 | 20 | 0.893 |
| 3 | Coverage percentage | *p*-uniform | .5 | .25 | 40 | 0.595 |
| 3 | Coverage percentage | *p*-uniform | .5 | .25 | 60 | 0.468 |
| 3 | Coverage percentage | *p*-uniform | .5 | .25 | 80 | 0.327 |
| 3 | Coverage percentage | *p*-uniform | .5 | .25 | 100 | 0.269 |

Note: A horizontal line separates unique simulation/metric/$\delta$/$q$ combinations. The principal results are that (a) the *p*-curve and *p*-uniform approaches have poor root mean square error (RMSE) in Simulation 2 when the effect size $\delta$ is small, when the relative likelihood of publication $q$ is moderate, and when the number of studies is small and (b) the *p*-uniform approach has poor coverage percentage throughout Simulation 3.

## Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

## Supplemental Material

Additional supporting information may be found at http://pps.sagepub.com/content/by/supplemental-data

## References


Becker, B. J. (2005). Failsafe n or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 111–125). Chichester, England: John Wiley & Sons.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*, 1088–1101.

Bickel, P., & Doksum, K. (2007). *Mathematical statistics: Basic ideas and selected topics* (Vol. 1., 2nd ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Boyle, R. (1965). *The sceptical chymist*. London, England: Dawsons of Pall Mall. (Original work published 1661)

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Pacific Grove, CA: Duxbury.

Coburn, K. M., & Vevea, J. L. (2016). *Weightr: Estimating weight-function models for publication bias in r*. R package version 1.0.0. Retrieved from https://CRAN.R-project.org/package=weightr

Cohen, J. (1994). The earth is round ($p$ <.05). *American Psychologist*, *49*, 997–1003.

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

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

Copas, J. B., & Li, H. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*, 55–95.

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

Dear, K. B., & Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*, *7*, 237–245.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 11–33). Chichester, England: John Wiley & Sons.

Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 127–144). Chichester, England: John Wiley & Sons.

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89–98.

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

Editors. (1909). The reporting of unsuccessful cases [Editorial]. *The Boston Medical and Surgical Journal*, *161*, 263–264.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, *5*(4), e10068. doi:10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.

Ferriar, J., & Simmons, W. (1792). *Medical histories and reflection*. London, England: Cadell and Davies.

Fisher, Ronald A (1938). Presidential address, First Indian Statistical Conference. *Sankhya*, *4*, 14–17.

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*, *41*, 632–643.

Hall, M. B. (1959). In defense of experimental essays. In R. Boyle & M. B. Hall (Eds.), *Robert Boyle on natural philosophy: An essay, with selections from his writings* (pp. 119–131). Bloomington, IN: Indiana University Press.

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

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, *7*, 246–255.

Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 145–174). Chichester, England: John Wiley & Sons.

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

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes [On convex functions and inequality between the average values]. *Acta Mathematica*, *30*, 175–193.

Jin, Z.-C., Zhou, X.-H., & He, J. (2015). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine*, *34*, 343–360.

Klein, R. A., Ratliff, K., Nosek, B. A., Vianello, M., Pilati, R., Devos, T., . . . Kappes, H. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*, 107–112.

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, *9*, 612–625.

McShane, B. B., & Böckenholt, U. (2016). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods*, *21*, 47–60.

McShane, B. B., & Gal, D. (2016). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science*, *62*, 1707–1718.

Pigott, T. (2012). *Advances in meta-analysis*. New York, NY: Springer.

R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, England: John Wiley & Sons.

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage Publications.

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

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

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association*, *54*, 30–34.

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

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 75–98). Chichester, England: John Wiley & Sons.

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). Chichester, England: John Wiley & Sons.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*, 100–116.

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

van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*, 293–309.

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

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