Estimation and inference for step-function selection models in meta-analysis with dependent effects

James E. Pustejovsky[1], Martyna Citkowicz[2], and & Megha Joshi[2]

[1] University of Wisconsin-Madison

[2] American Institutes for Research

November 19, 2025

Author Note

Correspondence concerning this article should be addressed to James E. Pustejovsky, 1082C Educational Sciences, 1025 W Johnson St. Madison, WI 53706-1706. E-mail: pustejovsky@wisc.edu

Abstract

Meta-analyses in social science fields face multiple methodological challenges arising from how primary research studies are designed and reported. One challenge is that many primary studies report multiple relevant effect size estimates, leading to a data structure with dependent observations. Another is selective reporting bias, which arises when the availability of study findings is influenced by the statistical significance of results. Although many selective reporting diagnostics and bias-correction methods have been proposed, few are suitable for meta-analyses involving dependent effect sizes. Among available methods, step-function selection models are conceptually appealing and have shown promise in previous simulations. We study methods for estimating step-function models from data involving dependent effect sizes, focusing specifically on estimating parameters of the marginal distribution of effect sizes and accounting for dependence using cluster-robust variance estimation or bootstrap resampling. We describe two estimation strategies, demonstrate them by re-analyzing data from a previous synthesis on ego depletion effects, and evaluate their performance through an extensive simulation study under single-step selection. Simulation findings indicate that selection models provide low-bias estimates of average effect size and that clustered bootstrap confidence intervals provide acceptable coverage levels. However, adjusting for selective reporting bias using step-function models involves a bias-variance trade-off. Estimates from selection models and other adjustment methods are less accurate than unadjusted average effects when the strength of selective reporting is mild.

Keywords: meta-analysis; dependent effect sizes; selection models; selective reporting; publication bias

Highlights

What is already known

- Selective reporting of primary study results can distort the findings from meta-analyses.
- In many fields of application, meta-analyses include primary studies providing multiple, statistically dependent effect size estimates.
- Step-function selection models are a theoretically appealing tool for correcting bias from selective reporting, but these models were developed for meta-analyses that include only one effect size estimate per study.

What is new

- We examine strategies for estimating step-function selection models in meta-analyses with dependent effect sizes.
- We report a large simulation study demonstrating that step-function models can be estimated using composite penalized likelihood methods, combined with bootstrapping techniques to account for dependent effect size estimates.

Potential impact for RSM readers

- The proposed estimation methods can be applied to adjust for selective reporting of study results in meta-analyses that include dependent effect sizes.
- This strategy complements other existing methods for examining selective reporting bias in meta-analyses with dependent effect sizes.

1 Estimation and inference for step-function selection models in meta-analysis with dependent effects

The validity of conclusions from meta-analytic syntheses depend critically on the reporting practices of researchers, journal editors, and peer reviewers. If the findings accessible to meta-analysts are not a representative record of the research that has been conducted on a topic, then meta-analytic summaries may be systematically biased.[1] Of particular concern is the possibility that results from primary studies are selectively reported in ways that can distort the evidence available for synthesis, such as reporting findings that are statistically significant but omitting findings that are null or not consistent with researchers' hypotheses.[2,3]

We can conceptualize selective reporting practices as occurring at the level of the entire study or at the level of a specific findings within a study. At the study level, publication bias occurs when full study reports are withheld from public availability as a function of some aspect of their findings.[4] Reports with statistical conclusions that support research hypotheses may be more likely to be published than those that do not. Even if a study passes the hurdle of publication, the study might include results for only a subset of analyses conducted or outcome measures examined, leading to outcome reporting bias at the level of the individual finding.[5–7]

Evidence from medical, educational, and social sciences provides indications of the prevalence of selective reporting. For example, studies have found that statistically significant outcomes were 2.4 to 4.7 times more likely to be published than non-significant outcomes in the medical sciences[8] and 2.4 times more likely in education.[9] Studies examining various fields, including clinical trials of antipsychotics[10], psychology research[11–13], management science[14], economics, and environmental sciences[15], have found widespread selective outcome reporting on the basis of statistical significance.

Because selective reporting is of such central concern for meta-analysis, a wide range of statistical tools have been developed for assessing the presence of selective reporting and reducing the biases it creates.[1,16] One common graphical diagnostic is the funnel plot, a simple scatterplot of effect size estimates versus a measure of their precision.[17,18] Widely used statistical diagnostics include the rank correlation test[19]; Egger's regression test[20–23]; the trim-and-fill adjustment[24,25]; and various regression adjustment methods including the precision effect test (PET), precision effect estimate with standard error (PEESE), and PET-PEESE technique[26], and the endogenous kink meta-regression.[27]

Another class of methods for assessing and correcting selective reporting are p-value selection models. Such models build on summary meta-analysis or meta-regression models by making specific, explicit assumptions about the selection function, or how the probability that an effect size estimate is reported relates to sign and statistical significance level of the effect. Building on earlier proposals[28–30], Vevea and Hedges proposed *step-function* models where the selection function is piece-wise constant, with steps at fixed significance thresholds.[31,32] Other forms involve selection functions based on beta densities[33], power curves, and a variety of other parametric forms.[34]

Within the class of p-value selection models, much attention has focused on the Vevea-Hedges step-function model[31,32] because it captures simple but plausible forms of selective reporting, such as shifts in selection probability at the psychologically salient thresholds of $p = 0.05$ or $p = 0.01$[see, e.g., 2,35–37]. Step-function selection models have several advantages over other available methods for diagnosing and adjusting for selective reporting bias. First, they are generative models with parameters that directly describe the selective reporting process, making them more interpretable than tests or adjustments for small-study effects, which are agnostic with respect to the specific mechanism of selective reporting. Second, they can incorporate both discrete and continuous moderators, enabling one to distinguish between selective reporting bias and systematic differences in effect size that can

be predicted by primary study characteristics. Third, findings from simulations indicate that simple forms of step-function models outperform alternative bias adjustment methods when effect sizes are heterogeneous because they allow for the inclusion of a random effect term[38–40].

Step-function models also have drawbacks that should be acknowledged. Notably, they are more complex—and perhaps less intuitive—than regression-based adjustments or trim-and-fill. The thresholds must be specified a priori, and so may not correspond to the true selective reporting process. Furthermore, step-function models are built on the assumption that the selection process is homogeneous across study features other than statistical significance, such as study size or funding status. Just as with other selection models, they require a larger number of effect sizes for accurate estimation of the model parameters, particularly when a complex selection process is assumed. Despite these complexities, the plausible selective reporting process expressed by the step-function model, coupled with its demonstrated performance in past simulations[38,41], makes it a promising approach for analyzing and correcting the bias arising from selective reporting.

1.1 Dependent effect sizes

The vast majority of the work on selective reporting has focused on methods appropriate for relatively simple summary meta-analyses in which each included study contributes a single independent effect size estimate. This presents a problem for syntheses in education, psychology, and many other areas, where meta-analyses routinely include studies with multiple, dependent effect sizes. Effect size dependencies occur when multiple effect sizes are extracted from the same sample, resulting in statistically dependent estimates. Dependent effect size estimates commonly occur (1) when multiple outcome measures are collected on the same sample; (2) when the same sample is measured over multiple time points; or (3) when multiple treatment groups are compared to the same control group.[42] Dependence can also arise when effect sizes are extracted from multiple samples involving the

same operational features, such as multiple studies conducted by the same research group.[43]

Effect size dependencies are very common in social science synthesis, as well as in other research areas.[44] For example, for the more than 1,000 educational intervention studies reviewed by the What Works Clearinghouse since 2017, most (73%) included more than one intervention effect estimate, with a median of four effect sizes per study (WWC, 2020). Surveys of systematic reviews on topics in psychology and education[45], environmental sciences[46], and neurobiology[47] have also documented a high prevalence of dependent effect sizes. Thus, multiple effects are the norm, rather than the exception, in many fields that use quantitative synthesis.

Meta-analysts now have access to an array of methods for summarizing and modeling dependent effect sizes, including multivariate meta-analysis[48–50], multi-level meta-analyses[51–53], robust variance estimation[RVE, 43], and combinations thereof.[54] Among these, RVE has proven to be an attractive strategy because it provides a means to assess uncertainty in model parameter estimates that does not rely on strong assumptions about the exact dependence structure of the effect size estimates. Instead, RVE involves specifying a tentative working model for the dependence, but calculating standard errors, hypothesis tests, and confidence intervals using sandwich estimators that do not require the working model to be correct. Although the original form of RVE required a relatively large number of independent studies, subsequent work has provided refinements to standard errors and hypothesis tests based on RVE to provide accurate inferences even when the number of studies is small.[55,56] A closely related strategy is to use bootstrap re-sampling to approximate the distribution of test statistics.[57] However, extant developments in RVE and bootstrapping methods are limited to summary meta-analysis and meta-regression models. Applications to selection models remain to be explored.

1.2 Investigating selective reporting with dependent effect sizes

Methodologists have only recently begun to examine selective reporting detection or bias correction methods in meta-analysis involving dependent effect sizes. Mathur and VanderWeele proposed a sensitivity analysis based on the simplest possible form of the step-function selection model.[58] This sensitivity analysis provides an estimate of the average effect size after correcting for selective reporting based on a single, threshold statistical significance level, where the maximum strength of selection is pre-specified by the analyst. It handles effect size dependence using RVE.[43] However, this approach is premised on an assumed degree of selective reporting; thus, it does not estimate the strength of selection, nor does it have extensions to more complex forms of selection models (such as step functions with multiple thresholds).

Another alternative is to use a regression test for small-study effects, or association between effect sizes and standard errors, combined with RVE or multilevel meta-analysis to handle dependent effect sizes.[59,60] This method has limited power to detect selective reporting under common meta-analytic scenarios.[60] It also has the same limitations as univariate Egger's regression, in that it tests for a pattern of small-study effects, which could have causes other than selective reporting.[61] Further, Egger's regression is not based on a generative model and is therefore not directly informative about the degree or pattern of selective reporting.

Chen and Pustejovsky reviewed a range of existing techniques for estimating average effect sizes in the presence of selective reporting and dependent effect sizes.[62] They also proposed adaptations of several existing methods that can be formulated as meta-regressions, such as PET/PEESE and the endogenous kink method, with dependence addressed using a particular working model combined with RVE. In an extensive simulation study, they examined the performance of proposed adaptations alongside existing methods that ignore effect size dependence, under scenarios where the selective reporting process was consistent

with a one-step or two-step selection model. Although no single bias-correction method performed best across all conditions examined, simple forms of the step-function selection model emerged as strong candidates. Across a wide range of conditions, one-step and two-step selection models yielded average effect size estimates with low bias that were usually more accurate than alternative bias-adjusted estimators. The strong performance of step-function models is partially attributable to alignment between the assumptions of the step-function model and the mechanism used to introduce selective reporting in the data-generating process of the simulations. However, because the step-function models involve the assumption that all effect sizes are independent, confidence intervals generated from the selection models did not have accurate coverage. These findings indicate a need to further develop selection models that can account for dependent effect sizes.[62]

To address this need, we investigate how to estimate selection models and provide valid assessments of uncertainty in parameter estimates for meta-analyses that involve dependent effect sizes. In applying selection models to datasets involving dependent effects, we propose to model the *marginal* distribution of the effect size estimates rather than the joint distribution of the dependent effects within each study. This means we model each estimate considered as a single observation, without explicitly accounting for its dependence on other effect sizes from the same study. To account for dependence, we consider cluster-robust variance estimation or clustered bootstrap inference methods that allow for dependent observations even though the dependence is not explicitly modeled. This strategy does have the limitation that the model parameter estimates pertain only to the marginal distribution and do not distinguish between selective publication of full studies versus selective reporting of individual outcomes.

We believe that the strategy of modeling the marginal distribution is worth pursuing for at least three reasons. First, this strategy has several precedents in the meta-analysis and broader statistical literature, including generalized estimating equations with working

independence structures for analysis of longitudinal data[63], the unrestricted weighted least squares approach for meta-regression[64], and pseudo-likelihood estimators for multivariate meta-analysis[65]. Second, focusing on the marginal distribution allows for computational tractability, whereas fitting even quite basic multivariate models would require computing selection probabilities over high-dimensional spaces of possible outcomes. Third, although this strategy is simple, it still captures a plausible form of selection, in which reporting is influenced by the significance level of individual effect size estimates. In contrast, developing a multivariate selection model would require specifying assumptions about the joint probability that different subsets of effect sizes are reported or censored. Little past work on reporting practices has considered reporting of multiple findings within a study, so there is currently little empirical basis to support assumptions about multivariate selection processes. Moreover, it seems reasonable to assume that the statistical significance level of individual effect size estimates would still be a major consideration in more nuanced, multivariate selection models. Thus, we focus on developing marginal step-function selection models as a practical and feasible tool, which could also serve as a building block for more complex multivariate models.

The remainder of the paper is organized as follows. In the next section, we describe the step-function selection model and detail two different strategies for estimating model parameters: one using penalized maximum likelihood estimation methods and a novel strategy based on a re-weighted random effects model with inverse probability of selection weights. We also describe extensions of RVE and bootstrap re-sampling techniques to assess uncertainty in selection model parameter estimates. In the following section, we provide an empirical example that illustrates the methods by re-analyzing data from a previously reported meta-analysis. In subsequent sections, we describe the methods and results from a simulation study that evaluates the performance of point estimators and confidence intervals across a wide range of meta-analytic conditions. In the final section, we discuss findings, limitations, and initial implications for practice.

2 Models and Estimation Methods

Step-function selection models involve two components.[66] The first component, which we shall call the evidence-generating process, involves assumptions about the distribution of effect size estimates prior to selective reporting. This component is usually a random effects model or meta-regression model. The second component, which we shall call the selection process, involves assumptions about how effect size estimates come to be observed and therefore available for inclusion in the meta-analysis. Combining the assumptions of both components leads to a model for the distribution of observed effect size estimates, with interpretable parameters describing both the evidence-generating process and the selection process.

We will use the following notation to describe the model and estimation methods. Consider a meta-analytic sample consisting of J studies, in which study j reports k_j effect size estimates. Let y_{ij} denote observed effect size estimate i from study j, defined so that positive values are consistent with theoretical expectations. Each estimate is accompanied by a standard error σ_{ij} and corresponding one-sided p-value p_{ij}, where the one-sided p-value is defined with respect to the null hypothesis that the effect is less than or equal to zero and the alternative hypothesis that the effect is positive. Let $\mathbf{x}_{ij}$ be a $1 \times x$ row-vector of predictors that encode characteristics of the effect sizes, samples, or study procedures. Let $\Phi()$ denote the standard normal cumulative distribution function and $\phi()$ denote the standard normal density function.

2.1 Evidence-generating process

We consider an evidence-generating process based on a standard meta-regression model. Let Y^* denote an effect size estimate that has been generated from primary study data but might or might not be reported; let σ^*, p^*, and $\mathbf{x}^*$ denote the corresponding standard error, one-sided p-value, and predictor vector. The evidence-generating process can

then be expressed as

$$(Y^*|\sigma^*, \mathbf{x}^*) \sim N\left(\mathbf{x}^*\boldsymbol{\beta},\ \tau^2 + \sigma^{*2}\right), \tag{1}$$

where $\boldsymbol{\beta}$ is an $x \times 1$ vector of regression coefficients that relate the predictors to average effect size and τ^2 is the variance of the distribution of effect size parameters. This random-effects meta-regression treats each observed effect size as if it were independent, even though the data may include multiple, statistically dependent effect size estimates generated from the same sample. As a result, the regression coefficients $\boldsymbol{\beta}$ describe the overall expected effect size (given the predictors) and the variance parameter τ^2 describes the marginal or *total* heterogeneity of the effect size distribution, rather than decomposing the heterogeneity into within-study and between-study components.

2.2 Selection process

At a general level, a p-value selection process is defined by a selection function, which specifies the probability that an effect size estimate is reported given its p-value. Letting O be an indicator for whether the effect size estimate Y^* is observed, the selection process defines $\Pr(O = 1 \mid p^*) = \Pr(O = 1 \mid Y^*, \sigma^*)$. In particular, such models assume that $\Pr(O = 1|p^*)$ is proportional to a function $w(p^*; \boldsymbol{\lambda})$ that maps p-values in the interval $[0, 1]$ to strictly positive weights and that involves an unknown $h \times 1$ parameter vector $\boldsymbol{\lambda}$.

Many different specific selection functions have been proposed in the literature. Building on work on random effects models without predictors[31], Vevea and Hedges[32] described a random effects meta-regression model where selection probabilities vary depending on a set of pre-specified thresholds for the one-sided p-value. The thresholds are chosen based on conventional, psychologically salient cut-offs for judging statistical significance. Let $\alpha_1, ..., \alpha_H$ be a set of thresholds for the one-sided p-values, and set $\alpha_0 = 0$, $\alpha_{H+1} = 1$, and $\lambda_0 = 1$. A step function selection model is then given by

$$w(p^*; \boldsymbol{\lambda}) = \lambda_h \quad \text{if} \quad \alpha_h < p^* \leq \alpha_{h+1}, \tag{2}$$

(a) *One-step selection with* $\lambda_1 = 0.4$ (b) *Two-step selection with* $\lambda_1 = 0.4, \lambda_2 = 0.2$

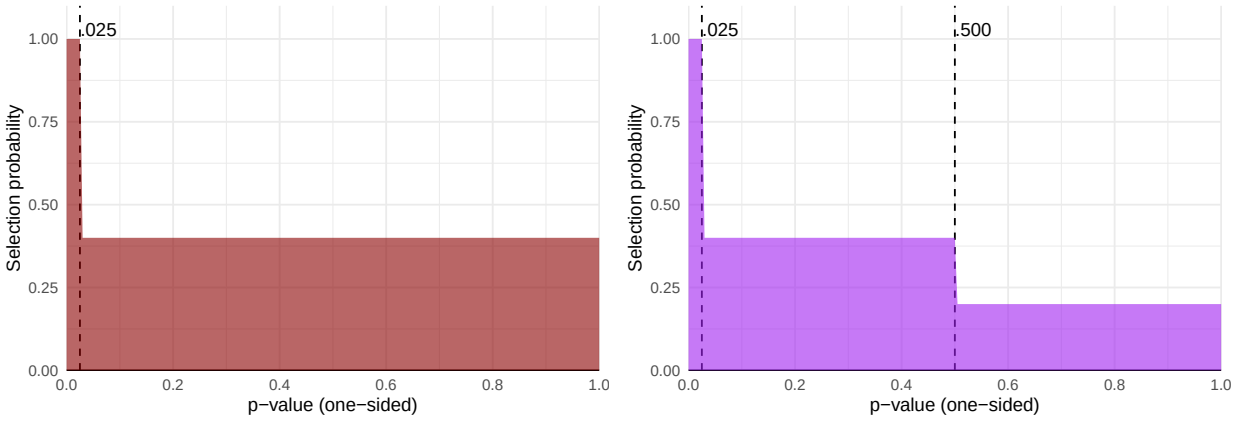

Figure 1
Examples of step functions

for $h = 0, ..., H$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_H)$. Equation (2) can be written equivalently as

$$w(Y^*, \sigma^*; \boldsymbol{\lambda}) = \lambda_h \quad \text{if} \quad \sigma^* \Phi^{-1}(1 - \alpha_{h+1}) \leq Y^* < \sigma^* \Phi^{-1}(1 - \alpha_h). \tag{3}$$

Setting $\lambda_0 = 1$ is necessary for identification because the absolute probabilities of selection cannot be estimated. The remaining parameters are therefore interpreted as *relative* probabilities of selection, compared to the probability of selection for an effect size with $p^* \leq \alpha_1$.

In practice, meta-analysts will often use only a small number of steps in the selection model. One common choice is the three-parameter selection model, which has a single step at $\alpha_1 = .025$, as depicted in Figure 1a. With this choice of threshold, positive effects that are statistically significant at the two-sided level of $p < .05$ have a different probability of selection than effects that are not statistically significant or not in the anticipated direction. Another possibility is to use two steps at $\alpha_1 = .025$ and $\alpha_2 = .500$, which allows for different probabilities of selection for effects that are positive but not statistically significant and effects that are negative (i.e., in the opposite the intended direction), as depicted in Figure 1b.

2.3 Distribution of observed effect size estimates

Combining the assumptions of the evidence-generating process and the selection process leads to a model for an observed effect size estimate. The distribution of an observed effect size estimate is equivalent to the distribution of Y^* given that $O = 1$. The marginal density of an observed effect size estimate Y with standard error σ has the form

$$f(Y = y|\sigma, \mathbf{x}) = \frac{1}{A(\mathbf{x}, \sigma; \boldsymbol{\beta}, \tau^2, \boldsymbol{\lambda})} \times w\left(y, \sigma; \boldsymbol{\lambda}\right) \times \frac{1}{\sqrt{\tau^2 + \sigma^2}} \phi\left(\frac{y - \mathbf{x}\boldsymbol{\beta}}{\sqrt{\tau^2 + \sigma^2}}\right), \tag{4}$$

where

$$A(\mathbf{x}, \sigma; \boldsymbol{\beta}, \tau^2, \boldsymbol{\lambda}) = \int_{\mathbb{R}} w\left(y, \sigma; \boldsymbol{\lambda}\right) \times \frac{1}{\sqrt{\tau^2 + \sigma^2}} \phi\left(\frac{y - \mathbf{x}\boldsymbol{\beta}}{\sqrt{\tau^2 + \sigma^2}}\right) dy. \tag{5}$$

If $w(y, \sigma; \boldsymbol{\lambda}) = 1$, then there is no selective reporting, $A(\mathbf{x}, \sigma; \boldsymbol{\beta}, \tau^2, \boldsymbol{\lambda}) = 1$, and the density reduces to the unweighted density of the evidence-generating process (i.e., the density of a random-effects meta-regression). The $A(\mathbf{x}, \sigma; \boldsymbol{\beta}, \tau^2, \boldsymbol{\lambda})$ term can be computed using the closed-form expression

$$A_{ij} = A(\mathbf{x}_{ij}, \sigma_{ij}; \boldsymbol{\beta}, \tau^2, \boldsymbol{\lambda}) = \sum_{h=0}^{H} \lambda_h B_{hij} \tag{6}$$

where

$$B_{hij} = \Phi\left(c_{hij}\right) - \Phi\left(c_{h+1,ij}\right), \tag{7}$$

and $c_{hij} = \left(\sigma_{ij}\Phi^{-1}\left(1 - \alpha_h\right) - \mathbf{x}_{ij}\boldsymbol{\beta}\right) / \sqrt{\tau^2 + \sigma_{ij}^2}$ for $h = 0, ..., H$.[32]

2.4 Estimation Methods

Past developments of selection models have focused either on maximum likelihood estimation under the assumption that all effect sizes are mutually independent[31–33] or on sensitivity analysis methods that treat the selection model as known[58,67]. We consider two estimation and inference strategies that build upon and generalize past approaches, including maximum composite marginal likelihood and an alternative based on re-weighting the Gaussian likelihood of the evidence-generating process. With both approaches, we allow for

incorporation of prior weights, which permits efficient calculation for a variety of bootstrapping techniques. Thus, let $a_{11}, ..., a_{Jk_j}$ be an arbitrary set of prior weights assigned to each effect size estimate; in a typical, unweighted analysis, all weights will be equal to $a_{ij} = 1$.

2.4.1 Maximum composite marginal likelihood

Composite marginal likelihood techniques involve working with the marginal distribution of each observed effect size estimate as if they were all mutually independent[68–70]. Thus, we assume that the observed effect size estimates were generated from Equation (4). For purposes of estimation, we write the likelihoods using natural log transformations of the variance parameter and selection parameters, with $\gamma = \log \tau^2$, $\zeta_h = \log \lambda_h$, and $\boldsymbol{\zeta} = [\zeta_1, ..., \zeta_H]'$. The log of the marginal likelihood contribution for effect size estimate i from study j is given by

$$\begin{aligned} l_{ij}^{M}\left(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right) &= \log f\left(Y = y_{ij} | \sigma_{ij}, \mathbf{x}_{ij}\right) \\ &\propto \log w\left(y_{ij}, \sigma_{ij}; \boldsymbol{\zeta}\right) - \frac{1}{2}\frac{\left(y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta}\right)^2}{\exp(\gamma) + \sigma_{ij}^2} \\ &\quad - \frac{1}{2}\log\left(\exp(\gamma) + \sigma_{ij}^2\right) - \log A\left(\mathbf{x}_{ij}, \sigma_{ij}; \boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right). \end{aligned} \tag{8}$$

The weighted composite marginal log-likelihood across all J studies is then

$$l^{M}\left(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right) = \sum_{j=1}^{J}\sum_{i=1}^{k_j} a_{ij} l_{ij}^{M}\left(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right). \tag{9}$$

The composite marginal likelihood (CML) estimators, denoted as $\hat{\boldsymbol{\beta}}$, $\hat{\gamma}$, and $\hat{\boldsymbol{\zeta}}$, are obtained as the set of parameter values that maximize the composite likelihood for the observed data, as given in Equation (9).

If the true parameter values are not at the extremes of their ranges, the CML

estimator can also be defined as the solution of the weighted score equations,

$$\sum_{j=1}^{J} \mathbf{S}_j \left(\hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\boldsymbol{\zeta}}\right) = \mathbf{0} \tag{10}$$

where $\mathbf{S}_j = \left(\mathbf{S}'_{\beta j} \; S_{\gamma j} \; \mathbf{S}'_{\zeta j}\right)'$ denotes the score vector from study j, consisting of the derivatives of the likelihood contributions for each study with respect to the component parameters:

$$\mathbf{S}_{\beta j}\left(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right) = \sum_{i=1}^{k_j} a_{ij} \frac{\partial l_{ij}^{M}\left(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right)}{\partial \boldsymbol{\beta}} \tag{11}$$

$$S_{\gamma j}\left(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right) = \sum_{i=1}^{k_j} a_{ij} \frac{\partial l_{ij}^{M}\left(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right)}{\partial \gamma} \tag{12}$$

$$\mathbf{S}_{\zeta j}\left(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right) = \sum_{i=1}^{k_j} a_{ij} \frac{\partial l_{ij}^{M}\left(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right)}{\partial \boldsymbol{\zeta}}. \tag{13}$$

Online Appendix A provides exact expressions for the score vectors of the step-function selection model.

Robust variance estimators, or sandwich estimators, are a commonly used technique for quantifying the uncertainty in CML estimators. Let $\mathbf{H}$ denote the Hessian matrix of the composite log-likelihood,

$$\mathbf{H}(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = \sum_{j=1}^{J} \frac{\partial \mathbf{S}_j(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta})}{\partial \left(\boldsymbol{\beta}' \; \gamma \; \boldsymbol{\zeta}'\right)}, \tag{14}$$

exact expressions for which are given in Online Appendix A. Let $\hat{\mathbf{S}}_j = \mathbf{S}_j(\hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\boldsymbol{\zeta}})$ and $\hat{\mathbf{H}} = \mathbf{H}(\hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\boldsymbol{\zeta}})$ denote the score vectors and Hessian matrix evaluated at the maximum of the composite likelihood. We then estimate the sampling variance of the CML estimator using a cluster-robust sandwich formula:

$$\mathbf{V}^{CML} = \hat{\mathbf{H}}^{-1} \left(\sum_{j=1}^{J} \hat{\mathbf{S}}_j \hat{\mathbf{S}}'_j\right) \hat{\mathbf{H}}^{-1}. \tag{15}$$

We construct confidence intervals for model parameters using $\mathbf{V}^{CML}$ with Wald-type large sample approximations. For instance, the $(1 - 2\alpha)$-level large-sample confidence interval for

a meta-regression parameter β_g is constructed as

$$\hat{\beta}_g \ \pm \ \Phi^{-1}(1-\alpha) \times \sqrt{V_{gg}^{CML}},$$

where V_{gg}^{CML} is the g^{th} diagonal entry of $\mathbf{V}^{CML}$.

2.4.2 Augmented, re-weighted Gaussian likelihood

Composite marginal likelihood is not the only possible basis for deriving estimators of selection model parameters. In the framework of a sensitivity analysis for worst-case selection bias, Mathur and VanderWeele[58] proposed using regular meta-analytic estimators for $\boldsymbol{\beta}$, but with weights defined by the inverse probability of selection under a step-function selection model with a single step at $\alpha_1 = .025$. Because they were working in the context of sensitivity analysis, they assumed a maximum plausible degree of selection rather than estimating the parameters of a selection model, so that the weights were fixed and known quantities. In contrast, here we will consider a more general model, possibly with multiple steps, using weights derived by estimating the selection model parameters. We describe the estimators as augmented, re-weighted Gaussian likelihood (ARGL) estimators because the Gaussian likelihood of the evidence-generating process is re-weighted based on the selection process, with selection process parameters identified by augmenting the likelihood with an additional estimating equation.

Given the parameters of the selection process, we can calculate relative probabilities of selection for each effect size estimate, $w_{ij} = w(y_{ij}, \sigma_{ij}; \boldsymbol{\zeta})$. We use these selection probabilities to form weighted estimating equations for the evidence-generating process. Under the evidence-generation model, the marginal log-likelihood of effect size estimate i from study j is Gaussian, given by

$$l_{ij}^{G}(\boldsymbol{\beta}, \gamma) \propto -\frac{1}{2}\frac{(y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta})^2}{\exp(\gamma) + \sigma_{ij}^2} - \frac{1}{2}\log(\exp(\gamma) + \sigma_{ij}^2).$$

Allowing for prior weights, the inverse selection-weighted Gaussian log-likelihood is therefore

$$l^G(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = \sum_{j=1}^{J} \sum_{i=1}^{k_j} \frac{a_{ij}}{w_{ij}} \times l_{ij}^G(\boldsymbol{\beta}, \gamma). \tag{16}$$

If the parameters of the selection process were known, we could find estimators for $\boldsymbol{\beta}$ and γ by maximizing $l^G(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta})$ for a fixed value of $\boldsymbol{\zeta}$. Equivalently, we could find the estimators as the solutions to the weighted score equations

$$\sum_{j=1}^{J} \mathbf{S}_{\beta j}^G(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = 0 \tag{17}$$

$$\sum_{j=1}^{J} \mathbf{S}_{\gamma j}^G(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = 0, \tag{18}$$

where the score contributions for study j are

$$\mathbf{S}_{\beta j}^G(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = \sum_{i=1}^{k_j} a_{ij} \times \mathbf{x}_{ij}' \frac{y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta}}{w_{ij}\left(\exp(\gamma) + \sigma_{ij}^2\right)} \tag{19}$$

$$\mathbf{S}_{\gamma j}^G(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = \sum_{i=1}^{k_j} a_{ij} \times \frac{\exp(\gamma)}{2w_{ij}} \left(\frac{(y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta})^2}{\left(\exp(\gamma) + \sigma_{ij}^2\right)^2} - \frac{1}{\exp(\gamma) + \sigma_{ij}^2} \right). \tag{20}$$

The question remains how to obtain an estimator for $\boldsymbol{\zeta}$. To estimate $\boldsymbol{\zeta}$, we augment the Gaussian log-likelihood with the marginal score equation with respect to $\boldsymbol{\zeta}$. Specifically, we define the ARGL estimators as the values that simultaneously solve Equations (17) and (18) together with the estimating equation for $\boldsymbol{\zeta}$ from the composite marginal likelihood approach. With $\mathbf{S}_{\zeta j}(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta})$ as given in Equation (13), the full set of estimating equations is

$$\mathbf{M}_j(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = \begin{bmatrix} \mathbf{S}_{\beta j}^G(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) \\ S_{\gamma j}^G(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) \\ \mathbf{S}_{\zeta j}(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) \end{bmatrix}, \tag{21}$$

based on which we define the ARGL estimator as the solution to the estimating equations

$$\sum_{j=1}^{J} \mathbf{M}_j(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = \mathbf{0}. \tag{22}$$

We will denote the ARGL parameter estimators as $\tilde{\boldsymbol{\beta}}$, $\tilde{\gamma}$, and $\tilde{\boldsymbol{\zeta}}$.

We consider conducting inferences for the ARGL estimators using cluster-robust sandwich variance estimators that have the same form as (15). Online Appendix B provides further details. We construct confidence intervals for model parameters using Wald-type large sample approximations, just as with the CML estimators.

2.5 Bootstrap inference

Sandwich estimators such as those in Equations (15) require a large number of independent clusters (i.e., large J) to provide accurate assessments of uncertainty. We consider alternative inference techniques based on bootstrapping, which might provide more accurate inference with a limited number of clusters. Bootstrap techniques involve generating many new pseudo-samples of observations by randomly perturbing the original sample, then re-calculating an estimator using each pseudo-sample. The distribution of the estimator across pseudo-samples is used as a proxy for the actual sampling distribution of the estimator, providing a basis for calculating standard errors and confidence intervals.[71]

Many different bootstrap sampling schemes have been described that apply to different data structures and require different assumptions[72]. For data involving dependent observations, it is crucial that the process used to generate pseudo-samples accounts for the dependence structure. Techniques that do so include the non-parametric clustered bootstrap, two-stage bootstrap[73,74], and fractional random weight bootstrap[75]. Here, we focus on the two-stage bootstrap; details about other bootstrap techniques can be found in Online Appendix C.1.

In the two-stage bootstrap, each pseudo-sample is generated by randomly drawing J

clusters of observations with replacement from the original sample and then randomly re-sampling observations with replacement from each selected cluster. This process amounts to simulating set of weights. Let $a_j^{(b)}$ be a first-stage weight for cluster j and $a_{ij}^{(b)}$ be the weight assigned to observation i in cluster j for pseudo-sample b. The two-stage bootstrap is equivalent to first drawing $a_1^{(b)}, ..., a_J^{(b)}$ from a multinomial distribution with J trials and equal probability on each of J categories, then drawing $a_{1j}^{(b)}, ..., a_{k_j j}^{(b)}$ from a multinomial distribution with $a_j^{(b)} \times k_j$ trials and equal probability on each of k_j categories, for $j = 1, ..., J$.

For constructing confidence intervals, bootstrapping entails generating a total of B pseudo-samples, where B is a large number such as 1999, and re-calculating the estimator for each pseudo-sample. There are several methods for constructing confidence intervals from a bootstrap distribution. We consider four standard methods[72], including the percentile CI, basic CI, studentized CI, and the bias-corrected-and-accelerated CI proposed by Efron[76]. Online Appendix C.2 provides further details about the bootstrap CI calculations.

3 Empirical Example

To demonstrate the proposed modeling strategy and examine potential differences between CML and ARGL estimation methods, we re-analyzed the data from a meta-analysis reported by[77]. The original meta-analysis examined a large corpus of primary studies on the ego depletion effect, which refers to the theory that an individual's ability to exercise self-control diminishes with repeated exertion[78].[77] argued that the apparent strength of ego-depletion effects may be overstated due to selective reporting. Their review included a variety of self-control manipulation tasks as well as a range of outcomes. Effect sizes were measured as standardized mean differences, defined so that positive effects correspond to depletion of self-control (i.e., consistent with the theory of ego depletion).

Based on a sample of 66 studies,[77] estimated an overall average effect of 0.43, with a 95% CI of [0.34, 0.52]. However, more recent, preregistered multilab replication studies present a stark contrast.[79] estimated a much smaller, non-significant average effect of 0.04,

95% CI [-0.07, 0.15], from 23 laboratory studies. Similarly,[80] found a non-significant effect of 0.06, 95% CI [-0.02, 0.14], using 36 laboratory studies. This discrepancy highlights the importance of using modeling strategies that account for selective reporting. Unlike the earlier studies, the replication studies were preregistered and are thus not susceptible to the same selective reporting issues.

To mitigate possible effects of selective reporting,[77] included many unpublished studies, so that the full meta-analysis included 116 effects from 66 studies.

For illustrative purposes, we re-analyzed the findings from the subset of published studies only; we also excluded a single outlying effect size estimate that was greater than 2. This analytic sample includes 66 effect size estimates from 45 distinct studies. About 40% of the studies contributed multiple effects per study, due to either multiple samples or multiple outcomes, leading to dependent effect sizes. We conducted the analyses using R Version 4.4.3[81].

If selective reporting were not a concern, a correlated-and-hierarchical effects (CHE) model, a flexible model that accounts for both correlated and hierarchical dependent data structures, would be one way to summarize the distribution of ego depletion effects.[54] Based on a CHE model, the overall average effect estimate was 0.46, 95% CI [0.34, 0.59]. Alternately, one could apply a CHE model with inverse sampling covariance weighting (CHE-ISCW), which places relatively more weight on larger studies (those with smaller sampling variances) and thus is less biased by selective reporting.[62] Applying CHE-ISCW reduces the overall effect estimate to 0.39, 95% CI [0.24, 0.54]. As a further point of comparison, we estimated the overall average effect using the PET/PEESE regression adjustment[26], clustering the standard errors by study. This yielded an overall average effect of -0.09, 95% CI [-0.76, 0.58]. The CHE and CHE-ISCW estimates are both positive, significant, and similar in magnitude. The PET-PEESE estimate is negative and much smaller than the CHE and CHE-ISCW estimates, indicating a pattern of small study effects.

We used the `selection_model()` function from the `metaselection` package to fit single-step and two-step selection models[82]. The single-step model used a threshold at $\alpha_1 = 0.025$; the two-step model used thresholds at $\alpha_1 = 0.025$ and $\alpha_2 = 0.5$. For comparison purposes, we estimated model parameters using both CML and ARGL and computed cluster-robust and percentile bootstrap confidence intervals. For bootstrapping, we used two-stage cluster bootstrap re-sampling with 1999 replicates.

Table 1 presents the parameter estimates from the one-step and two-step selection models. The estimated selection parameters are similar across the one- and two-step models and across both estimators, all indicating that non-significant or negative effect size estimates were less likely to be reported than statistically significant, affirmative ones. The one-step and two-step selection model estimates of average effect size are positive but substantially smaller than the CHE-ISCW estimates, ranging from 0.24 to 0.27 depending on the model specification and estimation method. In contrast to the PET/PEESE estimate, the selection model estimates using CML are positive and statistically distinct from zero. Thus, an analyst would reach different conclusions about overall average effect size depending on whether they use an unadjusted model, a step-function model, or the PET/PEESE adjustment.

The estimates in Table 1 point towards some potential differences between estimation methods. Generally, the CML and ARGL parameter estimates are similar in magnitude, but the confidence intervals based on the CML estimator are narrower than those for the ARGL estimator. For the CML estimator, the bootstrap CIs are similar or slightly wider than the cluster-robust CIs. These patterns suggest that there could be differences in the performance of the estimators, as well as differences in performance between the step-function estimators and alternative adjustment methods such as PET/PEESE. However, these results are based on a single empirical dataset where the true data-generating process is unknown. To draw firmer conclusion about these methods, we conducted simulations to evaluate their performance characteristics across a range of conditions.

Table 1

Single-step and two-step selection model parameter estimates fit to ego depletion effects data from Carter et al. (2015)

	CML estimator			ARGL estimator		
Parameter	Estimate (SE)	Cluster-Robust CI	Percentile Bootstrap CI	Estimate (SE)	Cluster-Robust CI	Percentile Bootstrap CI
One-step						
β	0.25 (0.09)	[0.08, 0.43]	[0.06, 0.47]	0.27 (0.07)	[0.14, 0.40]	[0.11, 0.45]
τ^2	0.10 (0.03)	[0.06, 0.20]	[0.03, 0.18]	0.11 (0.04)	[0.06, 0.21]	[0.02, 0.21]
λ_1	0.27 (0.13)	[0.11, 0.71]	[0.09, 0.82]	0.30 (0.09)	[0.16, 0.55]	[0.08, 1.11]
Two-step						
β	0.24 (0.09)	[0.06, 0.42]	[0.05, 0.48]	0.26 (0.43)	[-0.59, 1.10]	[0.09, 0.50]
τ^2	0.11 (0.03)	[0.06, 0.20]	[0.03, 0.18]	0.11 (0.10)	[0.02, 0.66]	[0.02, 0.20]
λ_1	0.27 (0.12)	[0.11, 0.67]	[0.11, 0.81]	0.29 (0.53)	[0.01,10.10]	[0.10, 1.00]
λ_2	0.24 (0.14)	[0.08, 0.73]	[0.09, 1.11]	0.27 (0.67)	[0.00,34.98]	[0.06, 2.12]

Note: ARGL = augmented, reweighted gaussian likelihood; CML = composite maximum likelihood; CI = confidence interval; SE = standard error.

4 Simulation Methods

n— title: “Simulation Methods” output: pdf_document bibliography: references.bib editor_options: markdown: wrap: sentence —

We conducted Monte Carlo simulation studies to examine the performance of the CML and ARGL estimators for a step-function selection model with a single step, under a wide range of conditions where primary studies contribute multiple, statistically dependent effect size estimates. We compare the performance of these novel estimators to two available alternatives: a summary meta-analysis that addresses the dependency structure but does not correct for selective reporting and the PET-PEESE adjustment[26]. We evaluated the estimators in terms of convergence rates, bias, accuracy, and confidence interval coverage for recovering the average effect size prior to selective reporting. To limit the computational burden, we evaluated the performance of bootstrap confidence intervals only for a subset of the conditions.

We ran the simulation in R Version 4.4.1[81] using the high-throughput computing

cluster at the University of Wisconsin - Madison[83]. The simulation code made use of several R packages, including `metafor`[84], `clubSandwich`[85], `simhelpers`[86], `optimx`[87], `nleqslv`[88], and `tidyverse`[89].

4.1 Data generation

We simulated meta-analyses based on a CHE working model, with individual effect size estimates selected for inclusion based on the step-function selection model with a single step at $\alpha_1 = .025$. For each replication, we generated a total of J^* studies with a two-group comparison design, where study j had an effective sample size of N_j and contributed $k_j^* \geq 1$ effect size estimates prior to selective reporting. To generate each meta-analytic dataset, we sampled effective sample sizes* and numbers of effect sizes per study from an empirical distribution based on the What Works Clearinghouse database. The total effective sample size per study was divided equally into two groups, treatment and control. We then generated r_j, an outcome correlation for study j, by drawing from a beta distribution with mean ρ and standard deviation 0.05. This approach assumes a plausible dependency structure in which all within-study correlations are positive, as multiple outcomes often measure the same underlying construct or the same construct at different time points. For simplicity in the data-generating process, we assumed a constant correlation between pairs of outcomes within a study but allowed the correlation to vary from study to study.

We then simulated raw outcomes for each primary study included in the meta-analytic dataset. To do so, we first generated an average effect size per study, δ_j, from a normal distribution with mean μ and variance of τ_B^2. Given δ_j, we generated k_j^* effect size parameters per study, $\boldsymbol{\delta}_j = \left(\delta_{1j}, ..., \delta_{k_j^* j}\right)'$, from a normal distribution with mean of δ_j and variance of ω^2. For each of the $N_j/2$ participants in the treatment and control groups, we

* For studies that involved cluster-level treatment assignment, we computed effective sample sizes that account for the dependence of observations nested within clusters rather than using the raw participant-level sample sizes.

then simulated vectors of multivariate normal outcomes as

$$\mathbf{Y}_{hj}^{T} \sim N(\boldsymbol{\delta}_j, \boldsymbol{\Psi}_j) \qquad \text{and} \qquad \mathbf{Y}_{hj}^{C} \sim N(\mathbf{0}, \boldsymbol{\Psi}_j).$$

Here, $\mathbf{Y}_{hj}^{T}$ and $\mathbf{Y}_{hj}^{C}$ are $k_j^* \times 1$ vectors of outcomes for participant h in study j in treatment and control group and $\boldsymbol{\Psi}_j$ is the covariance matrix for outcomes in study j, with the diagonal elements equal to 1 and off-diagonal elements equal to r_j. We assumed a compound symmetry structure for the true effect sizes and a sampling variance of 1 for the raw outcomes. These assumptions represent a plausible data structure for meta-analyses with dependent effects, while also simplifying the data-generating process for our simulation. The assumption of a unit sampling variance is a standard practice when simulating data for standardized effect sizes like Hedges's g, as it does not affect the final effect size estimates. From the raw outcome data, we calculated standardized mean differences using Hedges's g bias correction for each of the correlated outcomes, yielding effect size estimates y_{ij}^* for $i = 1, ..., k_j^*$ and $j = 1, ..., J^*$. We calculated the sampling variances using conventional formulas[90] and computed one-sided p-values based on two-sample t-tests for the null of $H_0 : \delta_{ij} \leq 0$.

After simulating results for study j, we applied a step-function selection model to the individual effect size estimates, with a one-sided selection threshold at $\alpha = 0.025$. Specifically, we let result $(y_{ij}^*, \sigma_{ij}^*, p_{ij}^*)$ be included in the observed dataset with probability 1 if p_{ij}^* was less than 0.025 and with probability $0 < \lambda_1 \leq 1$ if $p_{ij}^* \geq 0.025$. We repeated the process of generating studies until the database included a total of J studies with at least one observed result, where observed study j includes k_j effect size estimates for $1 \leq k_j \leq k_j^*$.

4.2 Estimation methods

We estimated step-function selection models with a single step at $\alpha_1 = 0.025$, so that the assumed marginal selection process is consistent with the actual selective reporting process used to generate meta-analytic datasets. We estimated CML and ARGL estimators

and calculated cluster-robust confidence intervals as described in Section 2.4. For a subset of simulation conditions, we also examined percentile, basic, studentized, and bias-corrected-and-accelerated bootstrap confidence intervals based on the non-parametric two-stage bootstrap, clustered bootstrap, and fractional random weight bootstrap as described in Online Appendix C. To maintain computational feasibility, we used $B = 399$ bootstrap replications of each estimator but extrapolated coverage rates for $B = 1999$ replications[91]; Online Appendix C.3 provides further details.

We compared the performance of the step-function CML and ARGL estimators to that of two alternative methods. First, we estimated a summary meta-analysis model without any correction for selective reporting, using a method that accounts for effect size dependency. Specifically, we used the CHE working model with inverse sampling-covariance weights (CHE-ISCW)[62]. We estimated the variance components using restricted maximum likelihood and assumed a sampling correlation of 0.8 for all pairs of effect size estimates from the same study, which leads to a degree of mis-specification when the average correlation used in the data-generating process differs from 0.80. This assumption aligns with the default correlation structure used by `robumeta`[92], the widely used R package for implementing RVE in meta-analysis. Second, we implemented a variation of the PET/PEESE estimator originally proposed by[26], adapted to accommodate dependent effect sizes.* Following[26], we combined the estimators by using PEESE if the PET estimator is statistically distinct from zero at an α-level of 0.10, and otherwise using PET. For the PET/PEESE, and CHE-ISCW

* The PET estimator is based on the working model

$$T_{ij} = \mu + \beta \times \frac{2}{\sqrt{N_j}} + e_{ij} \tag{23}$$

The PEESE estimator is similar, but uses the sampling variance instead of the sampling standard error:

$$T_{ij} = \mu + \beta \times \frac{4}{N_j} + e_{ij} \tag{24}$$

For both estimating equations, sampling errors are assumed to have fixed variances $\text{Var}(e_{ij}) = \sigma_{ij}^2$ and constant sampling correlation within studies, $\text{cor}(e_{hj}, e_{ij}) = \rho$.

estimators, we calculated confidence intervals using cluster-robust variance estimation with the CR2 small-sample correction and Satterthwaite degrees of freedom[62].

4.3 Experimental design

Table 2 summarizes the experimental design for this study. Manipulated parameters included overall average standardized mean difference (μ), between-study heterogeneity (τ_B), within-study heterogeneity ratio (ω^2/τ_B^2), average correlation between outcomes (ρ), probability of selection for non-affirmative results (λ_1), number of observed studies (J), and primary study sample size. We examined values for the overall average SMD (μ) ranging from 0.0 to 0.80, which covers the range of effects observed in a review of 747 randomized control trials of education interventions[93]. We used values of τ ranging from 0.05 (a very small degree of heterogeneity) to 0.45 (a large degree of heterogeneity). We specified the degree of within-study heterogeneity in relative terms, by setting the ratio of ω^2 to between-study heterogeneity τ^2 at either 0 (i.e., no within-study heterogeneity) or 0.5. For the average correlation between outcomes from the same study, we examined the values of 0.40 or 0.80. The default value of the average correlation in software packages that implement RVE is 0.80. Thus, in conditions where ρ is 0.80, the working model is approximately correctly specified. We included conditions where $\rho = 0.4$ to examine performance when the working model is not correctly specified.

We examined a wide range of values for the probability of selection for non-affirmative effect sizes, ranging from no selective reporting ($\lambda_1 = 1$) to very severe selective reporting ($\lambda_1 = 0.02$). We also examined a wide range of conditions for the number of primary studies included in the meta-analysis, ranging from relatively small databases of $J = 15$ to very large databases with $J = 120$ studies. We chose these values to cover the conditions found in real meta-analyses of education and psychology research[45].

Lastly, we investigated the primary study sample size. For the typical primary study sample sizes, we used the empirical distribution of sample sizes in the What Works

Table 2
Parameter values examined in the simulation study

Parameter	Full Simulation	Bootstrap Simulation
Overall average SMD (μ)	0.0, 0.2, 0.4, 0.8	0.0, 0.2, 0.4, 0.8
Between-study heterogeneity (τ_B)	0.05, 0.15, 0.30, 0.45	0.15, 0.30, 0.45
Heterogeneity ratio (ω^2/τ_B^2)	0.0, 0.5	0.0, 0.5
Average correlation between outcomes (ρ)	0.40, 0.80	0.80
Probability of selection for non-affirmative effects (λ_1)	0.02, 0.05, 0.10, 0.20, 0.50, 1.0	0.05, 0.20, 1.0
Number of observed studies (J)	15, 30, 60, 90, 120	15, 30, 60
Primary study sample size	Typical, Small	Typical, Small

Clearinghouse database of findings from educational intervention studies. The sample sizes in the database ranged from 37 to 2,295 with a median of 211. The number of effect sizes ranged from 1 to 48 with a median of 3. To explore the influence of the effective sample size distribution, we also ran conditions in which we divided the sample sizes from the What Works Clearinghouse database by three to represent primary studies with smaller sample sizes, such as those used in psychology laboratory studies.

Parameters were fully crossed for a total of $4 \times 4 \times 2 \times 2 \times 6 \times 5 \times 2 = 3,840$ conditions in the full simulation study. Due to the computational demands of bootstrapping, we focused the bootstrap simulations on conditions with fewer studies per meta-analysis, for which we expected large-sample cluster-robust CIs to be relatively less effective. We also reduced the number of parameter values for factors where we did not observe much variation in results in the full simulation (e.g., excluding $\tau = 0.30$). This resulted in $4 \times 2 \times 2 \times 1 \times 3 \times 3 \times 2 = 288$ conditions for the bootstrap simulations. For each condition, we generated 2,000 replications.

4.4 Performance criteria

We evaluated the performance of these methods in terms of convergence rates, bias, scaled root mean-squared error (RMSE), and 95% confidence interval coverage for the overall

average effect size μ. Because we expected that RMSE would decrease proportionally with the square-root of the number of studies, we scaled the RMSE of each estimator by $\sqrt{J}$ to reduce variation across the number of studies included in each meta-analysis. Because the sampling distribution of the CML and ARGL estimators sometimes included extreme outlying values, we calculated bias and scaled RMSE after winsorizing the distribution. Specifically, we defined a lower fence of 2.5 times the inter-quartile range below the 25th percentile and an upper fence of 2.5 times the inter-quartile range above the 75th percentile. Estimates falling below the lower fence or above the upper fence were set to the corresponding fence values.

For confidence intervals based on cluster-robust variance estimation, we calculated coverage rates as the proportion of simulated intervals that included the true parameter. For bootstrap confidence intervals, estimation of coverage rates is complicated by the fact that coverage is affected by the number of bootstrap replications. To mitigate computational demand, we used few bootstraps per replication than recommended for analysis of real data. To estimate coverage rates for confidence intervals as would be used in practice, we used an extrapolation technique.[91] For each replication, we computed bootstrap confidence intervals not only for $B = 399$, but also for $B = 49$, 99, 199, and 299 bootstraps, randomly selected without replacement. We computed coverage rates separately for each value of B, fit a linear regression of the coverage rate on $1/B$, and then used this regression to predict the coverage rate of confidence intervals based on $B = 1999$ bootstraps. Online Appendix C.3 provides further details and demonstrates the performance of the extrapolation technique.

5 Simulation Results

We organize our presentation of simulation results by first considering the properties of point estimators for the average effect size. For this parameter, we compare the bias and accuracy of the CML and ARGL estimators to that of the CHE-ISCW estimator and the PET/PEESE estimator. We then examine the calibration of cluster-robust and bootstrap

confidence intervals based on the CML and ARGL estimators. Online Appendix E includes additional results regarding the bias and accuracy of estimators of the marginal variance of the effect size distribution; Online Appendix F has additional results on the performance of the CML and ARGL estimators for the selection parameter.

5.1 Convergence

The CHE-ISCW and PET/PEESE estimators produced results for every replication in every condition. The CML and ARGL estimators for the step-function selection model had very high convergence rates across most conditions, although the CML estimator did exhibit rates of convergence below 99% under conditions with the lowest degree of heterogeneity $\tau = 0.05$, with the lowest convergence rate of 93.70%. For the ARGL estimator, convergence was 100% across all conditions. Supplementary Figure D1 depicts the range of convergence rates of the CML estimator. We evaluated the performance characteristics of each estimator across the replications where it converged.

5.2 Bias

Figure 2 depicts the bias (represented on the vertical axis of each plot) of each estimator of average effect size as a function of the strength of selective reporting (horizontal axis), average effect size parameter (varying by grid column), and between-study heterogeneity (τ, varying by grid row). The box plot for each estimator depicts variation in bias over the remaining factors in the simulation design, which include the heterogeneity ratio, correlation between effect size estimates, number of observed studies, and primary study sample size distribution. Note that the range of the vertical axis differs by grid row because the bias of some estimators is strongly influenced by the degree of heterogeneity.

Bias is generally lower for the CML and ARGL estimators than the comparison methods across all conditions, particularly when the average effect size is small to moderate ($\mu \leq 0.4$). While still substantially lower than the comparison methods, the maximum absolute bias reaches 0.34 for CML and 0.34 for ARGL. This maximum bias occurs for both

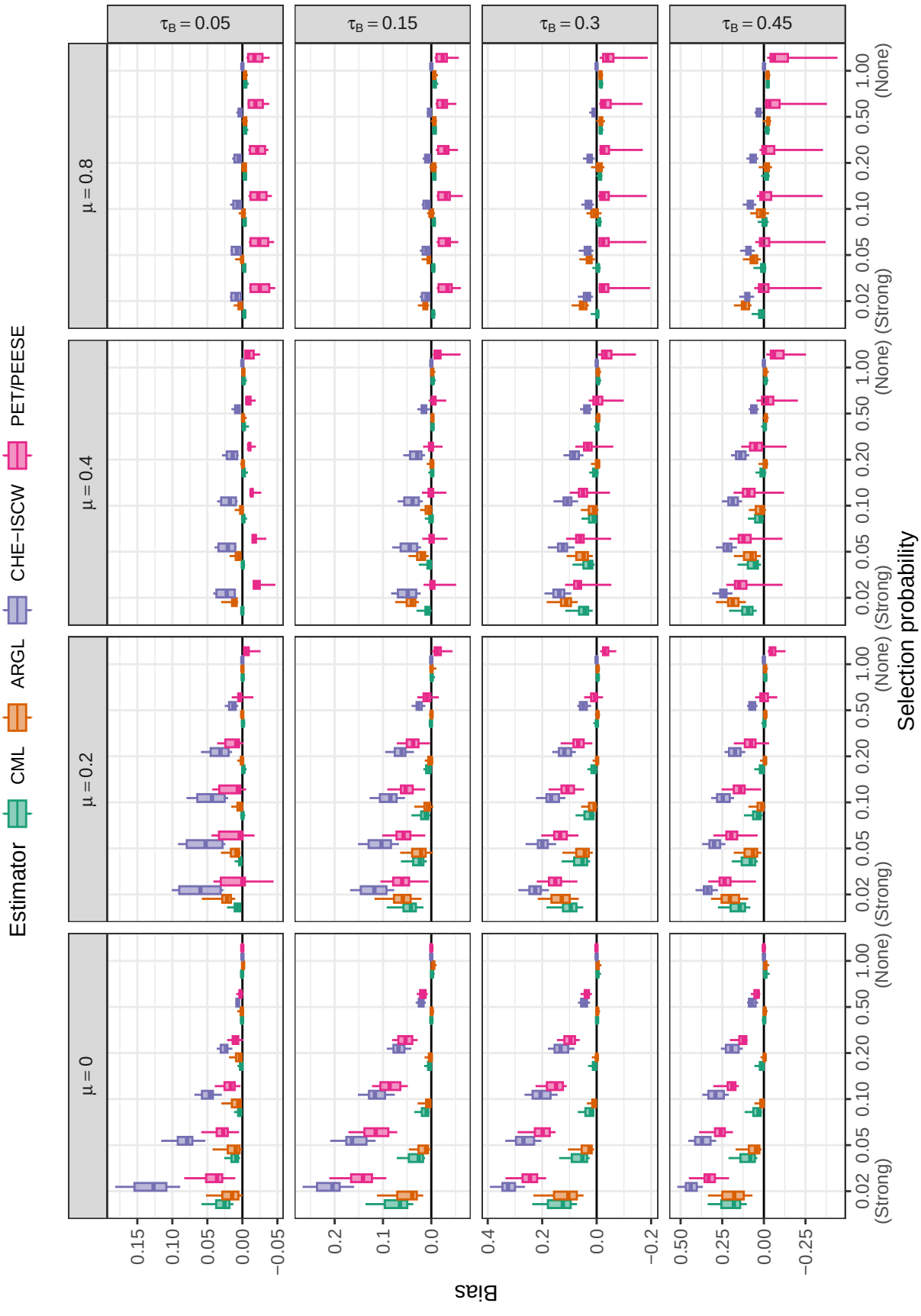

Figure 2

Bias for estimators of average effect size by selection probability, average SMD, and between-study heterogeneity

estimators when selective reporting is very strong, average effect size is zero, and heterogeneity is large. We observe a performance trade-off between the two estimators of the step-function selection model: Bias is lower for the ARGL estimator than the CML estimator when $\mu = 0.0$ but higher when ($\mu \geq 0.2$). For both estimators, bias decreases as the average effect increases, as heterogeneity decreases, and as the probability of selection decreases.

In contrast to the estimators based on the marginal selection model, the CHE-ISCW and PET/PEESE estimators are systematically biased under many conditions. The CHE-ISCW estimator, which does not directly adjust for selective reporting, is systematically biased under conditions with non-null selection. When average effect size is large ($\mu = 0.8$), its bias remains quite small even when selective reporting is very strong. This result is not surprising because the large average effect size and low heterogeneity mean that the vast majority of simulated results are statistically significant, leaving little opportunity for the selection mechanism to operate. However, the bias of CHE-ISCW grows stronger when selection is more extreme, when average effect size is smaller, and when heterogeneity is larger; its bias exceeds 0.50 when $\mu = 0.0$, $\tau = 0.45$, and $\lambda_1 = 0.02$. Although the PET/PEESE estimator uses a regression adjustment to account for possible selective reporting, it too becomes severely biased when selective reporting is strong. For smaller values of average effect size ($\mu \leq 0.2$), the bias of PET/PEESE tracks the bias of the CHE-ISCW estimator but is somewhat less pronounced. Its bias grows larger (and closer to that of CHE-ISCW) for smaller values of average effect size and higher levels of heterogeneity. For larger values of average effect size ($\mu = 0.8$), the PET/PEESE estimator is negatively biased, systematically under-estimating the average effect size—especially at high levels of heterogeneity. The selection model directly corresponds to the step-function data-generating mechanism used in the simulation, which could contribute to its estimators' strong performance relative to the other methods.

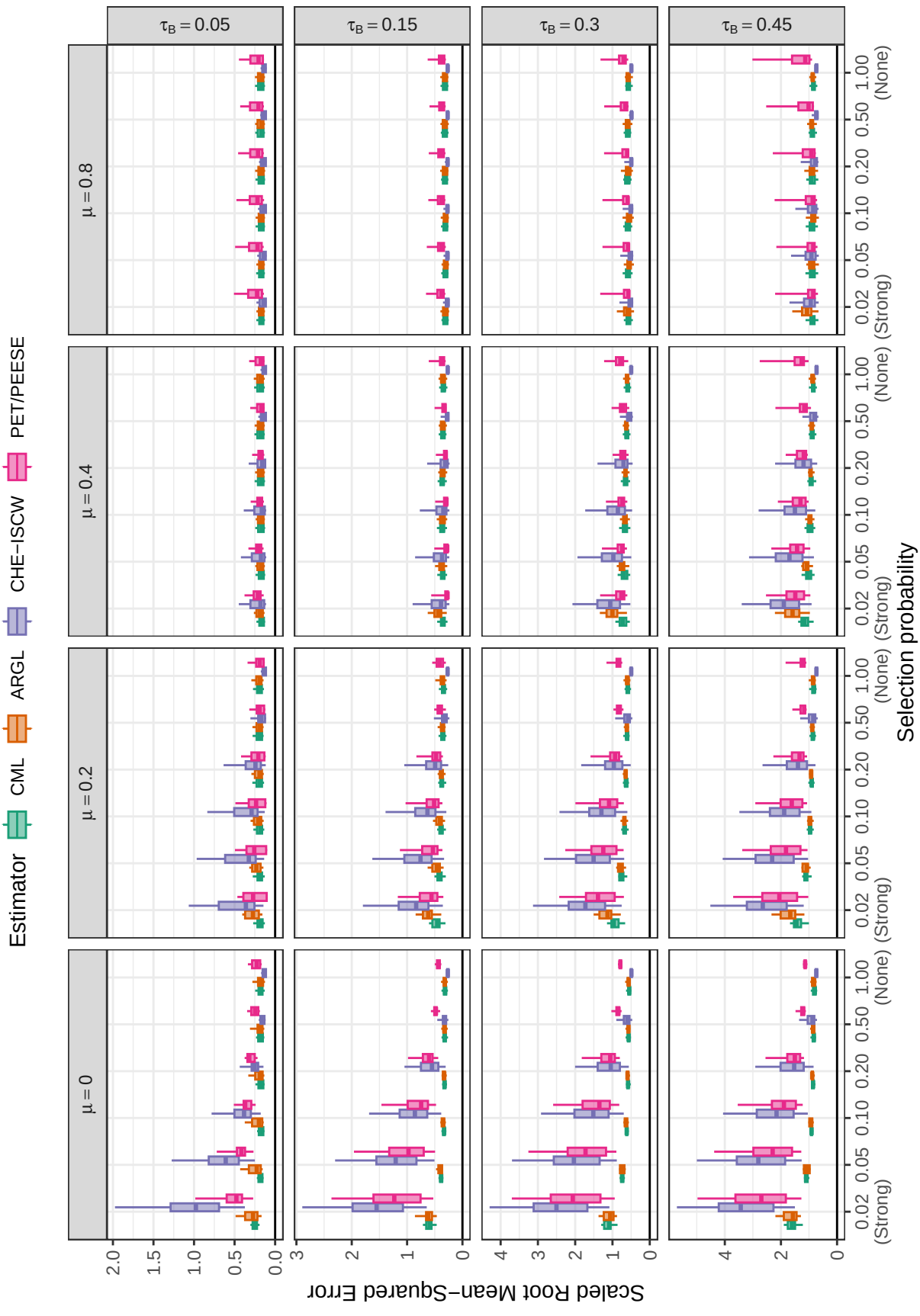

Figure 3

Scaled root mean-squared error for estimators of average effect size by selection probability, average SMD, and between-study heterogeneity

5.3 Scaled RMSE

Scaled RMSE combines both bias and variability into an overall measure of inaccuracy. Figure 3 depicts the scaled RMSE of each estimator of average effect size; it is constructed in the same way as Figure 2. Figures D2 through D6 in Online Appendix D provide greater detail about the relative accuracy of the four methods by plotting the ratio of scaled RMSEs for each pair of methods. In addition, Figure D7 depicts the non-scaled version of the RMSE of each estimator of average effect size. These figures illustrate several findings.

First, across most data-generating conditions, the ARGL estimator has higher RMSE than the CML estimator. As evident in Figure D2, the RMSE ratio comparing ARGL to CML is greater than one across most conditions examined. The ARGL estimator has lower RMSE only under conditions of very high heterogeneity and in databases with few studies. Thus, the CML estimator will typically be preferable to the ARGL estimator.

Second, considering both the selection model estimators and comparison methods, no single method achieves the lowest RMSE uniformly across all conditions examined. Instead, all methods face bias-variance trade-offs. Under conditions with small or moderate average effect size and moderate or strong selection, the selection model estimators generally have lower RMSE than the CHE-ISCW and PET/PEESE estimators. The CML estimator has lower RMSE than CHE-ISCW under most conditions where selective reporting creates meaningful bias—specifically, for $\lambda_1 <= 0.2$ and $\mu \leq 0.2$ (Figure D3). The relative accuracy of the ARGL estimator versus CHE-ISCW follows a similar pattern (Figure D4).

Third, the CML estimator also has lower RMSE than PET/PEESE under conditions where selective reporting creates meaningful bias, although it is not uniformly more accurate than PET/PEESE (Figure D5). Rather, PET/PEESE is more accurate under *some* conditions involving moderate or large effect size ($\mu \geq 0.4$) and varying degrees of between-study heterogeneity, which correspond to conditions where the bias of PET/PEESE

is small. The relative accuracy is difficult to characterize generally because it follows a non-linear pattern involving interactions among the data-generating parameters. The pattern of relative accuracy is very similar for the ARGL estimator (Figure D6).

5.4 Confidence Interval Coverage

Figure 4 shows the coverage rates of 95% CIs based on large-sample cluster-robust variance estimators for the CHE-ISCW, PET/PEESE, CML, and ARGL estimators.* Coverage rates are below the nominal rate of 0.95 for all methods across most conditions. The value of the average correlation between outcomes (ρ) has virtually no effect on the coverage results across all methods (Supplementary Figure**ref?**(fig:comparison-coverage-rho)). The CML and ARGL estimators based on the step-function selection model have higher coverage rates than the comparison methods under many conditions, particularly in conditions with higher between-study heterogeneity. Between the two selection model estimators, coverage rates are generally higher for the CML estimator than the ARGL estimator, though neither consistently achieves the nominal rate. Intervals based on the CML and ARGL estimators have coverage levels that improve towards 0.95 as the number of studies increases, but are often still unacceptably low even when J is 90 or greater. In contrast, intervals based on CHE-ISCW and PET/PEESE are often wildly mis-calibrated. Under conditions where CHE-ISCW and PET/PEESE are biased by selective reporting, their confidence intervals do not center on the true parameter. Consequently, as the number of studies increases, the standard error of the estimators decreases (as does the width of confidence intervals) and their coverage rates degrade towards zero.

Bootstrap intervals for the step-function model provide more accurate coverage levels. Due to the computational demands of bootstrapping, we evaluated the bootstrap confidence intervals under a more limited range of data-generating conditions, including a maximum

* To provide greater detail, the vertical axis of Figure 4 is limited to the range [0.5, 1.0], and coverage rates of the CHE-ISCW and PET/PEESE intervals are not depicted when they fall below 0.5. Supplementary Figure D8 depicts the full range of coverage rates.

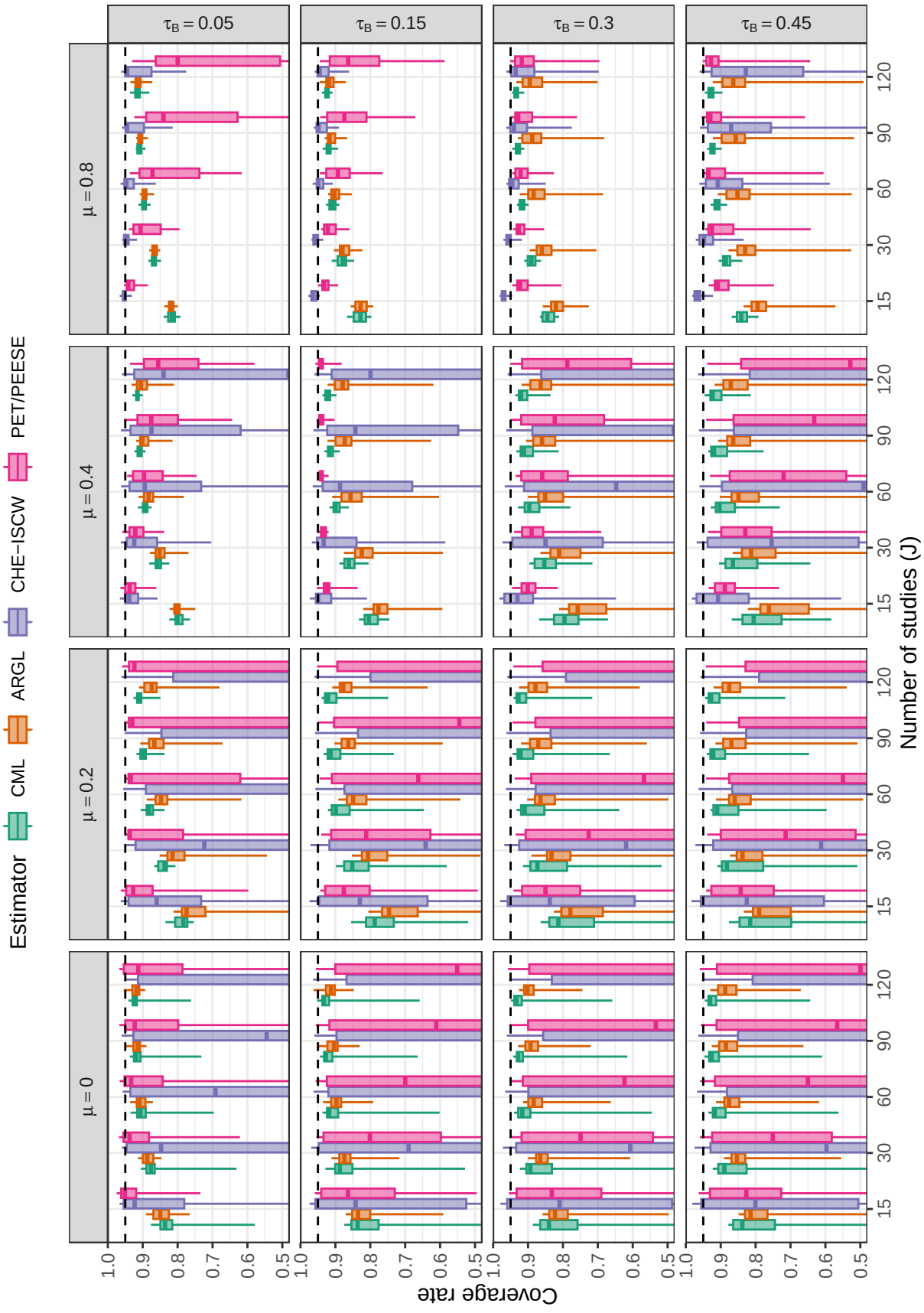

Figure 4

Coverage levels of confidence intervals based for average effect size based on cluster-robust variance approximations, by number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95. Coverage rates of the CHE-ISCW and PET/PEESE intervals are not depicted when they fall below 0.5

sample size of $J = 60$. Figure 5 depicts the coverage levels of confidence intervals based on the CML estimator, including intervals based on large-sample cluster-robust variance methods and percentile intervals using either two-stage, multinomial, or exponential (fractional reweighted) bootstrap resampling.* Although none of the intervals provide exactly nominal coverage, all versions of the percentile bootstrap intervals have coverage that is closer to nominal than the intervals based on cluster-robust variance estimation. The percentile intervals with two-stage clustered bootstrap re-sampling provided the best coverage levels, achieving or exceeding 90% coverage when selection is weak or non-existent or when the average effect size is large ($\mu = 0.8$), even with only $J = 15$ primary studies per meta-analysis. However, coverage variability increases substantially when selection is strong and the average effect size is small to moderate ($\mu \leq 0.04$). Coverage variability also increases as between-study heterogeneity increases, particularly when $\tau \geq 0.30$ and $\mu \leq 0.04$ (Supplementary Figure**ref?**(fig:CML-coverage-tau)). Coverage levels of intervals based on the ARGL estimator followed very similar patterns to those for the CML estimator (Supplementary Figures**ref?**(fig:ARGL-coverage-two-stage)**ref?**(fig:ARGL-coverage-exponential)).

6 Discussion

We have described and evaluated several methods for estimating step-function selection models while accounting for dependent effect sizes, a common feature of meta-analyses in social science fields. We focused on the step-function selection model because it offers a number of advantages over other available methods for diagnosing and correcting selective reporting bias. First, step-function models are generative, in that they include parameters describing the selective reporting process under simple yet plausible forms of selective reporting connected to statistical significance. In contrast, regression-based estimators such as PET/PEESE[26] or the endogenous kink meta-regression[27] are agnostic as

* Coverage levels of the other bootstrap intervals, including studentized, basic, and BCA intervals, were not as accurate as percentile intervals. See Supplementary Figures D11-D13 for detailed results.

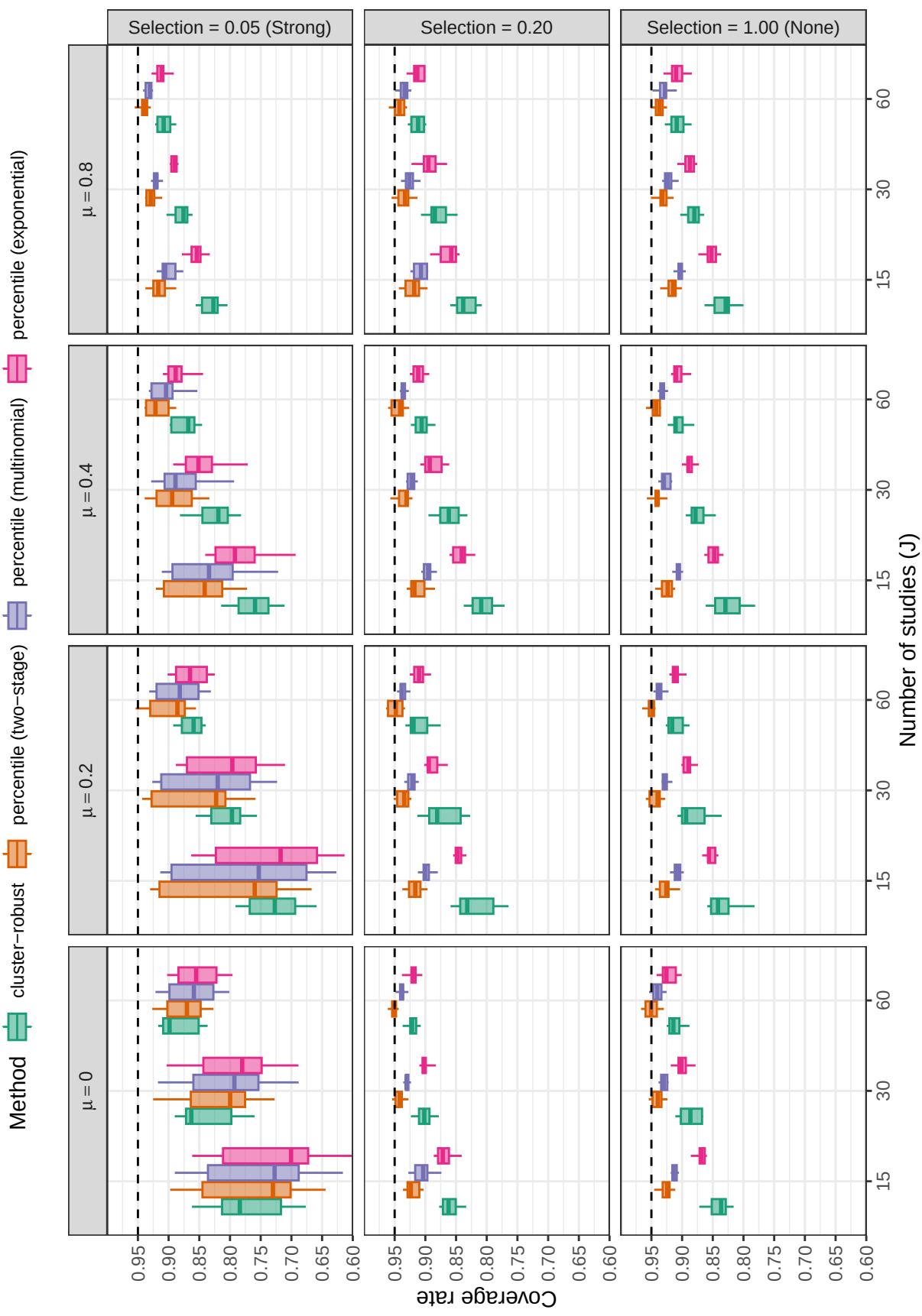

Figure 5
Coverage levels of confidence intervals based on the CML estimator of average effect size by number of studies, average SMD, and selection probability. Dashed lines correspond to the nominal confidence level of 0.95.

to selection mechanisms and thus are only indirectly informative about the strength or form of selection. The step-function model also embeds a familiar evidence-generating model that allows for heterogeneity of effects through inclusion of random effects and predictors of average effect size (i.e., meta-regression).[32] In contrast, well-known methods such as trim-and-fill[24,25] and more recent proposals such as p-curve[94] and p-uniform[95,96] are not as flexible and have been found to perform poorly when effects are heterogeneous[38].

We treated the step-function model as a description of the *marginal* distribution of the effect size estimates, effectively ignoring the dependence structure for purposes of estimating the model but accounting for it using cluster-robust sandwich estimation or clustered bootstrap inference. This strategy is appealing for its feasibility and because it connects to a selection process in which each individual effect size is selected on the basis of its statistical significance. We also studied two estimation methods for the marginal step-function model, composite marginal likelihood estimation and augmented-and-reweighted Gaussian likelihood estimation, and two inference strategies, based on either cluster-robust sandwich estimators or clustered bootstrap resampling.

Our simulations examined how well these estimators and inference techniques perform for recovering the average effect size under a selection process with a single step (at $\alpha_1 = .025$), compared to an estimator that accounts for dependence but not selection (i.e., the CHE-ISCW model) and to a variant of PET/PEESE that uses RVE. Across a broad range of conditions, we found that both estimators of the marginal step-function model show little bias overall and consistently out-perform PET/PEESE and CHE-ISCW under conditions with meaningful selective reporting. However, all the estimators face bias-variance trade-offs, which arise because the CHE-ISCW estimator (which does not directly adjust for selection) is substantially biased by selective reporting, whereas the CML and ARGL estimators have at most small biases. Under conditions where selection is absent or small and where average effect size is larger, the CHE-ISCW estimator has greater precision than

the estimators that adjust for selective reporting. Because selective reporting does not create much bias under such conditions, the additional variability that comes with estimating a selection model or PET/PEESE adjustment dominates the small reduction in bias that these methods provide. As a result, the step-function estimators are less accurate than the CHE-ISCW estimator under conditions where selective reporting is not strong or does not create meaningful bias.

The simulation results also demonstrated that the marginal step-function model estimators have better confidence interval coverage compared to the other methods, with coverage rates of the two-stage bootstrapped percentile confidence intervals approaching the nominal level of 0.95 for moderate sample sizes. Compared to the ARGL estimator, the CML estimator of the marginal mean is usually more accurate and had confidence interval coverage rates closer to nominal levels, although differences are fairly small. CML consistently out-performs ARGL for estimating between-study heterogeneity and the strength of selective reporting.

6.1 Limitations and Future Directions

Our approach of modeling the marginal distribution of effect sizes was motivated by the computational tractability of marginal models and by findings from prior simulations showing that univariate selection models perform well relative to alternative regression-based models to adjust for selective reporting bias.[62] However, this approach has several conceptual limitations that are important to note. First, such models do not reflect the structure of dependence among effect size estimates drawn from the same sample, but instead describe only the overall average and overall degree of heterogeneity of the effect size distribution. Because of this, they do not fully align with contemporary approaches for meta-analysis, which emphasize modeling the hierarchical structure of dependent effect sizes[52,54]. Second, focusing on the marginal distribution likely entails some loss of precision in parameter estimates. Better accounting for the dependence structure, such as through the use of

analytic weights, might allow for construction of more efficient estimators of the parameters of the evidence-generating process. Third, the marginal model provides no way to distinguish between study-level publication bias and effect-level selective outcome reporting. This strategy therefore precludes examination of more nuanced forms of selection, such as one where the probability that a given effect size is reported depends on the significance levels of other effect size estimates drawn from the same sample or on some broader feature of the study's results.

In addition to conceptual limitations, our simulation findings also need to be interpreted cautiously in light of the study's scope limitations. First, although our simulations covered a wide range of plausible conditions, the results remain generalizable only to the data-generating process examined. Of particular note, we generated data following a CHE effects model with primary study sample sizes and the number of effect sizes per study drawn from an empirical distribution of educational research studies. The performance of the step-function selection models and alternative selective reporting adjustments could change based on features of studies included in the synthesis, such as studies drawn from research areas that use smaller or larger samples or that tend to assess a smaller or larger number of outcomes. Likewise, there remains a need to investigate the robustness of the models to other evidence-generating processes, such as non-normal random effects distributions[97].

Second, the simulations examined the step-function selection model using a selection process that was compatible with the assumed model, in which the probability that an effect size was reported followed a step function in the one-sided p-value with a threshold at $\alpha_1 = .025$. Other simulation studies have shown that a one-step model can be more accurate than a more complex two-step model, even when the true data-generating process aligns with the latter model.[62] It may require a large number of primary studies to feasibly estimate models that include multiple steps in the selection function. Nonetheless, there may be meta-analytic datasets where a more complex set of steps is more appropriate, such as

when the data include a substantial number of negative effect size estimates.

Third and related to the previous point, the simulations were limited to evidence-generating processes that did not involve systematic predictors of the effect size distribution and where the strength of selective reporting was uniform and solely dependent on the p-value of each individual effect size. Other factors besides statistical significance of findings might affect a study's publication status. For instance, results from pre-registered replication studies might be insulated from selective reporting or subject to different reporting pressures than other forms of primary research[98]. Further evaluation of the CML and ARGL estimators is warranted to assess their performance in models involving moderators and their robustness to other selection mechanisms. In further development of step-function selection models, it may prove useful to model variation in the strength of reporting as a function of study characteristics such as pre-registration status[99].

Fourth and finally, the simulation findings we have reported here focused mostly on the performance of the estimators and confidence intervals for the overall average effect size, heterogeneity parameter, and selection parameter. We have not directly evaluated how well the estimators work as diagnostic *tests* for the presence of selective reporting of study results, although the below-nominal coverage rates of confidence intervals for the selection parameter suggests that they may not work well diagnostically. In univariate random effects models, likelihood ratio tests based on step-function selection models have been found to provide much stronger power for detecting selective reporting compared to alternatives such as Egger's regression or non-parametric symmetry tests[23]. Extension of such tests for meta-analyses of dependent effect sizes requires further development.

6.2 Conclusions

Selective reporting of positive, statistically significant findings in primary studies can potentially distort the results of meta-analyses. Detecting and adjusting for this form of bias is notoriously challenging—even in the simple setting where each sample contributes no more

than a single effect size estimate. These challenges are amplified with more complex data structures where the same study contributes multiple dependent effects. Nonetheless, meta-analysts must critically evaluate the evidence summarized in a synthesis, and this includes weighing the potential for bias from selective reporting and selective publication of primary study findings.

Based on the simulation results we have presented, we recommend using step function selection models with clustered bootstrap confidence intervals to assess selective reporting bias in syntheses of dependent effect sizes. As a parametric model built on specific assumptions about the selection process, the marginal step function model provides a useful complement to more agnostic techniques for identifying small-study effects, such as funnel plots and regression adjustment methods, the bias and accuracy of which are quite variable across data-generating processes. Likewise, estimated step function models could inform the use of sensitivity analysis[58] by using estimates of the strength of selection to guide assumptions about the maximal plausible degree of selection.

Consistent with recommendations from past work in the context of independent effect sizes[38,41], interpretation of any bias-corrected effect estimates needs to give consideration to the conditions under which the estimation method could be expected to perform well. Interpretation of marginal step-function models should focus mostly on the bias-adjusted average effect size, and selection parameter estimates should be interpreted cautiously in light of the mis-calibrated coverage levels of their cluster bootstrapped confidence intervals. More broadly, when applying and interpreting step function models, meta-analysts must consider the context of the evidence included in the synthesis, such as whether the effect size estimates are for focal results or merely incidental findings and whether studies were conducted under conditions where pressures to selectively reporting findings are present. As with inferences from any model, one's conclusions should be informed not only by the statistical results but also by knowledge of the research context.

Author Contributions

JEP: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Writing - original draft, Writing - review & editing, Supervision **MJ:** Methodology, Software, Validation, Formal Analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization **MC:** Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing, Project administration, Funding acquisition

Funding

This work was supported, in part, by the Institute of Educational Sciences, U.S. Department of Education through grant R305D220026 to the American Institutes of Research. The opinions expressed are those of the authors and do not represent the views of the Institute of the U.S. Department of Education.

Acknowledgements

We thank Laura Michaelson for feedback on a draft version of this article.

Data and Replication Materials

Code and data for replicating the empirical example and the Monte Carlo simulation study are available on the Open Science Framework at https://osf.io/v25rx/.

Conflict of Interest Statement

The authors declare no conflicts of interest.

References

1. Rothstein HR, Sutton AJ, Borenstein M. Publication bias in meta-analysis. In: Rothstein HR, Sutton AJ, Borenstein M, eds. *Publication Bias in Meta-Analysis: Prevention, Assessment, and Adjustments*. John Wiley & Sons; 2005:1-7. doi:10.1002/0470870168

2. Greenwald AG. Consequences of prejudice against the null hypothesis. *Psychological Bulletin*. 1975;82(1):1-20. doi:10.1037/h0076157

3. Sutton A. Publication bias. In: *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation; 2009:435-445.

4. Polanin JR, Tanner-Smith EE, Hennessy EA. Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of educational research*. 2016;86(1):207-236.

5. Williamson PR, Gamble C, Altman DG, Hutton JL. Outcome selection bias in meta-analysis. *Statistical Methods in Medical Research*. 2005;14(5):515-524. doi:10.1191/0962280205sm415oa

6. Hutton JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. *Journal of the Royal Statistical Society Series C: Applied Statistics*. 2000;49(3):359-370. doi:10.1111/1467-9876.00197

7. Dwan K, Gamble C, Williamson PR, Kirkham JJ, the Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias — an updated review. Boutron I, ed. *PLoS ONE*. 2013;8(7):e66844. doi:10.1371/journal.pone.0066844

8. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*. 2004;291(20):2457-2465.

9. Pigott TD, Valentine JC, Polanin JR, Williams RT, Canada DD. Outcome-reporting bias in education research. *Educational Researcher*. 2013;42(8):424-432.

10. Lancee M, Lemmens C, Kahn R, Vinkers C, Luykx J. Outcome reporting bias in randomized-controlled trials investigating antipsychotic drugs. *Translational psychiatry*. 2017;7(9):e1232-e1232.

11. John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*. 2012;23(5):524-532.

12. Franco A, Malhotra N, Simonovits G. Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science*. 2016;7(1):8-12. doi:10.1177/1948550615598377

13. Aert RCM van, Wicherts JM, Assen MALM van. Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. Macleod MR, ed. *PLOS ONE*. 2019;14(4):e0215052. doi:10.1371/journal.pone.0215052

14. O'Boyle Jr EH, Banks GC, Gonzalez-Mulé E. The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*. 2017;43(2):376-399.

15. Bartoš F, Maier M, Wagenmakers E, et al. Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics. *Research Synthesis Methods*. 2024;15(3):500-511. doi:10.1002/jrsm.1703

16. Marks-Anglin A, Chen Y. A historical review of publication bias. *Res Syn Meth*. 2020;11(6):725-742. doi:10.1002/jrsm.1452

17. Light RJ, Pillemer DB. *Summing Up*. Harvard University Press; 1984. https://books.google.com?id=qel3lAm4K6gC

18. Sterne JAC, Becker BJ, Egger M. The funnel plot. In: Rothstein HR, Sutton AJ, Borenstein M, eds. *Publication Bias in Meta-Analysis: Prevention, Assessment, and Adjustments*. John Wiley & Sons; 2005:73-98. doi:10.1002/0470870168

19. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics*. Published online 1994:1088-1101.

20. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629-634.

21. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in medicine*. 2006;25(20):3443-3457.

22. Moreno SG, Sutton AJ, Thompson JR, Ades A, Abrams KR, Cooper NJ. A generalized weighting regression-derived meta-analysis estimator robust to small-study effects and heterogeneity. *Statistics in medicine*. 2012;31(14):1407-1417.

23. Pustejovsky JE, Rodgers MA. Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*. 2019;10(1):57-71.

24. Duval S, Tweedie R. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the american statistical association*. 2000;95(449):89-98.

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

26. Stanley TD, Doucouliagos H. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*. 2014;5(1):60-78.

27. Bom PRD, Rachinger H. A kinked meta-regression model for publication bias correction. *Research Synthesis Methods*. 2019;10(4):497-514. doi:10.1002/jrsm.1352

28. Hedges LV. Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*. 1984;9(1):61-85.

29. Iyengar S, Greenhouse JB. Selection models and the file drawer problem. *Statistical Science*. 1988;3(1):109-117. doi:10.1214/ss/1177013012

30. Dear KBG, Begg CB. An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*. 1992;7(2):237-245.

31. Hedges LV. Modeling publication selection effects in meta-analysis. *Statistical Science*. 1992;7(2):246-255.

32. Vevea JL, Hedges LV. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*. 1995;60(3):419-435. doi:10.1007/BF02294384

33. Citkowicz M, Vevea JL. A parsimonious weight function for modeling publication bias. *Psychological Methods*. 2017;22(1):28-41. doi:10.1037/met0000119

34. Preston C, Ashby D, Smyth R. Adjusting for publication bias: Modelling the selection process. *Evaluation Clinical Practice*. 2004;10(2):313-322. doi:10.1111/j.1365-2753.2003.00457.x

35. Nelson N, Rosenthal R, Rosnow RL. Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*. 1986;41(11):1299-1301. doi:10.1037/0003-066X.41.11.1299

36. Rosenthal R, Gaito J. The interpretation of levels of significance by psychological researchers. *The Journal of Psychology: Interdisciplinary and Applied*. 1963;55(1):33-38. doi:10.1080/00223980.1963.9916596

37. Rosenthal R, Gaito J. Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports*. 1964;15(2):570. doi:10.2466/pr0.1964.15.2.570

38. Carter EC, Schönbrodt FD, Gervais WM, Hilgard J. Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*. 2019;2(2):115-144.

39. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*. 2003;22(13):2113-2126. doi:10.1002/sim.1461

40. Hong S, Reed WR. Using Monte Carlo experiments to select meta-analytic estimators. *Research Synthesis Methods*. 2021;12(2):192-215. doi:10.1002/jrsm.1467

41. McShane BB, Böckenholt U, Hansen KT. Adjusting for publication bias in meta-analysis an evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*. 2016;11(5):730-749. doi:10.1177/1745691616662243

42. Becker BJ. Multivariate meta-analysis. In: Brown SD, Tinsley HEA, eds. *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. Academic Press; 2000:499-525. doi:10.1016/B978-012691360-6/50018-5

43. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*. 2010;1(1):39-65. doi:10.1002/jrsm.5

44. Wu W, Duan J, Reed WR, Tipton E. What can we learn from 1000 meta-analyses across 10 different disciplines? *Working Papers in Economics*. 2025;(25/07). Accessed November 19, 2025. https://ideas.repec.org//p/cbt/econwp/25-07.html

45. Tipton E, Pustejovsky JE, Ahmadi H. Current practices in meta-regression in psychology, education, and medicine. *Research synthesis methods*. 2019;10(2):180-194.

46. Nakagawa S, Yang Y, Macartney EL, Spake R, Lagisz M. Quantitative evidence synthesis: A practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. *Environ Evid*. 2023;12(1):8. doi:10.1186/s13750-023-00301-6

47. Yang Y, Macleod M, Pan J, Lagisz M, Nakagawa S. Advanced methods and implementations for the meta-analyses of animal models: Current practices and future recommendations. *Neuroscience & Biobehavioral Reviews*. 2023;146:105016. doi:10.1016/j.neubiorev.2022.105016

48. Jackson D, Riley R, White IR. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*. Published online 2011:n/a-n/a. doi:10.1002/sim.4172

49. Kalaian HA, Raudenbush SW. A multivariate mixed linear model for meta-analysis. *Psychological Methods*. 1996;1(3):227-235. doi:10.1037/1082-989X.1.3.227

50. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine*. 2002;21(4):589-624. doi:10.1002/sim.1040

51. Konstantopoulos S. Fixed effects and variance components estimation in three-level meta-analysis: Three-level meta-analysis. *Research Synthesis Methods*. 2011;2(1):61-76. doi:10.1002/jrsm.35

52. Noortgate W van den, López-López JA, Marín-Martínez F, Sánchez-Meca J. Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*. 2013;45(2):576-594. doi:10.3758/s13428-012-0261-6

53. Noortgate W van den, López-López JA, Marín-Martínez F, Sánchez-Meca J. Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods*. 2015;47(4):1274-1294. doi:10.3758/s13428-014-0527-2

54. Pustejovsky JE, Tipton E. Meta-Analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Prevention Science*. 2022;23:425-438. doi:10.1016/j.jsp.2018.02.003

55. Tipton E. Small sample adjustments for robust variance estimation with meta-regression. *Psychological methods*. 2015;20(3):375.

56. Tipton E, Pustejovsky JE. Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*. 2015;40(6):604-634.

57. Joshi M, Pustejovsky JE, Beretvas SN. Cluster wild bootstrapping to handle dependent effect sizes in meta-analysis with a small number of studies. *Research Synthesis Methods*. 2022;13(4):457-477. doi:10.1002/jrsm.1554

58. Mathur MB, VanderWeele TJ. Sensitivity analysis for publication bias in meta-analyses. *J R Stat Soc C*. 2020;69(5):1091-1119. doi:10.1111/rssc.12440

59. Fernández-Castilla B, Declercq L, Jamshidi L, Beretvas SN, Onghena P, Noortgate W van den. Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *The Journal of Experimental Education*. Published online April 2019:1-20. doi:10.1080/00220973.2019.1582470

60. Rodgers MA, Pustejovsky JE. Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological methods*. 2021;26(2):141.

61. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*. 2011;343:d4002. doi:10.1136/bmj.d4002

62. Chen M, Pustejovsky JE. Adapting methods for correcting selective reporting bias in meta-analysis of dependent effect sizes. Published online October 23, 2024. doi:10.31222/osf.io/jq52s

63. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22.

64. Stanley TD, Doucouliagos H. Neither fixed nor random: Weighted least squares meta-analysis. *Statistics in Medicine*. 2015;34(13):2116-2127. doi:10.1002/sim.6481

65. Chen Y, Hong C, Riley RD. An alternative pseudolikelihood method for multivariate random-effects meta-analysis: An alternative pseudolikelihood method for multivariate random-effects meta-analysis. *Statistics in Medicine*. 2015;34(3):361-380. doi:10.1002/sim.6350

66. Hedges LV, Vevea J. Selection method approaches. In: Rothstein HR, Sutton AJ, Borenstein M, eds. *Publication Bias in Meta-Analysis*. John Wiley & Sons, Ltd; 2005:145-174. doi:10.1002/0470870168.ch9

67. Vevea JL, Woods CM. Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*. 2005;10(4):428-443. doi:10.1037/1082-989X.10.4.428

68. Cox DR, Reid N. A note on pseudolikelihood constructed from marginal densities. *Biometrika*. 2004;91(3):729-737. doi:10.1093/biomet/91.3.729

69. Lindsay BG. Composite likelihood methods. In: Prabhu NU, ed. *Contemporary Mathematics*. Vol 80. American Mathematical Society; 1988:221-239. doi:10.1090/conm/080/999014

70. Varin C. On composite marginal likelihoods. *AStA*. 2008;92(1):1-28. doi:10.1007/s10182-008-0060-7

71. Boos DD. Introduction to the bootstrap world. *Statistical science*. 2003;18(2):168-174.

72. Davison AC, Hinkley DV. *Bootstrap Methods and Their Applications*. Cambridge University Press; 1997.

73. Field CA, Welsh AH. Bootstrapping clustered data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2007;69(3):369-390. doi:10.1111/j.1467-9868.2007.00593.x

74. Leeden RVD, Meijer E, Busing FMTA. Resampling multilevel models. In: Leeuw JD, Meijer E, eds. *Handbook of Multilevel Analysis*. Springer New York; 2008:401-433. doi:10.1007/978-0-387-73186-5_11

75. Xu L, Gotwalt C, Hong Y, King CB, Meeker WQ. Applications of the fractional-random-weight bootstrap. *The American Statistician*. 2020;74(4):345-358. doi:10.1080/00031305.2020.1731599

76. Efron B. Better bootstrap confidence intervals. *Journal of the American Statistical Association*. 1987;82(397):171-185. doi:10.1080/01621459.1987.10478410

77. Carter EC, Kofler LM, Forster DE, McCullough ME. A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*. 2015;144(4):796.

78. Hagger MS, Wood C, Stiff C, Chatzisarantis NLD. Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*. 2010;136(4):495-525. doi:10.1037/a0019486

79. Hagger MS, Chatzisarantis NLD, Alberts H, et al. A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*. 2016;11(4):546-573. doi:10.1177/1745691616652873

80. Vohs KD, Schmeichel BJ, Lohmann S, et al. A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science*. 2021;32(10):1566-1581. doi:10.1177/0956797621989733

81. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2025. https://www.R-project.org/

82. Pustejovsky JE, Joshi M, Citkowicz M. *Metaselection: Meta-Analytic Selection Models with Cluster-Robust and Cluster-Bootstrap Standard Errors for Dependent Effect Size Estimates.*; 2025. https://github.com/jepusto/metaselection

83. Center for High Throughput Computing. Center for high throughput computing. Published online 2006. doi:10.21231/GNT1-HW21

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

85. Pustejovsky JE. *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections.*; 2024. https://CRAN.R-project.org/package=clubSandwich

86. Joshi M, Pustejovsky JE. *Simhelpers: Helper Functions for Simulation Studies.*; 2024. https://CRAN.R-project.org/package=simhelpers

87. Nash JC, Varadhan R. Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software*. 2011;43(9):1-14. doi:10.18637/jss.v043.i09

88. Hasselman B. *Nleqslv: Solve Systems of Nonlinear Equations.*; 2023. https://CRAN.R-project.org/package=nleqslv

89. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *Journal of Open Source Software*. 2019;4(43):1686. doi:10.21105/joss.01686

90. Borenstein M, Hedges LV. Effect sizes for meta-analysis. *The handbook of research synthesis and meta-analysis*. 2019;3:207-243.

91. Boos DD, Zhang J. Monte carlo evaluation of resampling-based hypothesis tests. *Journal of the American Statistical Association*. 2000;95(450):486-492. doi:10.1080/01621459.2000.10474226

92. Fisher Z, Tipton E, Zhang H. *Robumeta: Robust Variance Meta-Regression.*; 2024. https://CRAN.R-project.org/package=robumeta

93. Kraft MA. Interpreting effect sizes of education interventions. *Educational researcher*. 2020;49(4):241-253.

94. Simonsohn U, Nelson LD, Simmons JP. P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*. 2014;9(6):666-681. doi:10.1177/1745691614553988

95. Assen MALM van, Aert RCM van, Wicherts JM. Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*. 2015;20(3):293-309. doi:10.1037/met0000025

96. Aert RCM van, Wicherts JM, Assen MALM van. Conducting meta-analyses based on p values: Reservations and recommendations for applying p -uniform and p -curve. *Perspectives on Psychological Science*. 2016;11(5):713-729. doi:10.1177/1745691616650874

97. Hedges LV, Vevea JL. Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*. 1996;21(4):299. doi:10.3102/10769986021004299

98. Aert RCM van. Meta-analyzing nonpreregistered and preregistered studies. *Psychological Methods*. Published online February 17, 2025. doi:10.1037/met0000719

99. Coburn KM, Vevea JL. Publication bias as a function of study characteristics. *Psychological Methods*. 2015;20(3):310-330. doi:10.1037/met0000046