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

May 28, 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. 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 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, and unadjusted estimates of average effect sizes may be preferable if the strength of selective reporting is mild.

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

1 Introduction

Meta-analysis methods are used to synthesize quantitative findings across multiple sources of evidence, such as multiple studies that evaluate the effects of the same intervention. Because they rely on findings from primary studies as input data, the validity of conclusions from a meta-analytic synthesis depends critically on the reporting practices of researchers, journal editors, and peer reviewers. If the findings accessible to meta-analysts are not a complete or representative record of the research that has been conducted on a topic, then meta-analytic summaries could be systematically biased (Rothstein et al., 2005). 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 (Carter et al., 2019).

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 (Chan et al., 2004) and 2.4 times more likely in education (Pigott et al., 2013). Research across various fields, from clinical trials of antipsychotics (e.g., Lancee et al., 2017) to psychology studies (Franco et al., 2016; e.g., John et al., 2012) and management research (e.g., O'Boyle Jr et al., 2017), indicates that selective outcome reporting is widespread and often driven by 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 (Marks-Anglin & Chen, 2020; Rothstein et al., 2005). One common graphical diagnostic is the funnel plot, a simple scatterplot of effect size estimates versus a measure of their precision (Kossmeier et al., 2020; Light & Pillemer, 1984; Sterne et al., 2005). Widely used statistical diagnostics include the rank correlation test by Begg and

Mazumdar (1994); Egger's regression test (Egger et al., 1997; Harbord et al., 2006; Macaskill et al., 2001; Moreno et al., 2012; Peters et al., 2006; Pustejovsky & Rodgers, 2019; Stanley, 2008; Thompson & Sharp, 1999); the trim-and-fill adjustment (Duval & Tweedie, 2000a, 2000b); and various regression adjustment methods including the precision effect test (PET), precision effect estimate with standard error (PEESE), and PET-PEESE technique by Stanley and Doucouliagos (2014), and the endogenous kink meta-regression (Bom & Rachinger, 2019).

Another class of methods for assessing and correcting selective reporting are p-value selection models.[1] 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. Early proposals of this form include Hedges (1984), Iyengar and Greenhouse (1988), and Dear and Begg (1992). Hedges (1992) and Vevea and Hedges (1995) proposed *step-function* models where the selection function is piece-wise constant, with steps at psychologically salient significance levels such as $\alpha = .05$. Other forms involve selection functions based on beta densities (Citkowicz & Vevea, 2017), power curves, and a variety of other parametric forms (Preston et al., 2004).

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; they are therefore more interpretable than tests or adjustments for small-study effects, which are agnostic with respect to the specific mechanism of selective reporting. Second, selection models can allow for effect heterogeneity with a random effect term. Findings from simulations indicate that selection

[1] A further class of selection models exists that depends both on the effect size estimate and its standard error (e.g., Copas, 1999; Copas & Li, 1997; Copas & Shi, 2001). We focus here on the selection models that depend solely on the p-value because they have shown promise in prior simulation studies (Carter et al., 2019; Terrin et al., 2003) and because the second set of selection models have identification issues that make them more useful as sensitivity analyses than as estimation methods (Hedges & Vevea, 2005; Sutton, 2009).

models outperform simpler alternative methods when effect sizes are heterogeneous (Carter et al., 2019; Terrin et al., 2003). Third, selection models 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. In addition to these features, the Vevea and Hedges (1995) step-function model is particularly useful because it allows one to specify cut points that capture simple but plausible forms of selective reporting (e.g., $p < 0.01$ and $p < 0.05$, see, e.g., Greenwald, 1975; Nelson et al., 1986; Rosenthal & Gaito, 1963, 1964).

1.1 Dependent effect sizes

The vast majority of the work on selective reporting—including the development of selection models—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 (Becker, 2000). 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 (Hedges et al., 2010).

Effect size dependencies are very common in social science synthesis, as well as in other research areas. 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). In a survey of systematic reviews published in 2016 across several prominent journals, Tipton et

al. (2019) found that primary studies contributed an average of 3.1 effect sizes to systematic reviews published in *Psychological Bulletin*; 11.0 effect sizes to reviews published in *Journal of Applied Psychology*; and 5.0 effect sizes to reviews published in *Review of Educational Research*. Surveys of recent meta-analyses on topics in environmental sciences (Nakagawa et al., 2023) and neurobiological research involving animal models (Yang et al., 2023) 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 multi-level meta-analyses (Konstantopoulos, 2011; Van den Noortgate et al., 2013, 2015), robust variance estimation (RVE, Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015), and combinations thereof (Pustejovsky & Tipton, 2022). 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.

The original form of RVE was based on asymptotic approximations that required a relatively large number of independent studies. Tipton (2015) and Tipton and Pustejovsky (2015) developed small-sample adjustments for standard errors and hypothesis tests based on RVE so that the desirable asymptotic properties are maintained even when the number of studies is small. A closely related strategy is to use bootstrap re-sampling to approximate the distribution of test statistics. For instance, Joshi et al. (2022) examined a cluster-wild bootstrap method for hypothesis testing in meta-regression with dependent effects, finding that it led to refined Type I error rates and improved power compared to analytic approximations. 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 (2020) proposed a sensitivity analysis based on the simplest possible form of the Vevea and Hedges (1995) step-function selection model. 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 methods (Hedges et al., 2010). 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 (as in Egger et al., 1997), combined with RVE or multilevel meta-analysis to handle dependent effect sizes (Fernández-Castilla et al., 2019; Rodgers & Pustejovsky, 2021). Rodgers and Pustejovsky (2021) found that this method has limited power to detect selective reporting under common meta-analytic scenarios. It also has the same limitations as univariate Egger's regression, in that it tests for small-study effects (or asymmetry in the funnel plot distribution), which could have causes other than selective reporting. 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 (2024) reviewed a range of existing techniques for estimating average effect sizes in the presence of selective reporting and proposed adaptations of some existing methods to accommodate dependent effect sizes. The proposed adaptations applied to methods that could 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. 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, even though the selection models ignore the dependence structure of the data. However, because selection models rely on the assumption that all effect sizes are independent, confidence intervals generated from the selection models did not have accurate coverage. In light of these findings, Chen and Pustejovsky (2024) noted a need to further develop selection models that can account for dependent effect sizes.

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. We make three main contributions. First, we describe extensions of RVE and bootstrap re-sampling techniques to assess uncertainty in selection model parameter estimates. Second, we describe and evaluate two different strategies for estimating model parameters: one using conventional maximum likelihood estimation methods and a novel strategy based on a reweighted random effects model with inverse probability of selection weights. Third, we use simulation to evaluate the performance of RVE and bootstrap re-sampling for constructing confidence intervals for average effect sizes.

The remainder of the paper is organized as follows. In the next section, we describe the step-function selection model and detail our estimation and inference strategies. 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

Selection models involve two components (Hedges & Vevea, 2005). 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.

In applying selection models to datasets involving dependent effects, we propose to model the *marginal* distribution of the effect size estimates—that is, the distribution of each estimate considered singly—rather than modeling the joint distribution of the multiple observed effects reported in a primary 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 so would not, for instance, allow for a decomposition of heterogeneity into between-study and within-study variance. Likewise, models for the marginal distribution do not allow one to distinguish between selective publication of full studies versus selective reporting of individual outcomes. Nonetheless, we believe that the strategy is worth pursuing both because it is feasible and because it captures a plausible form of selection, in which reporting is influenced by the significance level of individual effect size estimates.

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

$$\Pr\left(O = 1 | p^*\right) \propto w\left(p^*; \boldsymbol{\lambda}\right) \tag{2}$$

for some known function $w\left(.; \boldsymbol{\lambda}\right)$ 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 by Hedges (1992) on random effects models without predictors, Vevea and Hedges (1995) 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

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

for $h = 0, ..., H$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_H)$. Equation (3) can be written equivalently as

$$w(Y^*, \sigma^*; \boldsymbol{\lambda}) = \lambda_h \quad \text{if} \quad \sigma^* \Phi^{-1}\left(1 - \alpha_{h+1}\right) \leq Y^* < \sigma^* \Phi^{-1}\left(1 - \alpha_h\right). \tag{4}$$

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 $p^* \leq \alpha_1$.

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

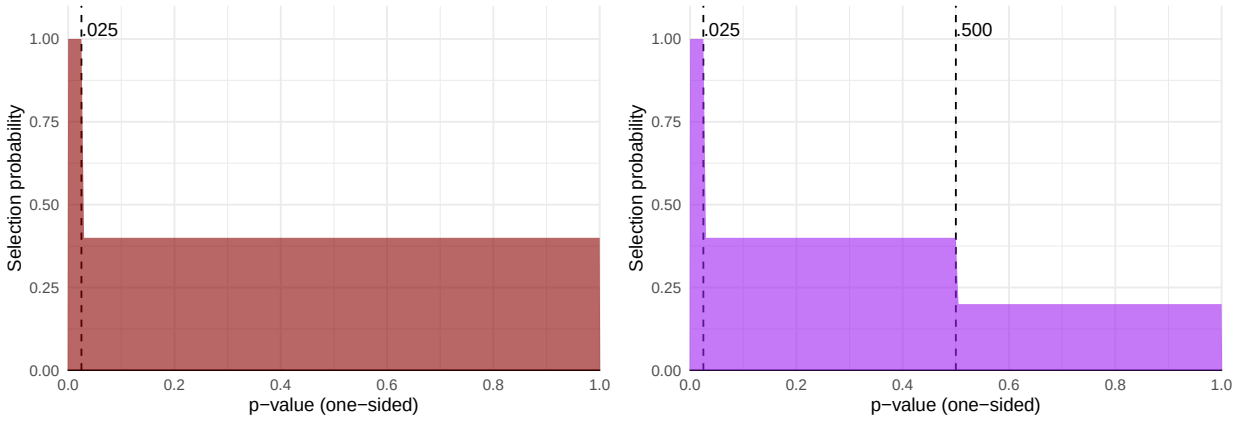

Figure 1
Examples of step functions

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 $O = 1$. By Bayes rule,

$$\Pr(Y^* = y | O = 1, \sigma^* = \sigma) = \frac{\Pr(O = 1 | Y^* = y, \sigma^* = \sigma) \times \Pr(Y^* = y | \sigma^* = \sigma)}{\Pr(O = 1 | \sigma^* = \sigma)},$$

where the selection process defines the first term in the numerator and the evidence-generating process defines the second term in the numerator. The marginal density of an observed effect size estimate Y with standard error σ therefore has the form

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

where

$$A(\mathbf{x}, \sigma; \boldsymbol{\beta}, \tau^2, \boldsymbol{\lambda}) = \int_{\mathbb{R}} w\left(y, \sigma; \boldsymbol{\lambda}\right) \times \frac{1}{\sqrt{\tau^2 + \sigma^2}} \phi\left(\frac{y - \mathbf{x}\boldsymbol{\beta}}{\sqrt{\tau^2 + \sigma^2}}\right) dy. \tag{6}$$

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).

For the step-function selection process, the $A(\mathbf{x}, \sigma; \boldsymbol{\beta}, \tau^2, \boldsymbol{\lambda})$ term in the step-function composite likelihood can be computed using the closed-form expression

$$A_{ij} = A(\mathbf{x}_{ij}, \sigma_{ij}; \boldsymbol{\beta}, \tau^2, \boldsymbol{\lambda}) = \sum_{h=0}^{H} \lambda_h B_{hij} \tag{7}$$

where

$$B_{hij} = \Phi\left(c_{hij}\right) - \Phi\left(c_{h+1,ij}\right), \tag{8}$$

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 $h = 0, ..., H$ (Vevea & Hedges, 1995). The quantity B_{hij} corresponds to the probability that, prior to selection, a generated effect size estimate with predictor $\mathbf{x}_{ij}$ and standard error σ_{ij} will have a p-value falling in the interval $\alpha_h < p^* \leq \alpha_{h+1}$.

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 (Citkowicz & Vevea, 2017; Hedges, 1992; Vevea & Hedges, 1995) or on sensitivity analysis methods that

treat the selection model as known (Mathur & VanderWeele, 2020; Vevea & Woods, 2005). We consider several 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.[2] 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 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 observation (here, each observed effect size estimate) as if all observations were mutually independent (Cox & Reid, 2004; Lindsay, 1988; Varin, 2008). Thus, we assume that the observed effect size estimates were generated from Equation (5). 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

$$\begin{aligned} l_{ij}^{M}(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) &= \log f\left(Y = y_{ij} | \sigma_{ij}, \mathbf{x}_{ij}\right) \\ &\propto \log w\left(y_{ij}, \sigma_{ij}; \boldsymbol{\zeta}\right) - \frac{1}{2}\frac{\left(y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta}\right)^2}{\exp(\gamma) + \sigma_{ij}^2} \\ &\quad - \frac{1}{2}\log\left(\exp(\gamma) + \sigma_{ij}^2\right) - \log A\left(\mathbf{x}_{ij}, \sigma_{ij}; \boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right). \end{aligned} \tag{9}$$

[2] Another reason to consider weights is that analytic weights could be used to improve the efficiency of the parameter estimators, similar to the working model weights proposed by Hedges et al. (2010) for random effects meta-analysis and meta-regression. For instance, consider the basic meta-analysis context with no predictors. If effect size estimates from the same study are correlated, then a study with k_j observed effect size estimates could contribute somewhat less than k_j independent pieces of information. Down-weighting the effect sizes from study j based on the number of reported effect sizes k_j might therefore improve the efficiency of the estimator for the average effect size β and variance τ^2.

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

$$l^M(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = \sum_{j=1}^{J} \sum_{i=1}^{k_j} a_{ij} l_{ij}^M(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}). \tag{10}$$

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 (10).

Under the assumption that 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,

$$\sum_{j=1}^{J} \mathbf{S}_j \left(\hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\boldsymbol{\zeta}}\right) = \mathbf{0} \tag{11}$$

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:

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

$$S_{\gamma j}(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = \sum_{i=1}^{k_j} a_{ij} \frac{\partial l_{ij}^M(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta})}{\partial \gamma} \tag{13}$$

$$\mathbf{S}_{\zeta j}(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = \sum_{i=1}^{k_j} a_{ij} \frac{\partial l_{ij}^M(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}}. \tag{14}$$

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,

$$\mathbf{H}(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = \sum_{j=1}^{J} \frac{\partial \mathbf{S}_j(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta})}{\partial \left(\boldsymbol{\beta}' \; \gamma \; \boldsymbol{\zeta}'\right)}, \tag{15}$$

exact expressions for which are given in 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:

$$\mathbf{V}^{CML} = \hat{\mathbf{H}}^{-1} \left(\sum_{j=1}^{J} \hat{\mathbf{S}}_j \hat{\mathbf{S}}_j' \right) \hat{\mathbf{H}}^{-1}. \tag{16}$$

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 (2020) 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, $w_{ij} = w(y_{ij}, \sigma_{ij}; \boldsymbol{\zeta})$. Following Mathur and VanderWeele (2020), 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

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

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 $\boldsymbol{\zeta}$.

We propose to estimate $\boldsymbol{\zeta}$ by augmenting 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 (14), the full set of estimating equations is

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

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 $\tilde{\boldsymbol{\zeta}}$.[3]

For the step-function selection process, the score equation with respect to $\boldsymbol{\zeta}$ has an interesting and intuitively interpretable form. Observe that the probability that an observed effect size estimate with predictor $\mathbf{x}_{ij}$ and standard error σ_{ij} will have a p-value falling in the interval $\alpha_h < p^* \leq \alpha_{h+1}$ is $E_{hij} = \exp(\zeta_h) \times B_{hij}/A_{ij}$, where B_{hij} is given in Equation (8). Accounting for the prior weights, the expected number of observed effect size estimates falling into the interval $(\alpha_h, \alpha_{h+1}]$ is therefore

$$E_h = \sum_{j=1}^{J} \sum_{i=1}^{k_j} a_{ij} \times \frac{\exp(\zeta_h) \times B_{hij}}{A_{ij}}. \tag{23}$$

---

[3] For computational purposes, it is useful to observe that the solution to Equation (17) involves a weighted least squares estimator for $\boldsymbol{\beta}$. Given values of γ and $\boldsymbol{\zeta}$, $\tilde{\boldsymbol{\beta}}$ is

$$\tilde{\boldsymbol{\beta}}(\gamma, \boldsymbol{\zeta}) = \left(\sum_{j=1}^{J} \sum_{i=1}^{k_j} v_{ij} \mathbf{x}'_{ij} \mathbf{x}_{ij} \right)^{-1} \sum_{j=1}^{J} \sum_{i=1}^{k_j} v_{ij} \mathbf{x}'_{ij} \mathbf{y}_{ij},$$

where $v_{ij} = \frac{a_j}{w_{ij}\left(\exp(\gamma) + \sigma^2_{ij}\right)}$. The weighted least squares solution allows $\boldsymbol{\beta}$ to be profiled out of the estimating equations, so that Equation (22) need be solved only for γ and $\boldsymbol{\zeta}$.

Let $K_h = \sum_{j=1}^{J} \sum_{i=1}^{k_j} a_{ij} \times I\left(\alpha_h < p_{ij} \leq \alpha_{h+1}\right)$ denote the weighted count of observed effect size estimates with p-values falling into the interval $(\alpha_h, \alpha_{h+1}]$, for $h = 0, ..., H$. The score of the step-function model with respect to the h^{th} component of $\boldsymbol{\zeta}$ can then be written simply as

$$S_{\zeta h} = K_h - E_h. \tag{24}$$

Thus, by setting $\mathbf{S}_{\boldsymbol{\zeta}} = \mathbf{0}$, the estimator of $\boldsymbol{\zeta}$ is taken to be the values that equate the observed number of effect size estimates in each interval with the expected number of estimates under the step-function model.

We will consider conducting inferences for the ARGL estimators using cluster-robust sandwich variance estimators that have the same form as (16). 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 (16) require a large number of independent clusters (i.e., large J) to provide accurate assessments of uncertainty. For regular meta-analysis or meta-regression models, small-sample refinements are available that provide accurate inference even with a limited number of clusters (Tipton, 2015; Tipton & Pustejovsky, 2015). However, these small-sample refinements have not been extended to step-function selection models. We instead consider alternative inference techniques, including several forms of bootstrapping, that 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 means to calculate standard errors and confidence intervals.

Many different bootstrap sampling schemes have been described that apply to different data structures and require different assumptions (Davison & Hinkley, 1997). 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 (Field & Welsh, 2007; Leeden et al., 2008), and fractional random weight bootstrap (Rubin, 1981; Xu et al., 2020). Here, we focus on the two-stage bootstrap; details about other bootstrap techniques can be found in Appendix C.

In the two-stage bootstrap, each pseudo-sample is generated by, first, randomly drawing J clusters of observations with replacement from the original sample and then, second, 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, all as described by Davison and Hinkley (1997), including the percentile CI, basic CI, studentized CI, and the bias-corrected-and-accelerated CI proposed by Efron (1987). Appendix C 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 Carter et al. (2015). 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 (Hagger et al., 2010). Carter et al. (2015) 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. Some primary studies reported effects for multiple outcome tasks, leading to dependent effects. 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).

To mitigate possible effects of selective reporting, Carter et al. (2015) 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. We conducted the analyses using R Version 4.4.3 (R Core Team, 2025).

If selective reporting were not a concern, a correlated-and-hierarchical effects model (CHE, Pustejovsky & Tipton, 2022) would be one way to summarize the distribution of ego depletion effects. Based on a CHE model, the overall average effect estimate was 0.46, 95% CI [0.34, 0.59]. Alternately, Chen and Pustejovsky (2024) proposed estimating average effect sizes using 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. 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 (Stanley & Doucouliagos, 2014), 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 (Pustejovsky et al., 2025). 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

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

Parameter	CML estimator			ARGL estimator		
	Estimate (SE)	Cluster-Robust CI	Percentile Bootstrap CI	Estimate (SE)	Cluster-Robust CI	Percentile Bootstrap CI
One-step						
β	0.25 (0.09)	[0.07, 0.44]	[0.05, 0.47]	0.27 (0.07)	[0.14, 0.41]	[0.11, 0.47]
τ^2	0.11 (0.04)	[0.06, 0.20]	[0.02, 0.19]	0.11 (0.04)	[0.06, 0.21]	[0.01, 0.21]
λ_1	0.27 (0.14)	[0.10, 0.75]	[0.07, 0.87]	0.30 (0.10)	[0.16, 0.56]	[0.05, 1.20]
Two-step						
β	0.23 (0.11)	[0.02, 0.44]	[-0.01, 0.48]	0.25 (1.45)	[-2.59, 3.10]	[-0.05, 0.54]
τ^2	0.11 (0.04)	[0.06, 0.21]	[0.03, 0.19]	0.12 (0.33)	[0.00, 32.00]	[0.00, 0.20]
λ_1	0.27 (0.14)	[0.10, 0.74]	[0.07, 0.86]	0.29 (1.65)	[0.00,>100]	[0.05, 1.28]
λ_2	0.22 (0.16)	[0.06, 0.90]	[0.04, 1.18]	0.26 (2.17)	[0.00,>100]	[0.01, 3.14]

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

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 0.23 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.

4 Simulation Methods

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 (Stanley & Doucouliagos, 2014). We evaluated the point estimators and robust variance estimators in terms of convergence rates, bias, accuracy, and confidence interval coverage for recovering the average

effect size of the unselected distribution. To limit the computational burden, we evaluated the performance of bootstrap confidence intervals only in a subset of the conditions.

We ran the simulation in R Version 4.4.1 (R Core Team, 2025) using the high-throughput computing cluster at the University of Wisconsin - Madison (Center for High Throughput Computing, 2006). The simulation code drew on functionality from several R packages, including `metafor` (Viechtbauer, 2010), `clubSandwich` (Pustejovsky, 2024), `simhelpers` (Joshi & Pustejovsky, 2024), `optimx` (Nash & Varadhan, 2011), `nleqslv` (Hasselman, 2023), and `tidyverse` (Wickham et al., 2019).

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 contributed $k_j^* \geq 1$ effect size estimates prior to selective reporting. Let T_{ij} denote effect size estimate i from study j and let σ_{ij} denote its standard error, for $i = 1, ..., k_j^*$ and $j = 1, ..., J^*$. Let N_j denote the effective sample size from study j.

To generate each meta-analytic dataset, we sampled effective sample sizes[4] 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. We assumed 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

[4] 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.

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^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 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. 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 (Borenstein & Hedges, 2019) 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 as described in Section 2.4. For both estimators, we calculated cluster-robust standard errors using large-sample sandwich formulas. 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. To maintain computational feasibility, we used $B = 399$ bootstrap replications of each estimator.

We compared the performance of the step-function CML and ARGL estimators to 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) as proposed by Chen and Pustejovsky (2024).[5] 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. Second, we implemented a variation of the PET/PEESE estimator originally proposed by Stanley and Doucouliagos (2014), adapted to accommodate dependent effect sizes.[6] Following Stanley and Doucouliagos (2014), we combined the estimators by using

[5] CHE-ISCW is based on the working model

$$y_{ij} = \mu + u_j + v_{ij} + e_{ij}, \tag{25}$$

where $u_j \sim N(0, \tau^2)$, $v_{ij} \sim N(0, \omega^2)$, $e_{ij} \sim N(0, \sigma_{ij}^2)$, and $\text{cor}(e_{hj}, e_{ij}) = \rho$, with the sampling variances treated as known and assuming a common correlation between sampling errors within the same study. Rather than estimating μ using weights based on the full working model, CHE-ISCW instead uses generalized least squares with weighting matrices that are the inverse of the variance-covariance matrix of the sampling errors $e_{1j}, ..., e_{k_j j}$ only. This has the effect of allocating more weight to studies with smaller sampling variances, providing some robustness to selective reporting of study results.

[6] The PET estimator is based on the working model

$$T_{ij} = \mu + \beta \times \frac{2}{\sqrt{N_j}} + e_{ij} \tag{26}$$

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

$$T_{ij} = \mu + \beta \times \frac{4}{N_j} + e_{ij} \tag{27}$$

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, $\text{cor}(e_{hj}, e_{ij}) = \rho$.

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 estimators, we calculated confidence intervals using cluster-robust variance estimation with the CR2 small-sample correction and Satterthwaite degrees of freedom (Chen & Pustejovsky, 2024).

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), within-study heterogeneity ratio (ω^2/τ^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. 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.

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)	0.05, 0.15, 0.30, 0.45	0.15, 0.45
Heterogeneity ratio (ω^2/τ^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

In the full simulation, 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 by Kraft (2020). 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 (Tipton et al., 2019).

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 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.

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. Convergence measures the proportion of replications in which an estimation algorithm provides a valid numerical solution. Bias measures whether an estimator generates values that fall systematically above or below the true parameter. RMSE measures the overall accuracy of an estimator, capturing both systematic bias and sampling variance. 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. Due to the computational demands of bootstrapping, we used $B = 399$ bootstraps per replication—fewer 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 similar to one proposed by Boos and Zhang (2000). 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 from the full set of $B = 399$ bootstraps. We computed coverage rates separately for each value of B and fit a linear regression of the coverage rate on $1/B$. We

used the intercept from this regression as the predicted coverage rate of confidence intervals based on $B = 1999$ bootstraps.

5 Simulation Results

We organize our presentation of simulation results by first considering the properties of point and interval 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 also examine the calibration of cluster-robust and bootstrap confidence intervals based on the CML and ARGL estimators. We then briefly consider the bias and accuracy of estimators of the marginal variance of the effect size distribution and the selection parameter in the step-function model.

5.1 Average Effect Size

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 94.40%. For the ARGL estimator, convergence was above 99.80% across all conditions. Supplementary Figure D1 depicts the range of convergence rates of the CML and ARGL estimators. We evaluated the performance characteristics of each estimator across the replications where it converged.

5.1.1 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

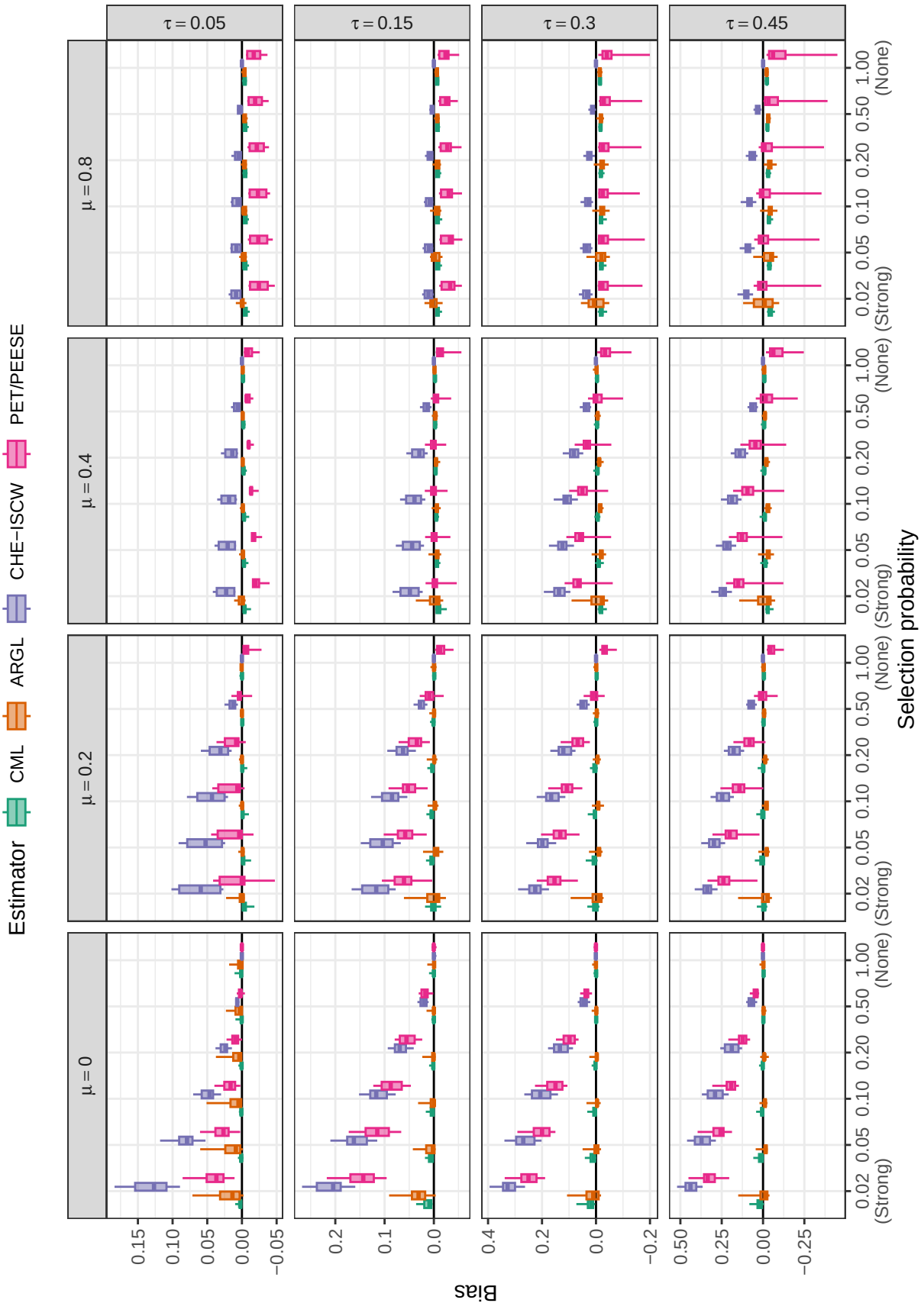

Figure 2

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

because the bias of some estimators is strongly influenced by the degree of heterogeneity.

The CML estimator has negligible or small bias across all conditions. Its largest bias was 0.08, occurring when selective reporting is very strong, average effect size is zero, and heterogeneity is large. The small bias of the CML estimator is stable across varying degrees of outcome correlation, within-study heterogeneity, number of observed studies, and primary study sample size. Similar to the CML estimator, the ARGL estimator also has negligible or small bias across most conditions, although its bias increases when average effect is zero and selection is very strong.

In contrast to the estimators based on the marginal selection model, the comparison 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. 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.

5.1.2 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

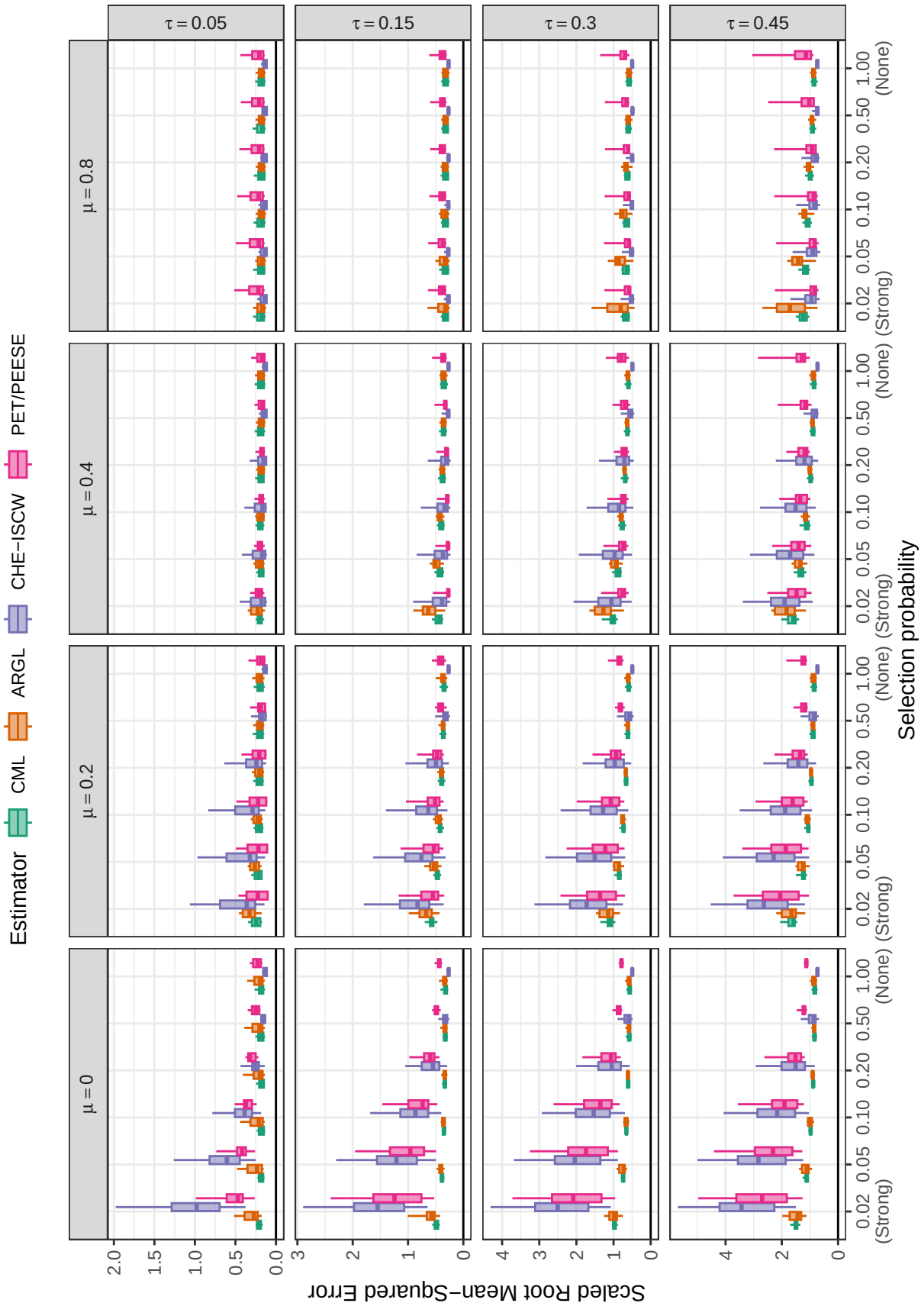

Figure 3

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

constructed in the same way as Figure 2. Figures D2 through D6 in Appendix D provide greater detail about the relative accuracy of the four methods by plotting the ratio of RMSEs for each pair of methods. 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 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).

The bias-variance trade-offs faced by all the estimators 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. However, 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.

5.1.3 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.[7] Coverage rates are below the nominal rate of 0.95 for all methods across most conditions. 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,

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

[7] 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 D7 depicts the full range of coverage rates.

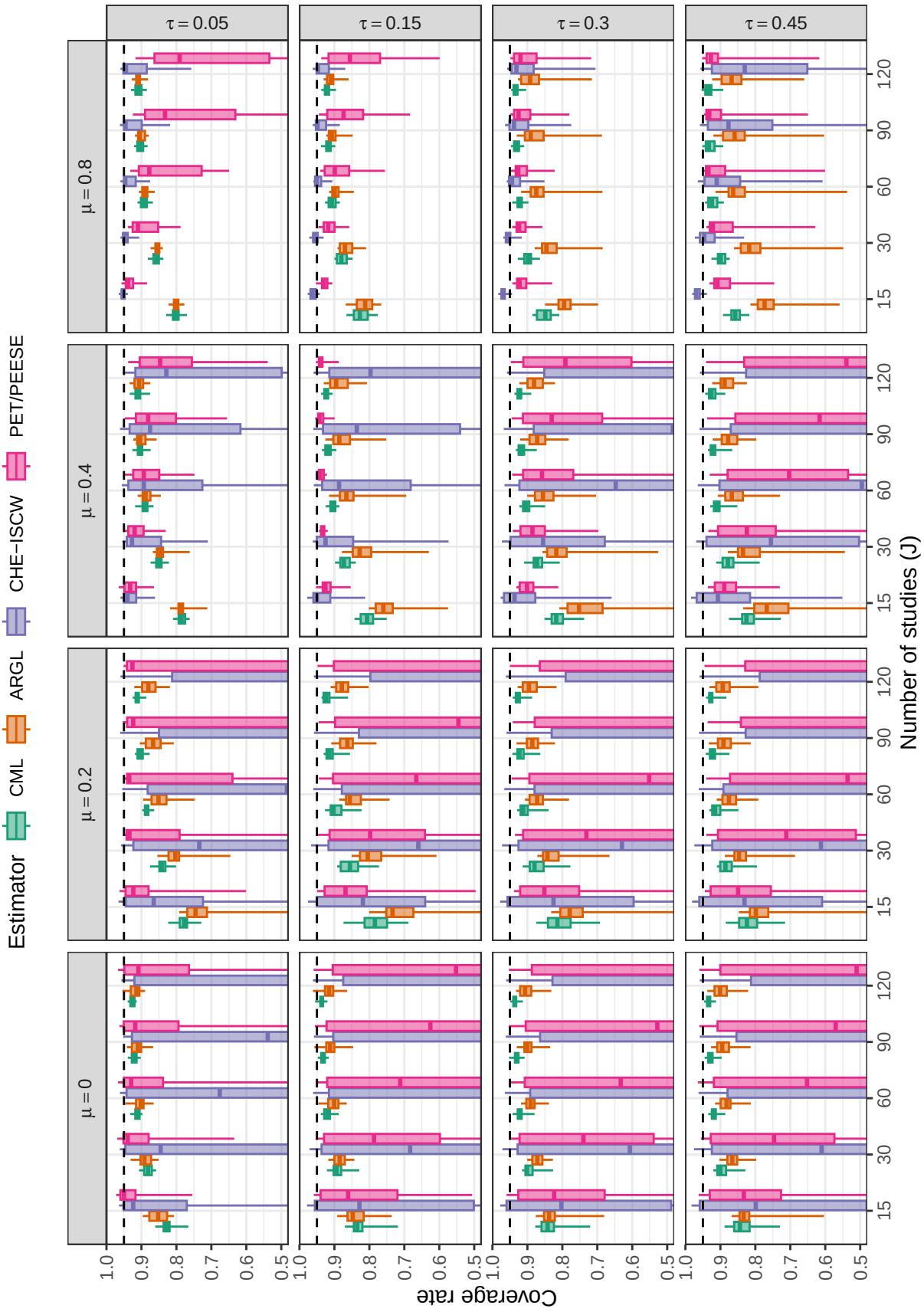

Figure 4

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

intervals under a more limited range of data-generating conditions, including a maximum 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 (non-parametric), 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. In particular, the percentile intervals with two-stage clustered bootstrap re-sampling provided the best coverage levels, exceeding 90% coverage across nearly all data-generating conditions, even with only $J = 15$ primary studies per meta-analysis. Coverage levels of the other bootstrap intervals, including studentized, basic, and BCa intervals, were not as accurate as percentile intervals (see Supplementary Figures D8-D10 for detailed results). Coverage levels of intervals based on the ARGL estimator followed very similar patterns to those for the CML estimator (Supplementary Figures D11-D13).

5.2 Effect Size Variance

We briefly consider estimation of the marginal heterogeneity of the effect size distribution, for which the CHE-ISCW, CML, and ARGL methods are all relevant. Figure E1 depicts the bias for the CHE-ISCW, CML, and ARGL estimators of log-heterogeneity $\gamma = \log(\tau^2)$. In most conditions, the estimators are biased in the negative direction. Bias is high for all three estimators in conditions where the between-study heterogeneity is low ($\tau = 0.05$) with bias improving as between-study heterogeneity increases. The CML estimator is less strongly biased than the CHE-ISCW estimator under conditions where there is strong selection. However, the ARGL estimator is badly biased downward, especially under conditions where selection is strong, average SMD is low and between-study heterogeneity is low.

Figure E2 depicts the scaled RMSE for estimators of log-heterogeneity, with Figures

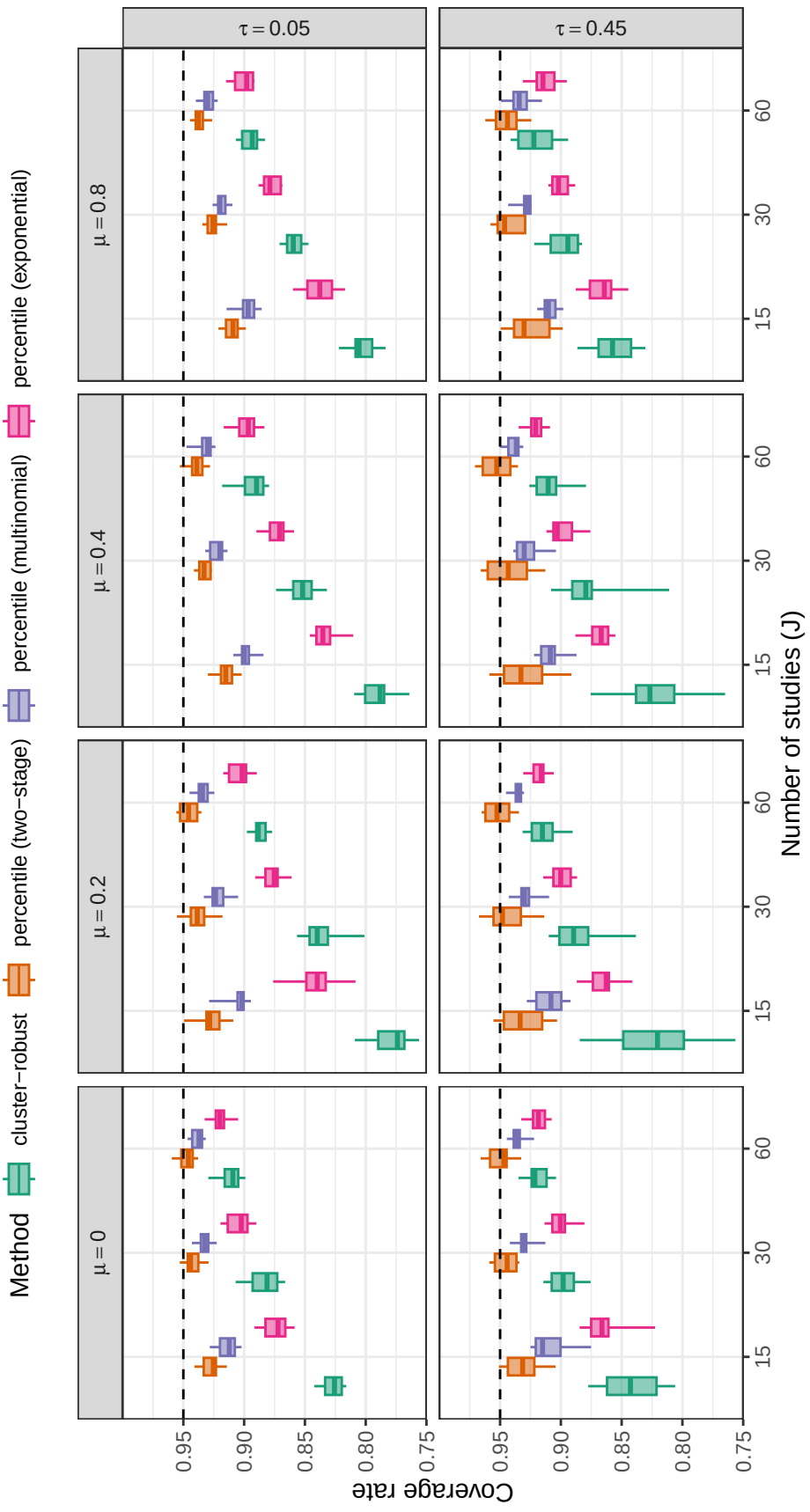

Figure 5
Coverage levels of confidence intervals based on the CML estimator of average effect size by number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95.

E4 and E5 providing greater detail about the relative accuracy of the three methods. Under nearly all conditions, the CML estimator is much more accurate than the ARGL estimator. The RMSE of CML is smaller than that of CHE under some but not all conditions; the former performs better under conditions where selection is strong, average SMD is low, between-study heterogeneity is large, and sample size is large. However, just as with the estimators of average effect size, the CHE estimator is more accurate when selection is mild or absent, leading to a bias-variance trade-off.

5.3 Selection Parameter

The CML and ARGL methods both estimate the log-selection parameter of the step-function model, a parameter that may be of substantive interest. Figure F1 shows that the bias of the two estimators is similar: Both show bias close to zero under conditions with low average SMD and high between-study heterogeneity but have large negative biases under conditions with strong selection probability and high average SMD, especially with low between-study heterogeneity. Figure F2 depicts the RMSE for the ARGL and CML estimators of the log-selection parameter with Figure F3 providing a more detailed view of the relative accuracy of the ARGL versus the CML estimators. The CML estimator of the selection parameter generally outperforms the ARGL estimator, especially under conditions with strong selection. Figures F4 and F5 depict the confidence interval coverage of the two estimators with various bootstrap confidence interval estimation approaches, using two-stage clustered bootstrap resampling. Broadly, none of the methods provide coverage rates that are close to the nominal 95% level across all conditions examined.

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 (Stanley & Doucouliagos, 2014) or the endogenous kink meta-regression (Bom & Rachinger, 2019) are agnostic as to selection mechanisms and thus are only indirectly informative about the strength or form of selection. The Vevea and Hedges (1995) 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). In contrast, well-known methods such as trim-and-fill (Duval & Tweedie, 2000a, 2000b) and more recent proposals such as *p*-curve (Simonsohn et al., 2014) and *p*-uniform (Aert et al., 2016; Assen et al., 2015) are not as flexible and have been found to perform poorly when effects are heterogeneous (Carter et al., 2019).

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 an broad range of conditions, we found that both estimators of the marginal step-function model

showed little bias overall and consistently out-performed PET/PEESE and CHE-ISCW under conditions with meaningful selective reporting. However, the selection model estimators face a bias-variance trade-off: the reduced bias that they provide comes at the expense of increased variance from using a marginal model. 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 marginal step-function model estimators have better 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 was usually more accurate and had confidence interval coverage rates closer to nominal levels, although differences are fairly small. CML consistently out-performed 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 (Chen & Pustejovsky, 2024) indicating that univariate selection models perform well relative to alternative regression-based models to adjust for selective reporting bias. 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 to summary meta-analysis and meta-regression analysis, which emphasize use of models that align with the hierarchical structure of dependent effect sizes (Pustejovsky & Tipton, 2022; Van den Noortgate et al., 2013).

Second, focusing on the marginal distribution likely entails some loss of precision in parameter estimates. Accounting for the dependence structure would allow for construction

of more efficient estimators of the parameters of the evidence-generating process. A useful direction for further research would be to explore how to refine the CML and ARGL estimators by incorporating analytic weights connected with the dependence structure of the effect sizes.

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. This strikes us as an area in need of further theoretical development—even simply to catalog a wider variety of plausible selective reporting mechanisms. However, developing more nuanced models would require shifting to estimation frameworks that can handle multivariate models, such as multivariate likelihood-based estimation or pairwise composite likelihood methods (e.g., Rao et al., 2013; Yi et al., 2016).

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 (3,840 conditions for the full simulation and 288 conditions for the bootstrap simulation), 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 (Hedges & Vevea, 1996).

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 $\alpha_1 = .025$. Chen and Pustejovsky (2024) examined the performance of one-step and two-step selection model estimators under conditions when the true data-generating process involved a two-step selection process with an additional threshold at $\alpha_2 = .500$. Their results indicated that a one-step model could be more accurate (i.e., lower RMSE) than a more complex two-step model—even if the former is mis-specified. 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. There may be other factors besides statistical significance of findings that 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 (Van Aert, 2025). 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 (Coburn & Vevea, 2015).

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 (Pustejovsky & Rodgers, 2019). 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

sensitivity analysis approach proposed by Mathur and VanderWeele (2020) by using estimates of the strength of selection to inform assumptions about the maximal plausible degree of selection.

Consistent with recommendations from past work in the context of independent effect sizes (Carter et al., 2019; McShane et al., 2016), 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, it remains critical that meta-analysts 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, when applying and interpreting step function models. 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

Aert, R. C. M. van, Wicherts, J. M., & Assen, M. A. L. M. van. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p -uniform and p -curve. *Perspectives on Psychological Science*, *11*(5), 713–729. https://doi.org/10.1177/1745691616650874

Assen, M. A. L. M. van, Van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293–309. https://doi.org/http://dx.doi.org/10.1037/met0000025

Becker, B. J. (2000). Multivariate Meta-analysis. In S. D. Brown & H. E. A. Tinsley (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499–525). Academic Press. https://doi.org/10.1016/B978-012691360-6/50018-5

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

Bom, P. R. D., & Rachinger, H. (2019). A kinked meta-regression model for publication bias correction. *Research Synthesis Methods*, *10*(4), 497–514. https://doi.org/10.1002/jrsm.1352

Boos, D. D., & Zhang, J. (2000). Monte carlo evaluation of resampling-based hypothesis

tests. *Journal of the American Statistical Association*, *95*(450), 486–492. https://doi.org/10.1080/01621459.2000.10474226

Borenstein, M., & Hedges, L. V. (2019). Effect sizes for meta-analysis. *The Handbook of Research Synthesis and Meta-Analysis*, *3*, 207–243.

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*(4), 796.

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

Center for High Throughput Computing. (2006). *Center for high throughput computing*. Center for High Throughput Computing. https://doi.org/10.21231/GNT1-HW21

Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Jama*, *291*(20), 2457–2465.

Chen, M., & Pustejovsky, J. E. (2024). *Adapting methods for correcting selective reporting bias in meta-analysis of dependent effect sizes*. https://doi.org/10.31222/osf.io/jq52s

Citkowicz, M., & Vevea, J. L. (2017). A parsimonious weight function for modeling publication bias. *Psychological Methods*, *22*(1), 28–41. https://doi.org/10.1037/met0000119

Coburn, K. M., & Vevea, J. L. (2015). Publication bias as a function of study characteristics. *Psychological Methods*, *20*(3), 310–330. https://doi.org/10.1037/met0000046

Copas, J. B. (1999). What works?: Selectivity models and meta-analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *162*(1), 95–109. https://doi.org/10.1111/1467-985X.00123

Copas, J. B., & Li, H. G. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *59*(1), 55–95.

https://doi.org/10.1111/1467-9868.00055

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

Cox, D. R., & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, *91*(3), 729–737. https://doi.org/10.1093/biomet/91.3.729

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge University Press.

Dear, K. B. G., & Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*, *7*(2), 237–245.

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*(449), 89–98.

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

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. https://doi.org/10.1080/01621459.1987.10478410

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634.

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2019). Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *The Journal of Experimental Education*, 1–20. https://doi.org/10.1080/00220973.2019.1582470

Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *69*(3), 369–390. https://doi.org/10.1111/j.1467-9868.2007.00593.x

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

Science, *7*(1), 8–12. https://doi.org/10.1177/1948550615598377

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20. https://doi.org/10.1037/h0076157

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, *136*(4), 495–525. https://doi.org/10.1037/a0019486

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

Hasselman, B. (2023). *Nleqslv: Solve systems of nonlinear equations*. https://CRAN.R-project.org/package=nleqslv

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*(1), 61–85.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, *7*(2), 246–255.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Hedges, L. V., & Vevea, J. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 145–174). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470870168.ch9

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, *21*(4), 299. https://doi.org/10.3102/10769986021004299

Iyengar, S., & Greenhouse, J. B. (1988). Selection Models and the File Drawer Problem.

Statistical Science, *3*(1), 109–117. https://doi.org/10.1214/ss/1177013012

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Joshi, M., & Pustejovsky, J. E. (2024). *Simhelpers: Helper functions for simulation studies*. https://meghapsimatrix.github.io/simhelpers/

Joshi, M., Pustejovsky, J. E., & Beretvas, S. N. (2022). Cluster Wild Bootstrapping to Handle Dependent Effect Sizes in Meta-Analysis with a Small Number of Studies. *Research Synthesis Methods*, *13*(4), 457–477. https://doi.org/10.1002/jrsm.1554

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

Kossmeier, M., Tran, U. S., & Voracek, M. (2020). Power-enhanced funnel plots for meta-analysis. *Zeitschrift Für Psychologie*. https://econtent.hogrefe.com/doi/10.1027/2151-2604/a000392

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253.

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

Leeden, R. V. D., Meijer, E., & Busing, F. M. T. A. (2008). Resampling multilevel models. In J. D. Leeuw & E. Meijer (Eds.), *Handbook of Multilevel Analysis* (pp. 401–433). Springer New York. https://doi.org/10.1007/978-0-387-73186-5_11

Light, R. J., & Pillemer, D. B. (1984). *Summing Up*. Harvard University Press. https://books.google.com?id=qel3lAm4K6gC

Lindsay, B. G. (1988). Composite likelihood methods. In N. U. Prabhu (Ed.), *Contemporary Mathematics* (Vol. 80, pp. 221–239). American Mathematical Society. https://doi.org/10.1090/conm/080/999014

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, *20*(4), 641–654. https://doi.org/10.1002/sim.698

Marks-Anglin, A., & Chen, Y. (2020). A historical review of publication bias. *Research Synthesis Methods*, *11*(6), 725–742. https://doi.org/10.1002/jrsm.1452

Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(5), 1091–1119. https://doi.org/10.1111/rssc.12440

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis an evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*(5), 730–749. http://pps.sagepub.com/content/11/5/730.short

Moreno, S. G., Sutton, A. J., Thompson, J. R., Ades, A., Abrams, K. R., & Cooper, N. J. (2012). A generalized weighting regression-derived meta-analysis estimator robust to small-study effects and heterogeneity. *Statistics in Medicine*, *31*(14), 1407–1417.

Nakagawa, S., Yang, Y., Macartney, E. L., Spake, R., & Lagisz, M. (2023). Quantitative evidence synthesis: A practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. *Environmental Evidence*, *12*(1), 8. https://doi.org/10.1186/s13750-023-00301-6

Nash, J. C., & Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software*, *43*(9), 1–14. https://doi.org/10.18637/jss.v043.i09

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

O'Boyle Jr, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, *43*(2),

376–399.

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

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, *42*(8), 424–432.

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

Pustejovsky, J. E. (2024). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections*. https://CRAN.R-project.org/package=clubSandwich

Pustejovsky, J. E., Joshi, M., & Citkowicz, M. (2025). *Metaselection: Meta-analytic selection models with cluster-robust and cluster-bootstrap standard errors for dependent effect size estimates*. https://github.com/jepusto/metaselection

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, *10*(1), 57–71.

Pustejovsky, J. E., & Tipton, E. (2022). Meta-Analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Prevention Science*, *23*, 425–438. https://doi.org/10.1016/j.jsp.2018.02.003

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rao, J. N. K., Verret, F., & Hidiroglou, M. A. (2013). A weighted composite likelihood approach to inference for two-level models from survey data. *Survey Methodology*, *39*(2), 263–282.

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, *26*(2), 141.

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

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

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment, and Adjustments* (pp. 1–7). John Wiley & Sons. https://doi.org/10.1002/0470870168

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, *9*(1). https://doi.org/10.1214/aos/1176345338

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681. https://doi.org/10.1177/1745691614553988

Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and Statistics*, *70*(1), 103–127.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78.

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment, and Adjustments* (pp. 73–98). John Wiley & Sons. https://doi.org/10.1002/0470870168

Sutton, A. (2009). Publication bias. In *The handbook of research synthesis and meta-analysis* (pp. 435–445). Russell Sage Foundation.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the

presence of heterogeneity. *Statistics in Medicine*, *22*(13), 2113–2126. https://doi.org/10.1002/sim.1461

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

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375.

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

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, *10*(2), 180–194.

Van Aert, R. C. M. (2025). Meta-analyzing nonpreregistered and preregistered studies. *Psychological Methods*. https://doi.org/10.1037/met0000719

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

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

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, *92*(1), 1–28. https://doi.org/10.1007/s10182-008-0060-7

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. https://doi.org/10.1007/BF02294384

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428–443. https://doi.org/10.1037/1082-989X.10.4.428

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

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Xu, L., Gotwalt, C., Hong, Y., King, C. B., & Meeker, W. Q. (2020). Applications of the fractional-random-weight bootstrap. *The American Statistician*, *74*(4), 345–358. https://doi.org/10.1080/00031305.2020.1731599

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

Yi, G. Y., Rao, J. N. K., & Li, H. (2016). A weighted composite likelihood approach for analysis of survey data under two-level models. *Statistica Sinica*, *26*(2), 569–587. https://doi.org/10.5705/ss.2013.383

Appendix A

Score and Hessian of the step-function marginal log likelihood

From Equation (9), the log of the marginal likelihood contribution for effect size estimate i from study j is given by

$$l_{ij}^{M}\left(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right) \propto \log w\left(y_{ij}, \sigma_{ij}; \boldsymbol{\zeta}\right) - \frac{1}{2}\frac{\left(y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta}\right)^2}{\exp(\gamma) + \sigma_{ij}^2} - \frac{1}{2}\log\left(\exp(\gamma) + \sigma_{ij}^2\right) - \log A\left(\mathbf{x}_{ij}, \sigma_{ij}; \boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right). \tag{A1}$$

The components of the score contribution of study j given in Equations (12), (13), and (14) involve derivatives of $l_{ij}^{M}\left(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right)$ with respect to all model parameters. Let $w_{ij} = w\left(y_{ij}, \sigma_{ij}; \boldsymbol{\lambda}\right)$, $\mu_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta}$, $\eta_{ij} = \exp(\gamma) + \sigma_{ij}^2$, and $A_{ij} = A\left(\mathbf{x}_{ij}, \sigma_{ij}; \boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}\right)$ as defined in Equation (7). The score contribution of study j for the meta-regression coefficients has the general form

$$\mathbf{S}_{\beta j} = \sum_{i=1}^{k_j} a_{ij}\mathbf{x}_{ij}'\left(\frac{\left(y_{ij} - \mu_{ij}\right)}{\eta_{ij}} - \frac{1}{A_{ij}}\frac{\partial A_{ij}}{\partial \mu_{ij}}\right). \tag{A2}$$

The score for the variance parameter has the general form

$$S_{\gamma j} = \sum_{i=1}^{k_j} a_{ij}\tau^2\left(\frac{\left(y_{ij} - \mu_{ij}\right)^2 - \eta_{ij}}{2\eta_{ij}^2} - \frac{1}{A_{ij}}\frac{\partial A_{ij}}{\partial \eta_{ij}}\right). \tag{A3}$$

The score for the h^{th} selection parameter has the general form

$$S_{\zeta_h j} = \sum_{i=1}^{k_j} a_{ij}\lambda_h\left(\frac{1}{w_{ij}}\frac{\partial w_{ij}}{\partial \lambda_h} - \frac{1}{A_{ij}}\frac{\partial A_{ij}}{\partial \lambda_h}\right), \tag{A4}$$

for $h = 1, ..., H$. For the step-function selection model, the derivatives of w_{ij} and A_{ij} are given by

$$\frac{\partial w_{ij}}{\partial \lambda_h} = I\left(\alpha_h < p_{ij} \leq \alpha_{h+1}\right) \tag{A5}$$

$$\frac{\partial A_{ij}}{\partial \mu_{ij}} = \frac{1}{\eta_{ij}^{1/2}} \sum_{h=0}^{H} \lambda_h \left[\phi\left(c_{h+1,ij}\right) - \phi\left(c_{hij}\right)\right] \tag{A6}$$

$$\frac{\partial A_{ij}}{\partial \eta_{ij}} = \frac{1}{2\eta_{ij}} \sum_{h=0}^{H} \lambda_h \left[c_{h+1,ij}\phi\left(c_{h+1,ij}\right) - c_{hij}\phi\left(c_{hij}\right)\right] \tag{A7}$$

$$\frac{\partial A_{ij}}{\partial \lambda_h} = B_{hij}, \tag{A8}$$

with c_{hij} and B_{hij} as defined in Equation (8).

The sub-components of the Hessian matrix from study j have the following forms:

$$\mathbf{H}_j^{\beta\beta'} = \frac{\partial}{\partial \boldsymbol{\beta}'}\mathbf{S}_{\beta j} = \sum_{i=1}^{k_j} a_{ij}\mathbf{x}_{ij}'\mathbf{x}_{ij}\left(\frac{1}{A_{ij}^2}\left(\frac{\partial A_{ij}}{\partial \mu_{ij}}\right)^2 - \frac{1}{A_{ij}}\frac{\partial^2 A_{ij}}{\partial \mu_{ij}^2} - \frac{1}{\eta_{ij}}\right) \tag{A9}$$

$$\mathbf{H}_j^{\beta\gamma} = \frac{\partial}{\partial \gamma}\mathbf{S}_{\beta j} = \sum_{i=1}^{k_j} a_{ij}\mathbf{x}_{ij}'\tau^2\left(\frac{1}{A_{ij}^2}\frac{\partial A_{ij}}{\partial \mu_{ij}}\frac{\partial A_{ij}}{\partial \eta_{ij}} - \frac{1}{A_{ij}}\frac{\partial^2 A_{ij}}{\partial \mu_{ij}\partial \eta_{ij}} - \frac{y_{ij} - \mu_{ij}}{\eta_{ij}^2}\right) \tag{A10}$$

$$\mathbf{H}_j^{\beta\zeta_h} = \frac{\partial}{\partial \zeta_h}\mathbf{S}_{\beta j} = \sum_{i=1}^{k_j} a_{ij}\mathbf{x}_{ij}'\lambda_h\left(\frac{1}{A_{ij}^2}\frac{\partial A_{ij}}{\partial \mu_{ij}}\frac{\partial A_{ij}}{\partial \lambda_h} - \frac{1}{A_{ij}}\frac{\partial^2 A_{ij}}{\partial \mu_{ij}\partial \lambda_h}\right) \tag{A11}$$

$$\begin{aligned}\mathbf{H}_j^{\gamma\gamma} = \frac{\partial}{\partial \gamma}S_{\gamma j} = \sum_{i=1}^{k_j} a_{ij}\Bigg[&\tau^2\left(\frac{\left(y_{ij} - \mu_{ij}\right)^2 - \eta_{ij}}{2\eta_{ij}^2} - \frac{1}{A_{ij}}\frac{\partial A_{ij}}{\partial \eta_{ij}}\right) \\ &+ \tau^4\left(\frac{1}{A_{ij}^2}\left(\frac{\partial A_{ij}}{\partial \eta_{ij}}\right)^2 - \frac{1}{A_{ij}}\frac{\partial^2 A_{ij}}{\partial \eta_{ij}^2} - \frac{2(y_{ij} - \mu_{ij})^2 - \eta_{ij}}{2\eta_{ij}^3}\right)\Bigg]\end{aligned} \tag{A12}$$

$$\mathbf{H}_j^{\gamma\zeta_h} = \frac{\partial}{\partial \zeta_h}S_{\gamma j} = \sum_{i=1}^{k_j} a_{ij}\tau^2\lambda_h\left(\frac{1}{A_{ij}^2}\frac{\partial A_{ij}}{\partial \eta_{ij}}\frac{\partial A_{ij}}{\partial \lambda_h} - \frac{1}{A_{ij}}\frac{\partial^2 A_{ij}}{\partial \eta_{ij}\partial \lambda_h}\right) \tag{A13}$$

$$\begin{aligned}\mathbf{H}_j^{\zeta_f\zeta_h} = \frac{\partial}{\partial \zeta_h}S_{\zeta_f j} = \sum_{i=1}^{k_j} a_{ij}\Bigg[&\lambda_f\lambda_h\left(\frac{1}{A_{ij}^2}\frac{\partial A_{ij}}{\partial \lambda_f}\frac{\partial A_{ij}}{\partial \lambda_h} - \frac{1}{w_{ij}^2}\frac{\partial w_{ij}}{\partial \lambda_f}\frac{\partial w_{ij}}{\partial \lambda_h}\right) \\ &+ I(f = h)\lambda_h\left(\frac{1}{w_{ij}}\frac{\partial w_{ij}}{\partial \lambda_h} - \frac{1}{A_{ij}}\frac{\partial A_{ij}}{\partial \lambda_h}\right)\Bigg],\end{aligned} \tag{A14}$$

where the second partial derivatives of A_{ij} are given by

$$\frac{\partial^2 A_{ij}}{\partial \mu_{ij}^2} = \frac{1}{\eta_{ij}} \sum_{h=0}^{H} \lambda_h \left[c_{h+1,ij} \phi(c_{h+1,ij}) - c_{hij} \phi(c_{hij}) \right] \tag{A15}$$

$$\frac{\partial^2 A_{ij}}{\partial \mu_{ij} \partial \eta_{ij}} = \frac{1}{2\eta_{ij}^{3/2}} \sum_{h=0}^{H} \lambda_h \left[(c_{h+1,ij}^2 - 1) \phi(c_{h+1,ij}) - (c_{hij}^2 - 1) \phi(c_{hij}) \right] \tag{A16}$$

$$\frac{\partial^2 A_{ij}}{\partial \mu_{ij} \partial \lambda_h} = \frac{\phi\left(c_{h+1,ij}\right) - \phi\left(c_{hij}\right)}{\eta_{ij}^{1/2}} \tag{A17}$$

$$\frac{\partial^2 A_{ij}}{\partial \eta_{ij}^2} = \frac{1}{4\eta_{ij}^2} \sum_{h=0}^{H} \lambda_h \left[(c_{h+1,ij}^3 - 3c_{h+1,ij}) \phi(c_{h+1,ij}) - (c_{hij}^3 - 3c_{hij}) \phi(c_{hij}) \right] \tag{A18}$$

$$\frac{\partial^2 A_{ij}}{\partial \eta_{ij} \partial \lambda_h} = \frac{c_{h+1,ij} \phi\left(c_{h+1,ij}\right) - c_{hij} \phi\left(c_{hij}\right)}{2\eta_{hij}}. \tag{A19}$$

Appendix B

Sandwich estimator for the augmented, re-weighted Gaussian likelihood estimator

The augmented, re-weighted Gaussian likelihood (ARGL) estimators are defined as the solution to the estimating equations given in Equation (21) and (22). Sandwich variance approximations for the ARGL estimators involve the estimating equations and their Jacobian.

Let $\mathbf{J}$ denote the Jacobian matrix of the ARGL estimating equations, given by

$$\mathbf{J}(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta}) = \sum_{j=1}^{J} \frac{\partial \mathbf{M}_j(\boldsymbol{\beta}, \gamma, \boldsymbol{\zeta})}{\partial(\boldsymbol{\beta}' \gamma \boldsymbol{\zeta}')}. \tag{B1}$$

The exact form of the Jacobian is described below. Let $\tilde{\mathbf{M}}_j = \mathbf{M}_j(\tilde{\boldsymbol{\beta}}, \tilde{\gamma}, \tilde{\boldsymbol{\zeta}})$ and $\tilde{\mathbf{J}} = \mathbf{J}(\tilde{\boldsymbol{\beta}}, \tilde{\gamma}, \tilde{\boldsymbol{\zeta}})$ denote the score vectors and Jacobian matrix evaluated at the solutions to the ARGL estimating equations. A cluster-robust sandwich estimator is then given by:

$$\mathbf{V}^{ARGL} = \tilde{\mathbf{J}}^{-1} \left(\sum_{j=1}^{J} \tilde{\mathbf{M}}_j \tilde{\mathbf{M}}_j' \right) \left(\tilde{\mathbf{J}}^{-1} \right)'. \tag{B2}$$

The Jacobian of (22) is given by

$$\mathbf{J} = \sum_{j=1}^{J} \begin{bmatrix} \mathbf{J}_j^{\beta\beta'} & \mathbf{J}_j^{\beta\gamma'} & \mathbf{J}_j^{\beta\zeta'} \\ \mathbf{J}_j^{\gamma\beta'} & \mathbf{J}_j^{\gamma\gamma'} & \mathbf{J}_j^{\gamma\zeta'} \\ \mathbf{J}_j^{\zeta\beta'} & \mathbf{J}_j^{\zeta\gamma'} & \mathbf{J}_j^{\zeta\zeta'}, \end{bmatrix}$$

where

$$
\begin{aligned}
\mathbf{J}_j^{\beta\beta'} &= -\sum_{i=1}^{k_j} a_{ij} \frac{\mathbf{x}_{ij}'\mathbf{x}_{ij}}{w_{ij}\eta_{ij}} \\
\mathbf{J}_j^{\beta\gamma} &= -\sum_{i=1}^{k_j} a_{ij} \mathbf{x}_{ij}'\tau^2 \left(\frac{y_{ij}-\mu_{ij}}{w_{ij}\eta_{ij}^2}\right) \\
\mathbf{J}_j^{\beta\zeta_h} &= -\sum_{i=1}^{k_j} a_{ij} \mathbf{x}_{ij}'\lambda_h \left(\frac{y_{ij}-\mu_{ij}}{w_{ij}^2\eta_{ij}}\right) \\
\mathbf{J}_j^{\gamma\beta'} &= \left(\mathbf{J}_j^{\beta\gamma}\right)' \\
\mathbf{J}_j^{\gamma\gamma} &= -\sum_{i=1}^{k_j} a_{ij} \left[\frac{\tau^4}{w_{ij}}\left(\frac{2(y_{ij}-\mu_{ij})^2-\eta_{ij}}{2\eta_{ij}^3}\right) - \frac{\tau^2}{w_{ij}}\left(\frac{(y_{ij}-\mu_{ij})^2-\eta}{2\eta_{ij}^2}\right)\right] \\
\mathbf{J}_j^{\gamma\zeta_h} &= -\sum_{i=1}^{k_j} a_{ij} \frac{\tau^2\lambda_h}{w_{ij}^2}\left(\frac{(y_{ij}-\mu_{ij})^2-\eta_{ij}}{2\eta_{ij}^2}\right) \\
\mathbf{J}_j^{\zeta_h\beta'} &= \left(\mathbf{H}_j^{\beta\zeta_h}\right)' \\
\mathbf{J}_j^{\zeta_h\gamma} &= \mathbf{H}_j^{\gamma\zeta_h} \\
\mathbf{J}_j^{\zeta_f\zeta_h} &= \mathbf{H}_j^{\zeta_f\zeta_h'},
\end{aligned}
$$

with $\mathbf{H}_j^{\beta\zeta_h}$, $\mathbf{H}_j^{\gamma\zeta_h}$, and $\mathbf{H}_j^{\zeta_f\zeta_h}$ as given in Equations (A11), (A13), and (A14).

Appendix C

Further details about bootstrapping

Other bootstrap sampling methods

The main manuscript described a two-stage cluster bootstrapping technique. We also considered two other bootstrap sampling schemes.

In the clustered bootstrap scheme, each pseudo-sample is generated by randomly drawing J clusters of observations with replacement from the original sample. In contrast to the two-stage cluster bootstrap, sampled clusters are left intact rather than perturbed. Because clusters are drawn with replacement, some will necessarily be included multiple times and some clusters might not be included in a given pseudo-sample. Following the notation in the main text, 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 clustered bootstrap process is then equivalent to drawing $a_1^{(b)}, ..., a_J^{(b)}$ from a multinomial distribution with J trials and equal probability on each of J categories and then setting $a_{ij}^{(b)} = a_j^{(b)}$ for $i = 1, ..., k_j$ and $j = 1, ..., J$.

The fractional random weight bootstrap follows a very similar process in which each pseudo-sample is generated by assigning a random weight to every cluster. Rather than using a multinomial distribution, the weights follow independent exponential distributions with mean 1, so that the weights for pseudo-sample b are generated as $a_j^{(b)} \sim Exp(1)$, with $a_{ij}^{(b)} = a_j^{(b)}$ for $i = 1, ..., k_j$ and $j = 1, ..., J$. A crucial difference between these bootstrap techniques is that the fractional random weight bootstrap puts strictly positive weight on every cluster of observations in every pseudo-sample, whereas the clustered bootstrap and two-stage bootstrap can assign zero weight to some clusters (Xu et al., 2020). If the existence of an estimator hinges on the inclusion of one or a small number of clusters, this will create computational problems for the non-parametric bootstrap but not for the

fractional random weight bootstrap.

Bootstrap confidence interval construction

We describe three methods for constructing a $1-2\alpha$ confidence interval (CI) from a set of B bootstrap replications. Consider a parameter θ that is a scalar component of $\boldsymbol{\beta}$, γ, or $\boldsymbol{\zeta}$. Let $\hat{\theta}$ denote an estimator of θ with sandwich variance estimator V. Let $\hat{\theta}_b^*$ denote the same estimator computed from bootstrap pseudo-sample b, with corresponding sandwich variance estimator V_b^*. Let $\hat{\theta}_{(1)}^*, ..., \hat{\theta}_{(B)}^*$ denote the pseudo-sample estimators sorted in ascending order. An estimator for the standard error of $\hat{\theta}$ can be computed by taking the standard deviation of the $\hat{\theta}_1^*, ..., \hat{\theta}_B^*$.

First, the percentile CI is calculated by taking the α and $1-\alpha$ quantiles of the bootstrap distribution, with end-points

$$\left[\hat{\theta}_{((B+1)\alpha)}^*, \hat{\theta}_{((B+1)(1-\alpha))}^*\right].$$

Second, the so-called "basic" CI pivots the bootstrap distribution around the original estimator $\hat{\theta}$. Its end-points are given by

$$\left[2\hat{\theta} - \hat{\theta}_{((B+1)(1-\alpha)}^*, 2\hat{\theta} - \hat{\theta}_{((B+1)\alpha)}^*\right].$$

Third, a studentized CI uses the bootstrap distribution of t-statistics rather than point estimators. The t statistic for pseudo-sample b is computed as $t_b^* = \left(\hat{\theta}_b^* - \hat{\theta}\right) / \sqrt{V_b^*}$. The studentized CI is computed using the percentiles of the bootstrap distribution of $t_1^*, ..., t_B^*$, taking

$$\left[\hat{\theta} - \sqrt{V} \times t_{((B+1)(1-\alpha)}^*,\ \hat{\theta} - \sqrt{V} \times t_{((B+1)\alpha)}^*\right].$$

Fourth, the bias-corrected-and-accelerated (BCa) CI is similar to the percentile CI in that its end-points are defined by quantiles of the bootstrap distribution. However, instead using the α and $1-\alpha$ quantiles, it uses quantiles that are adjusted to take into account the bias of the

estimator and the degree to which its sampling variance is related to the underlying parameter, as measured using an acceleration coefficient. These adjustments are defined in terms of the empirical influence function, which we approximate using a leave-one-cluster-out jackknife. The jackknife influence value for cluster j is $\hat{\theta} - \hat{\theta}^{+}_{-j}$, where $\hat{\theta}^{+}_{-j}$ denotes the estimator of θ computed while leaving out the observations in cluster j for $j = 1, ..., J$. The acceleration coefficient is then

$$\hat{a} = \frac{\sum_{j=1}^{J} \left(\hat{\theta} - \hat{\theta}^{+}_{-j}\right)^3}{6\left[\sum_{j=1}^{J} \left(\hat{\theta} - \hat{\theta}^{+}_{-j}\right)^2\right]^{3/2}}.$$

The bias coefficient is calculated as the proportion of the bootstrap distribution that falls below the original estimator:

$$\hat{\beta} = \frac{1}{B}\sum_{b=1}^{B} I\left(\hat{\theta}^{*}_{b} < \hat{\theta}\right).$$

With the acceleration and bias coefficients defined, define the adjustment function $f(\alpha)$ as

$$f(\alpha) = \Phi\left(\Phi^{-1}(\hat{\beta}) + \frac{\Phi^{-1}(\hat{\beta}) + \Phi^{-1}(\alpha)}{1 - \hat{a}\left[\Phi^{-1}(\hat{\beta}) + \Phi^{-1}(\alpha)\right]}\right)$$

for $0 < \alpha < 1$. The end-points of the BCa CI are then given by

$$\left[\hat{\theta}^{*}_{((B+1)\times f(\alpha))}, \hat{\theta}^{*}_{((B+1)\times f(1-\alpha))}\right].$$

Notably, the basic and studentized confidence intervals depend on the scale of the parameter θ, and the accuracy of their coverage levels therefore depends on the parameterization. In contrast, the percentile and bias-corrected-and-accelerated confidence intervals are invariant to transformation of θ.

Appendix D

Additional simulation results for estimators of average effect size (μ)

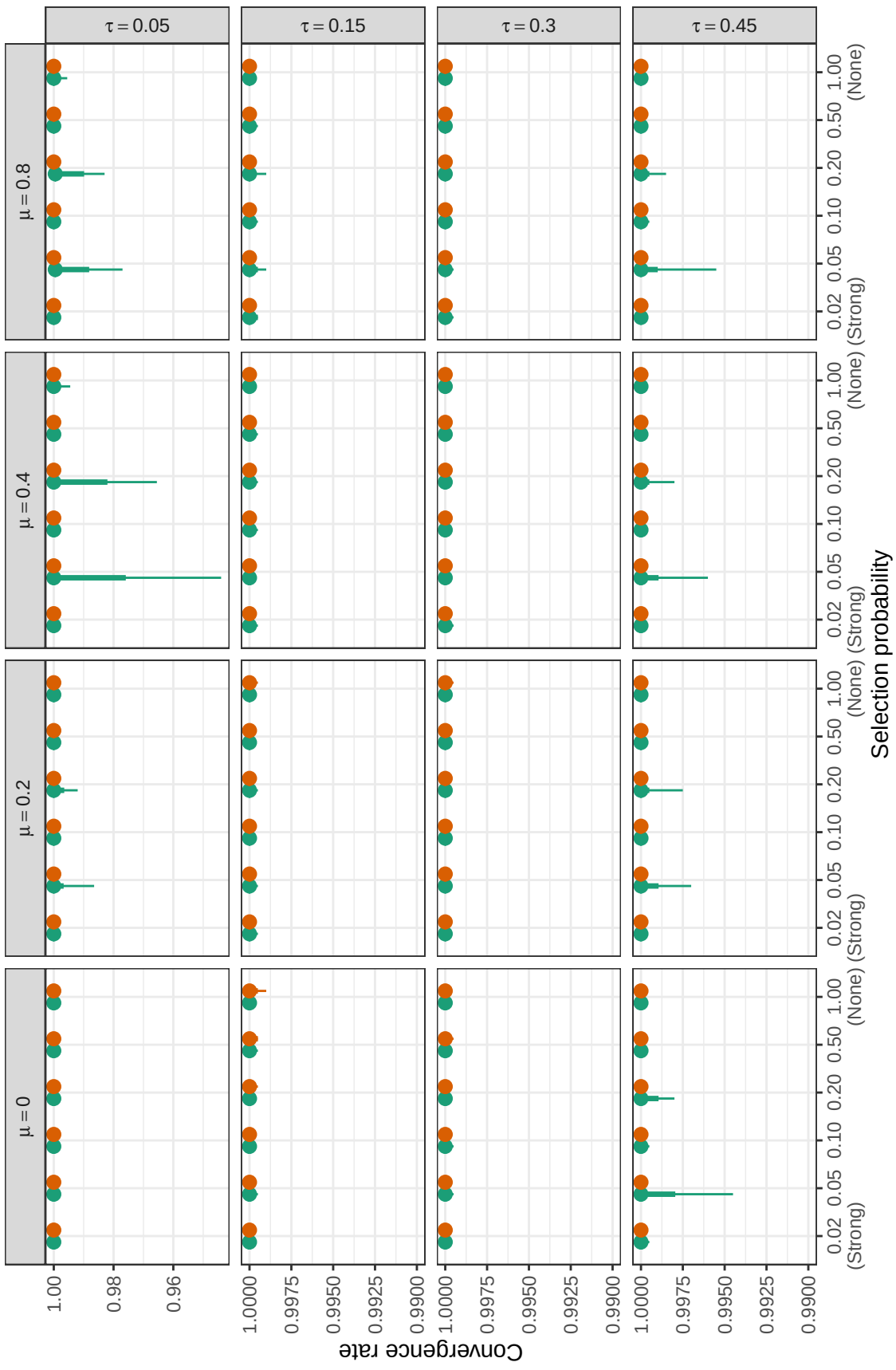

Figure D1

Convergence rates of CML and ARGL estimators by selection probability, average SMD, and between-study heterogeneity. Points correspond to median convergence rates; thin lines correspond to range of convergence rates; thick lines correspond to inter-decile range.

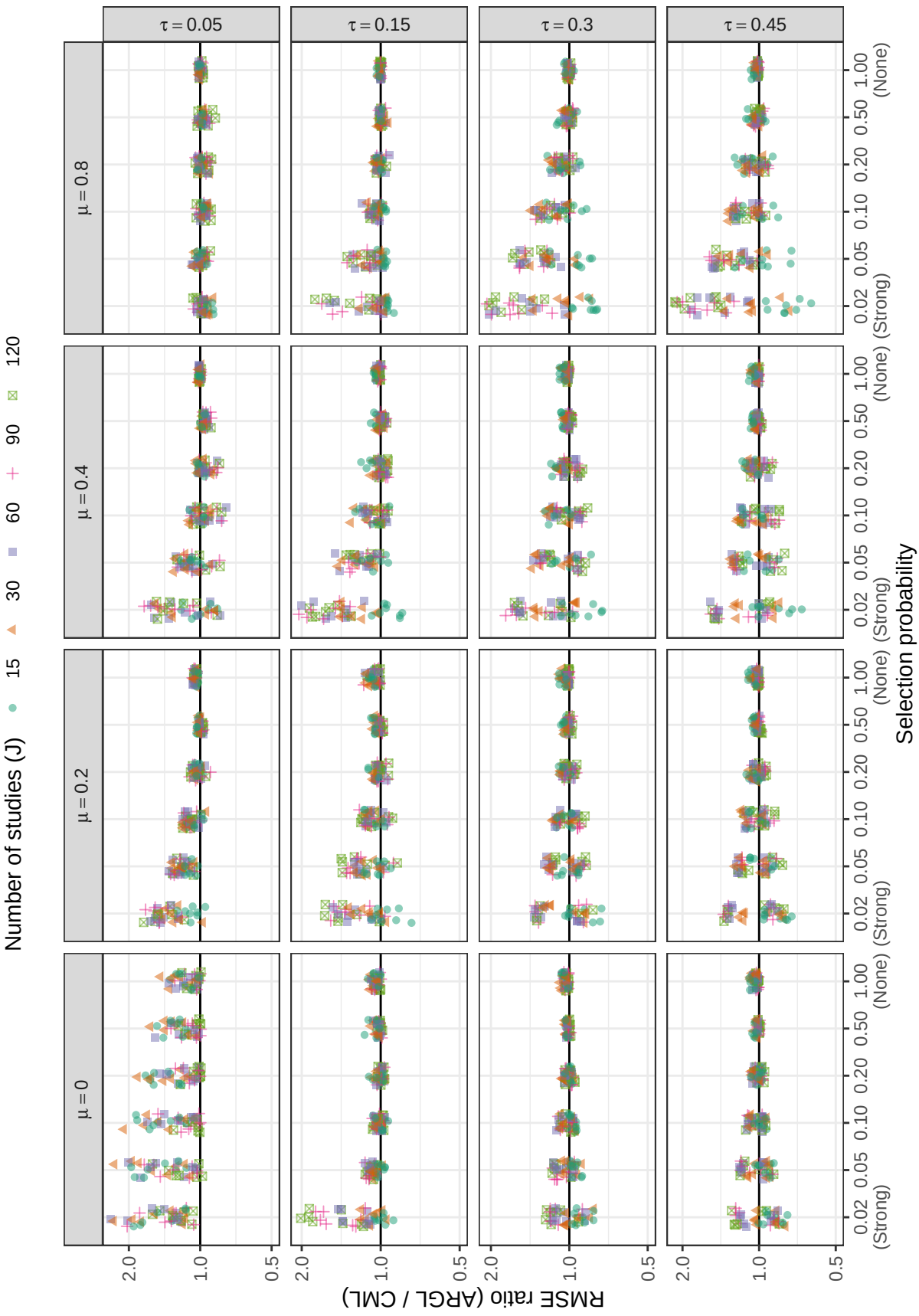

Figure D2

Ratio of root mean-squared error for ARGL estimator to root mean-squared error of CML estimator by selection probability, number of studies, average SMD, and between-study heterogeneity

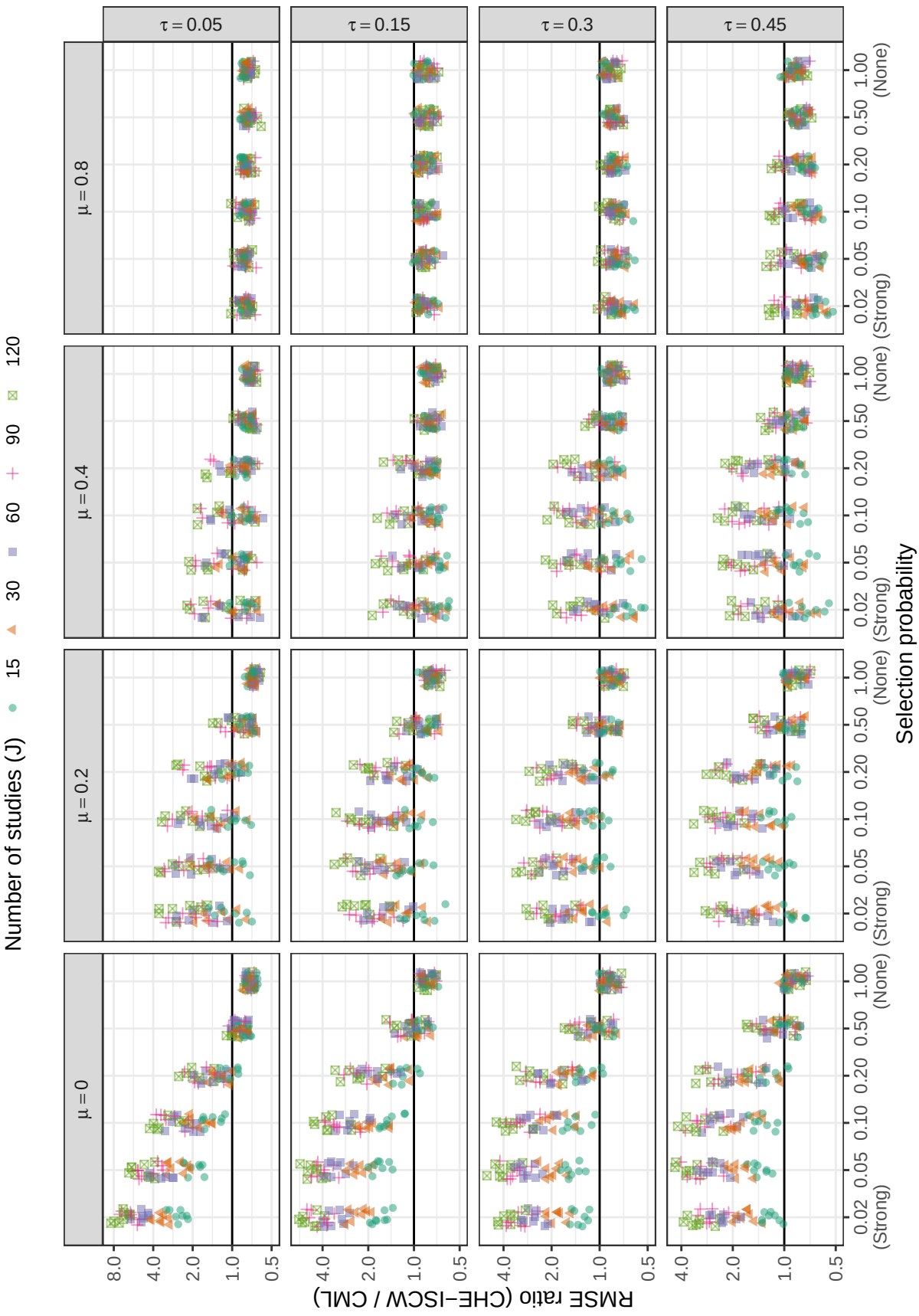

Figure D3

Ratio of root mean-squared error for CHE-ISCW estimator to root mean-squared error of CML estimator by selection probability, number of studies, average SMD, and between-study heterogeneity

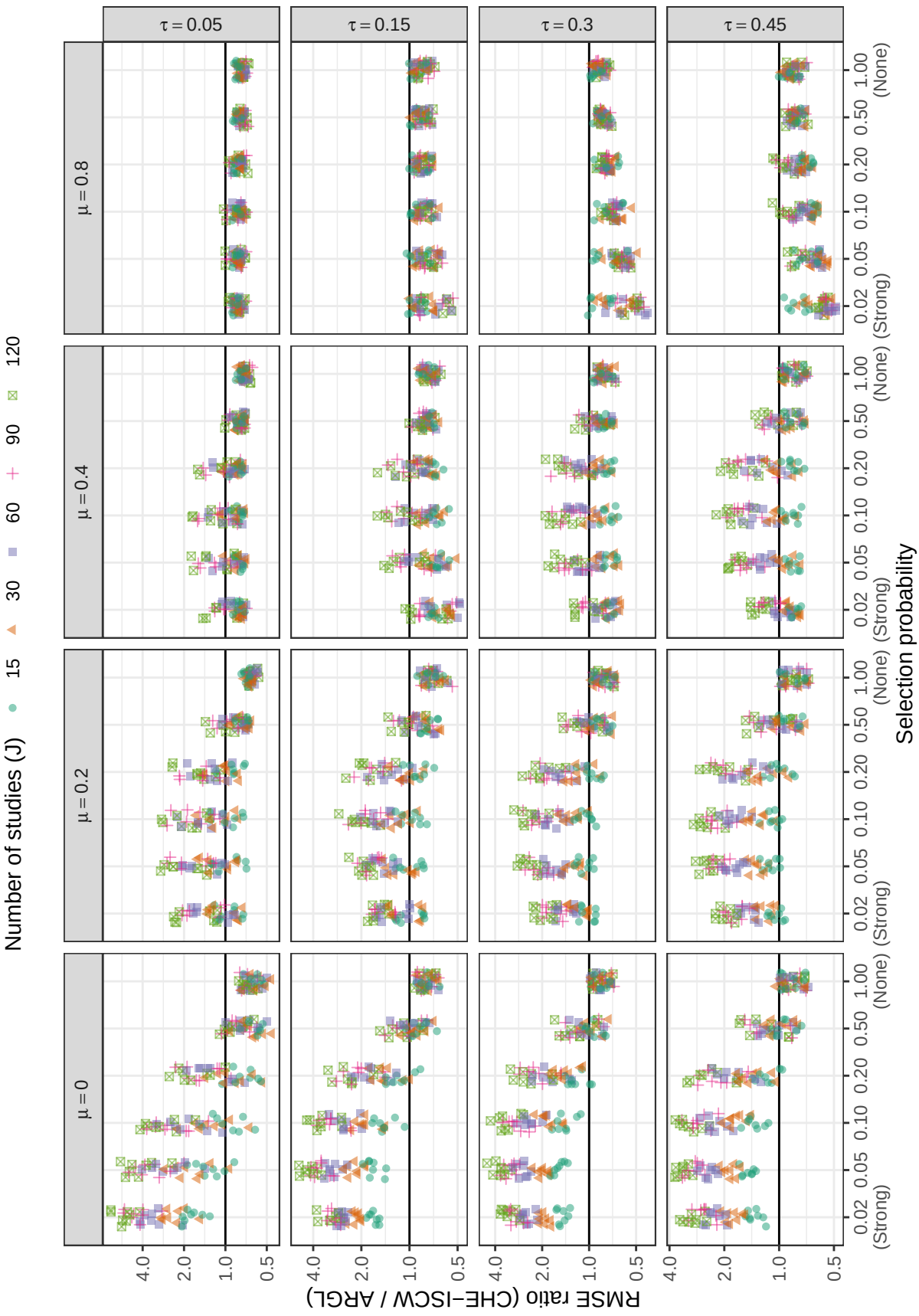

Figure D4

Ratio of root mean-squared error for CHE-ISCW estimator to root mean-squared error of ARGL estimator by selection probability, number of studies, average SMD, and between-study heterogeneity

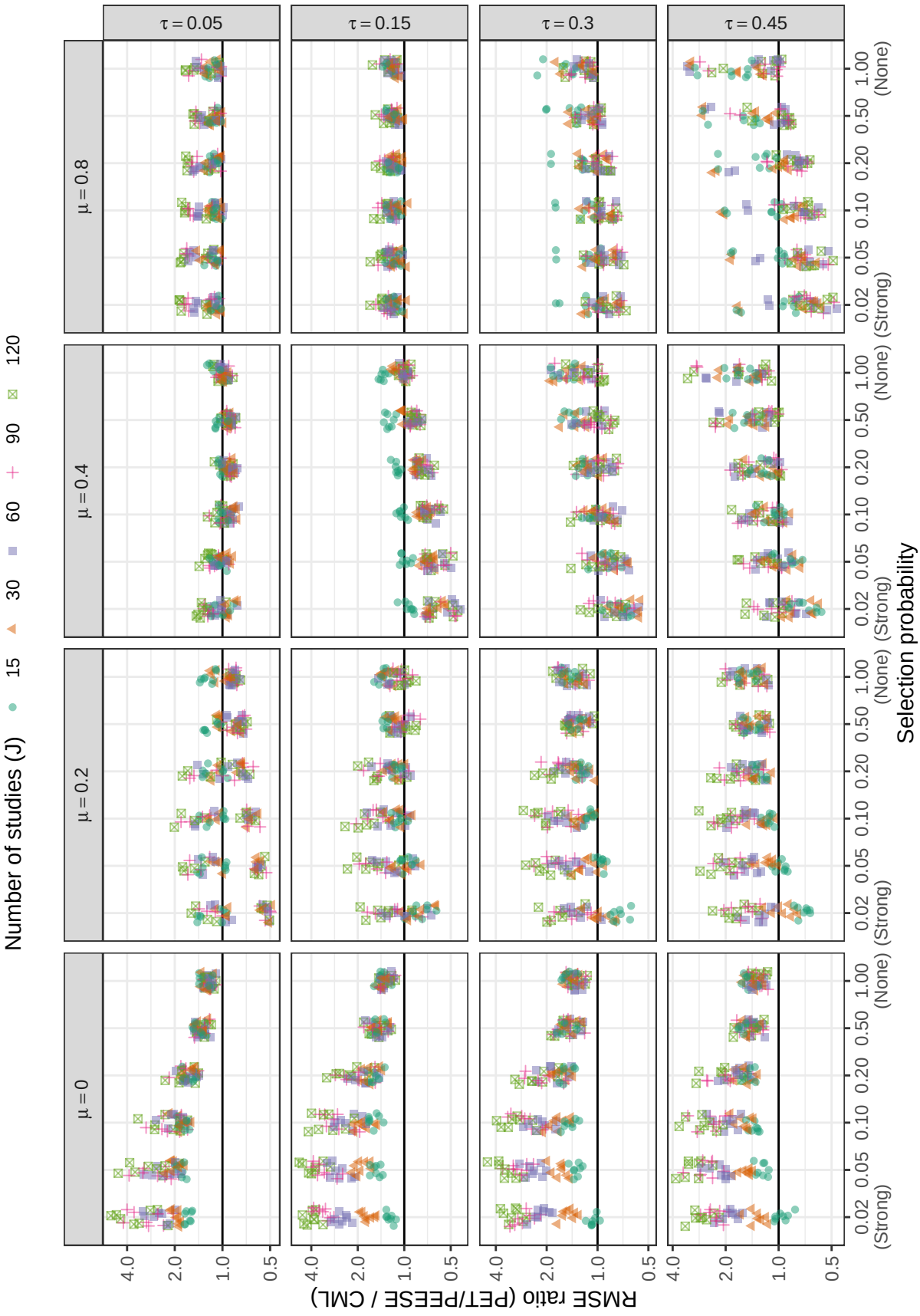

Figure D5

Ratio of root mean-squared error for PET/PEESE estimator to root mean-squared error of CML estimator by selection probability, number of studies, average SMD, and between-study heterogeneity

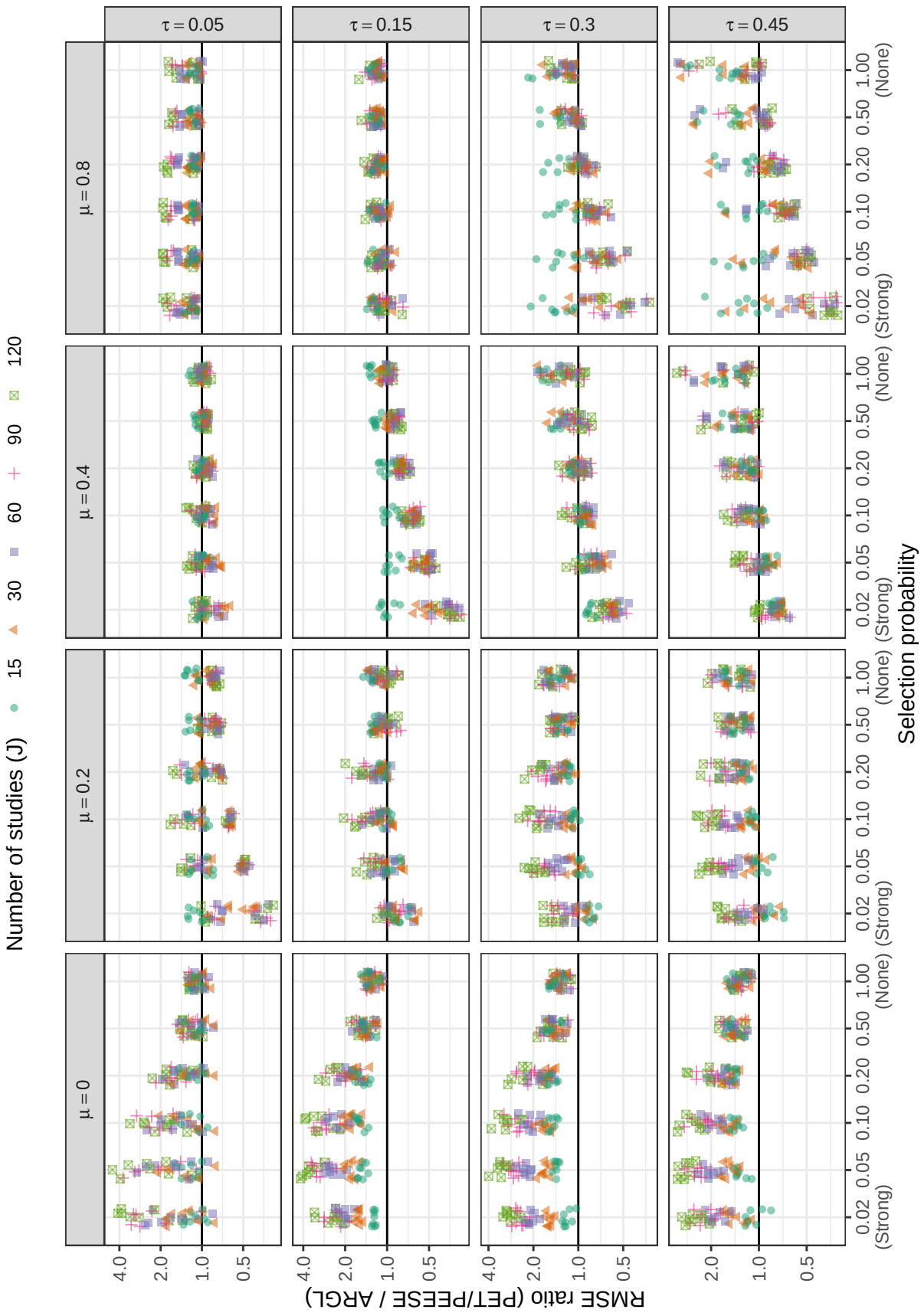

Figure D6

Ratio of root mean-squared error for PET/PEESE estimator to root mean-squared error of ARGL estimator by selection probability, number of studies, average SMD, and between-study heterogeneity

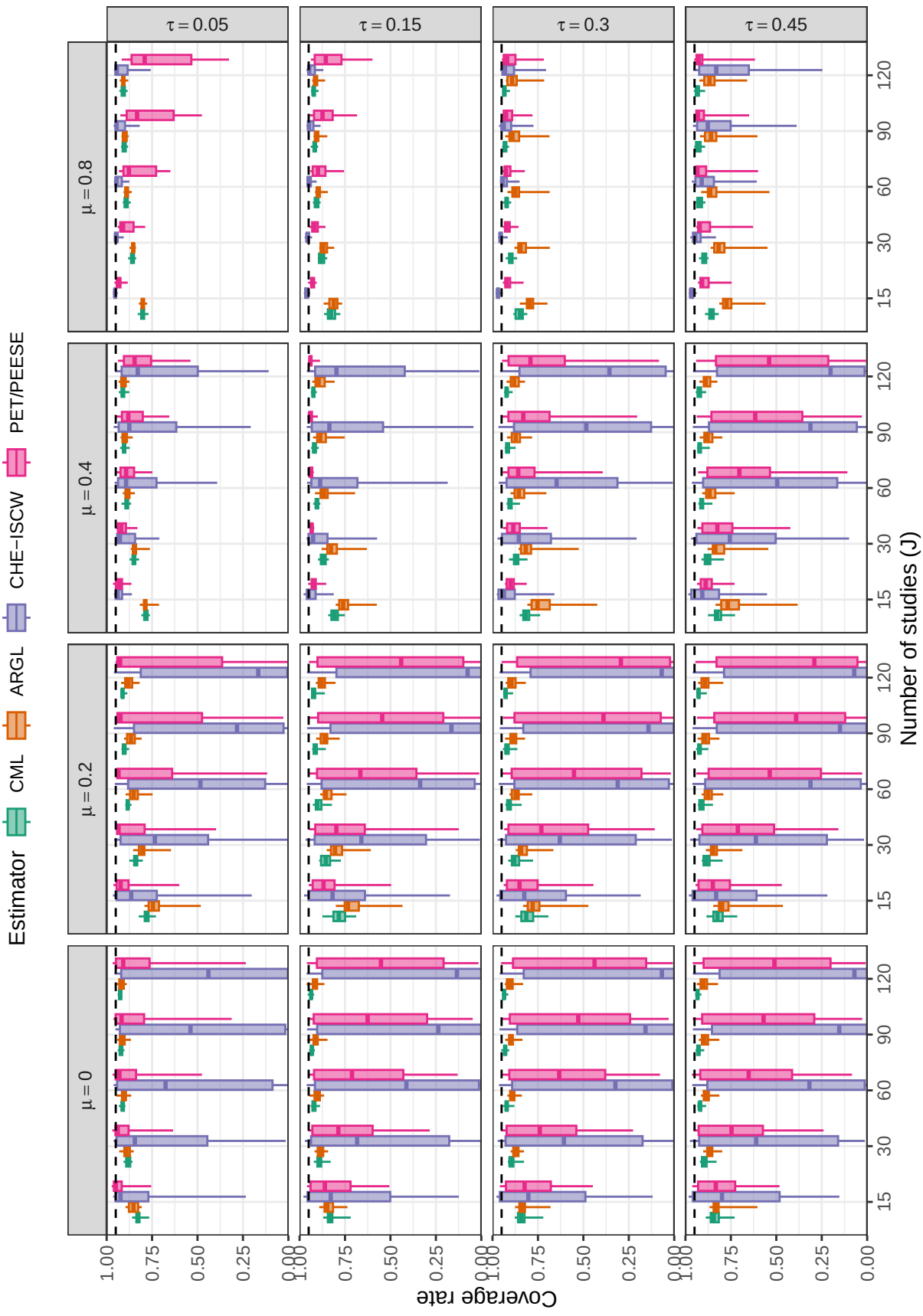

Figure D7

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.

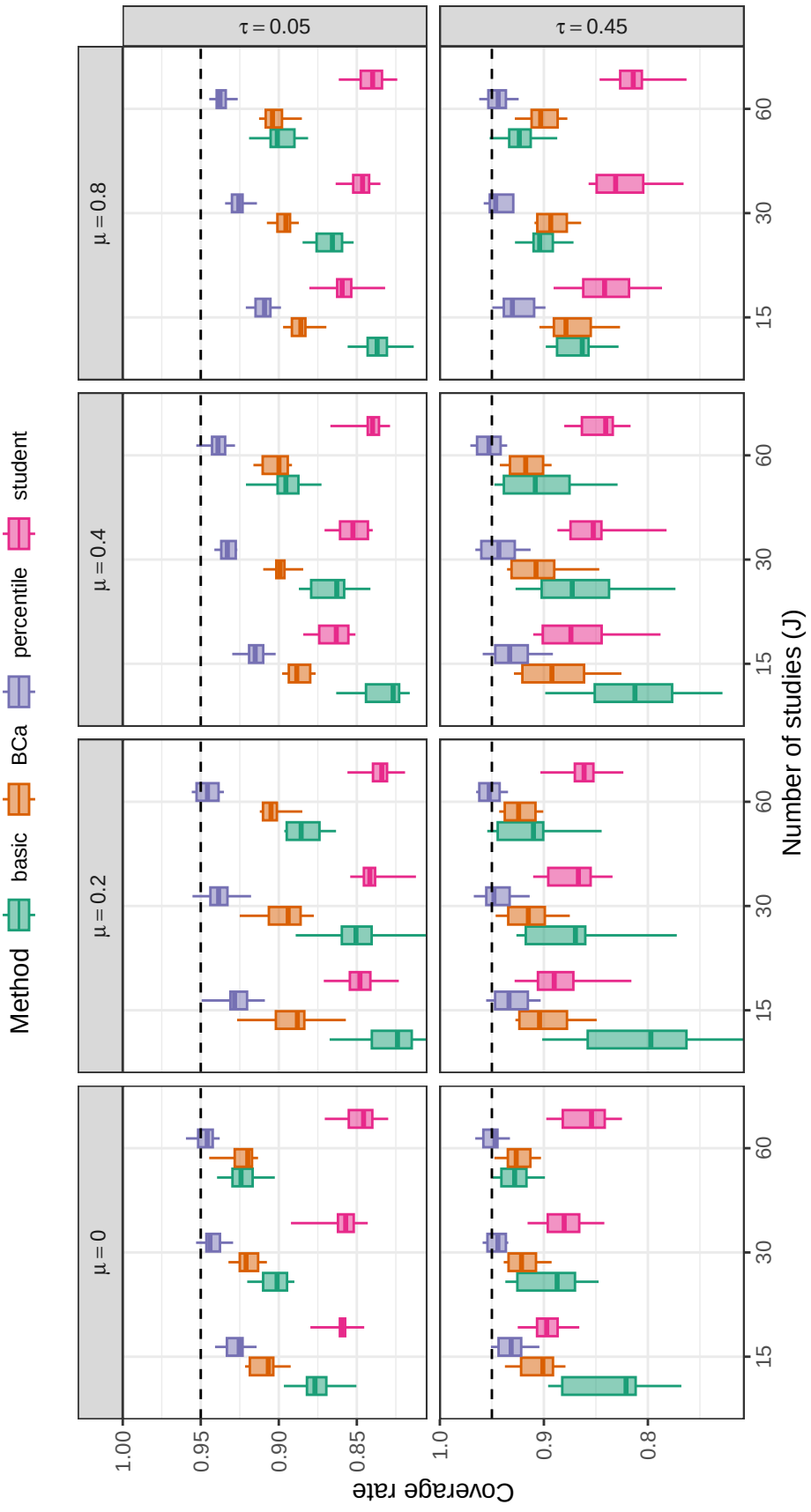

Figure D8

Coverage levels of two-stage bootstrap confidence intervals based on the CML estimator of average effect size by number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95.

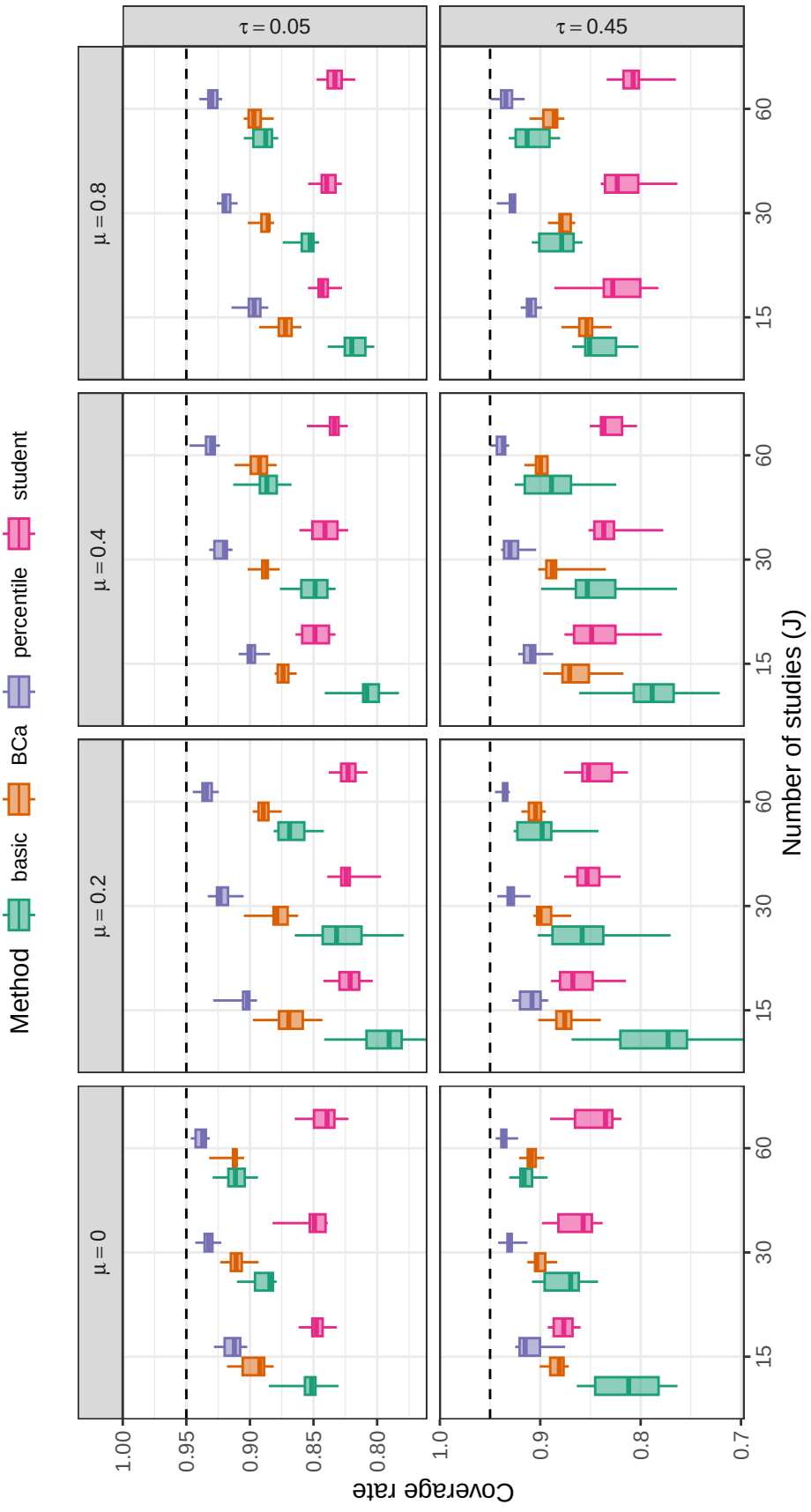

Figure D9

Coverage levels of multinomial bootstrap confidence intervals based on the CML estimator of average effect size by number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95.

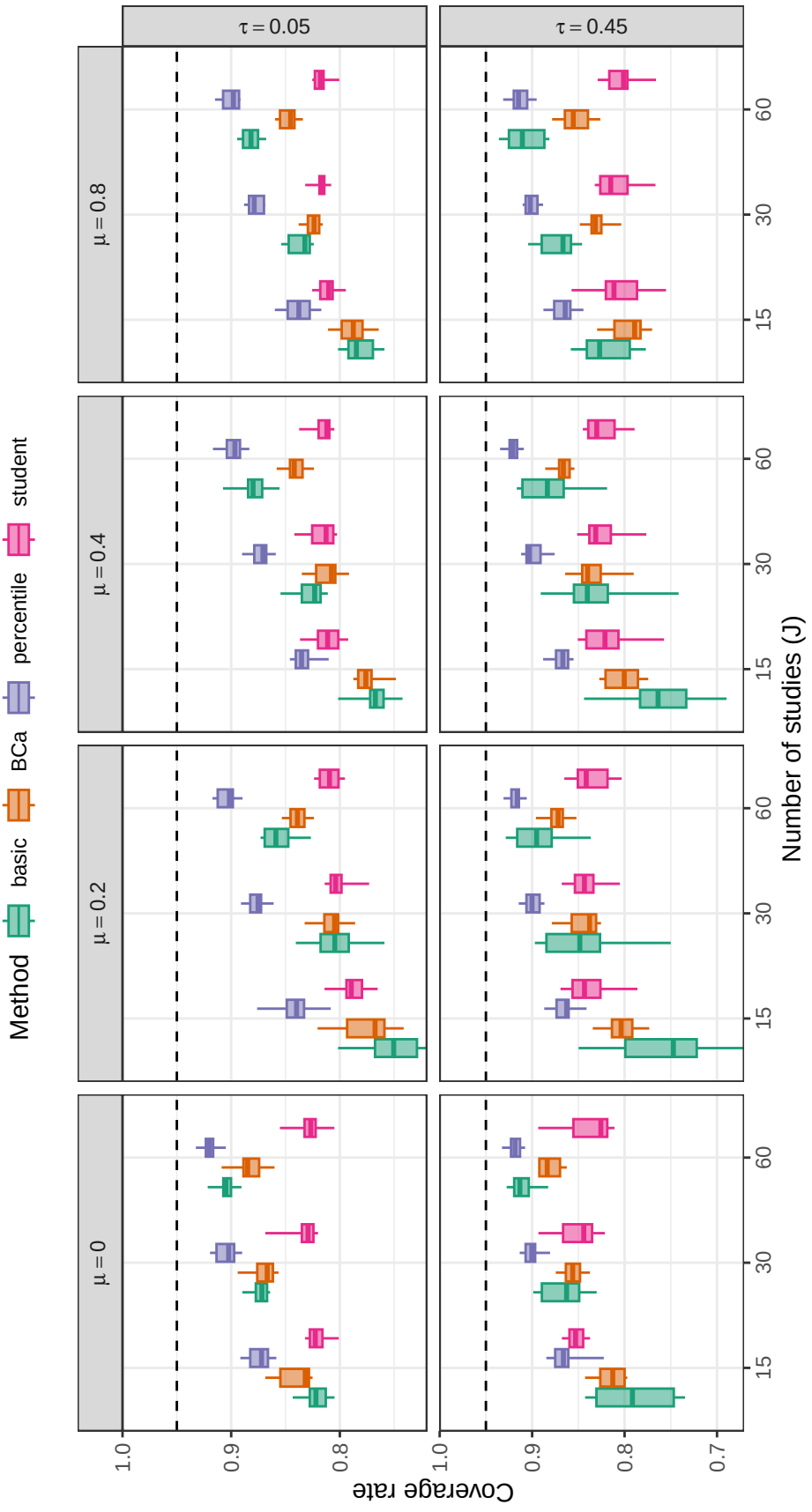

Figure D10
Coverage levels of fractional random weight bootstrap confidence intervals based on the CML estimator of average effect size by number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95.

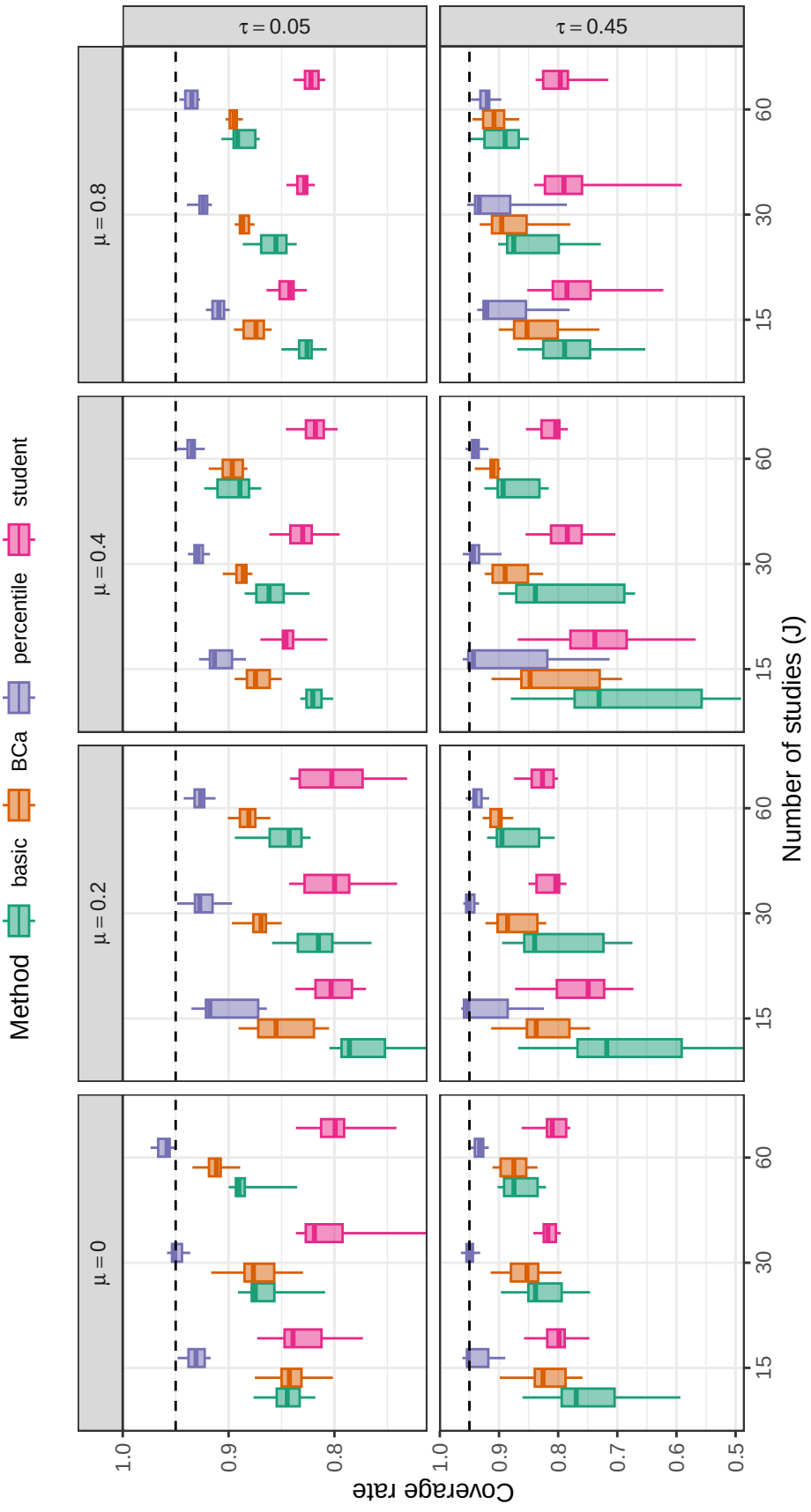

Figure D11
Coverage levels of two-stage bootstrap confidence intervals based on the ARGL estimator of average effect size by number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95.

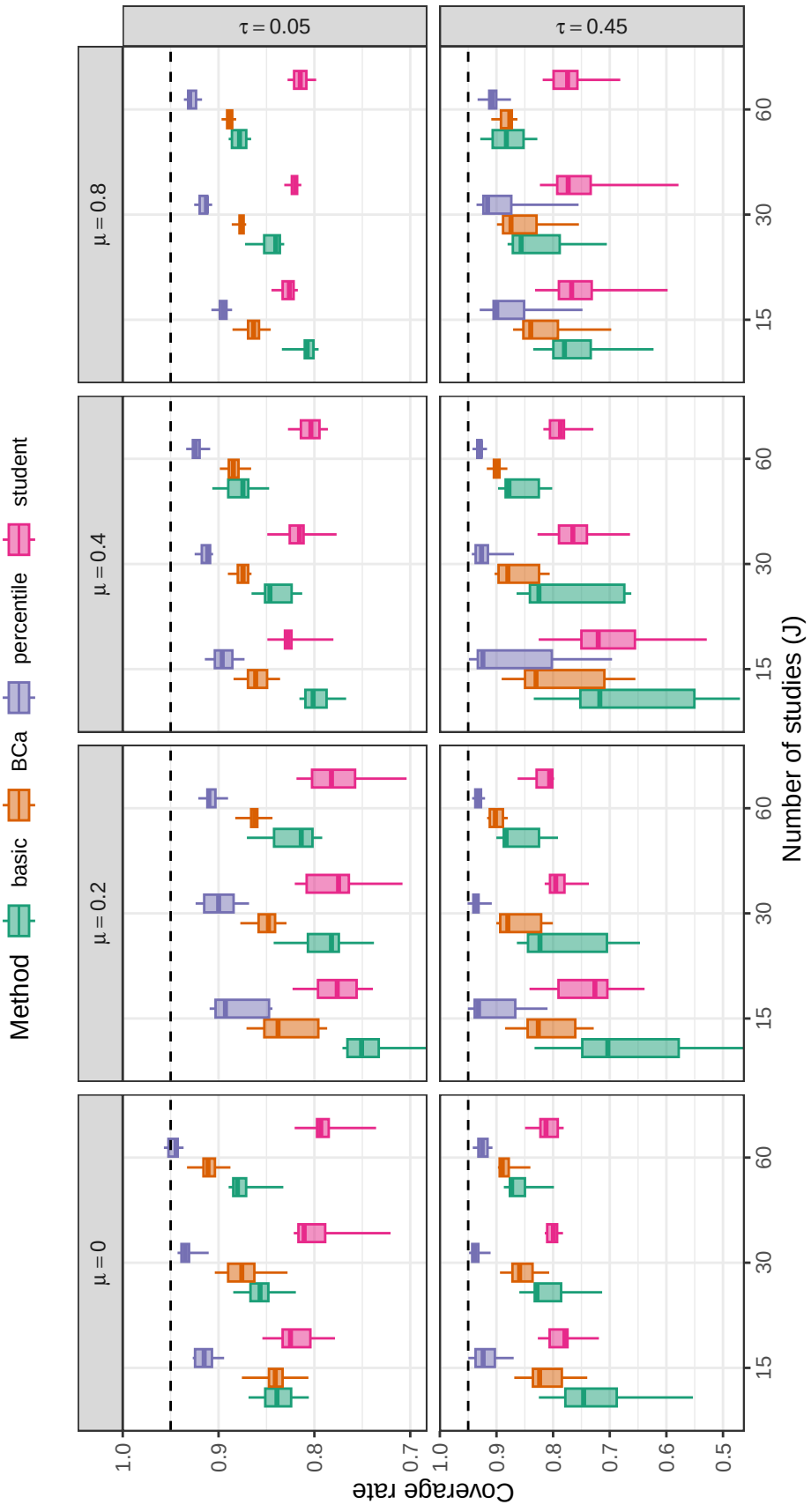

Figure D12
Coverage levels of multinomial bootstrap confidence intervals based on the ARGL estimator of average effect size by number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95.

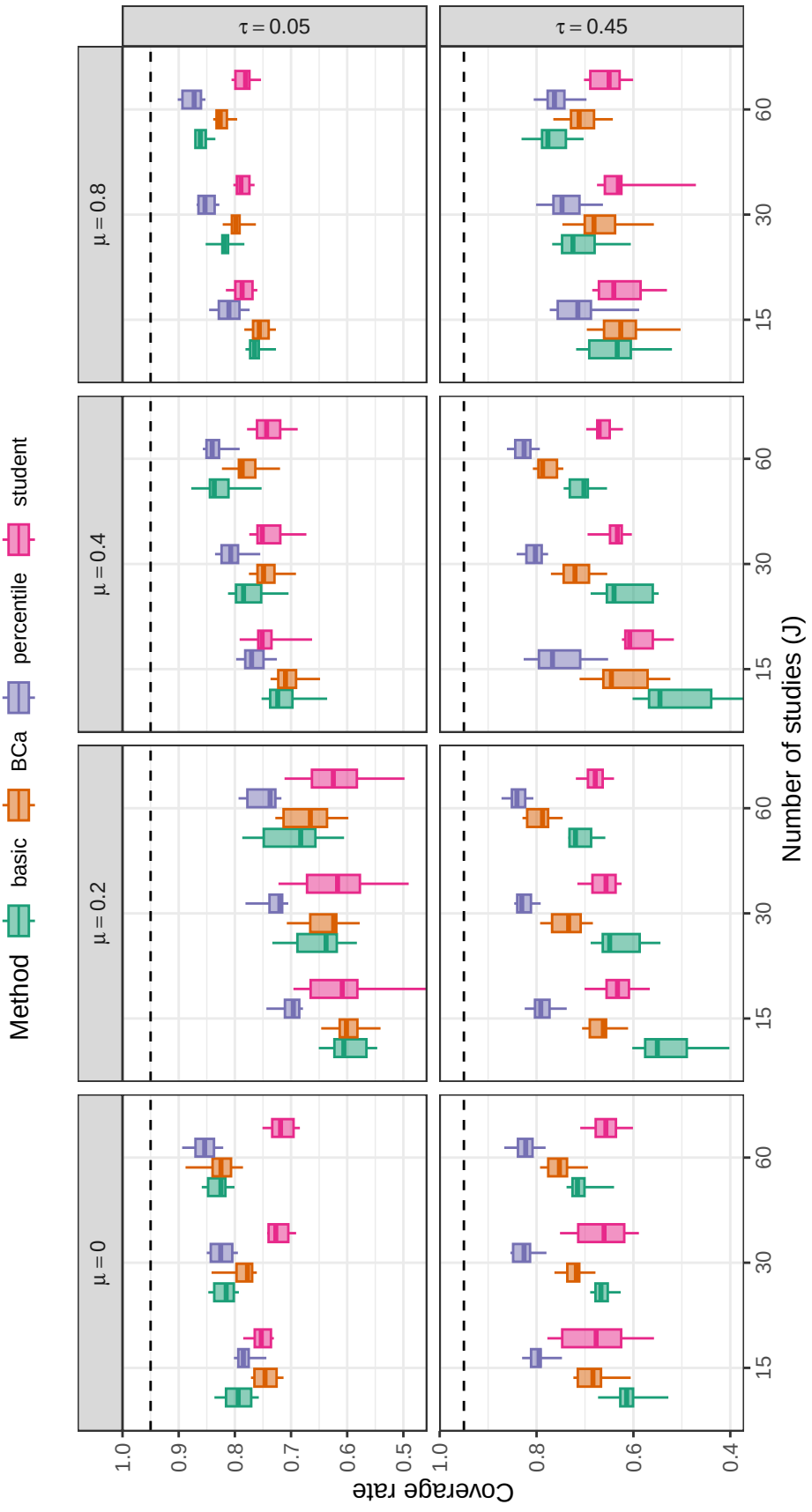

Figure D13
Coverage levels of fractional random weight bootstrap confidence intervals based on the ARGL estimator of average effect size by number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95.

Appendix E

Additional simulation results for estimators of log-heterogeneity (γ)

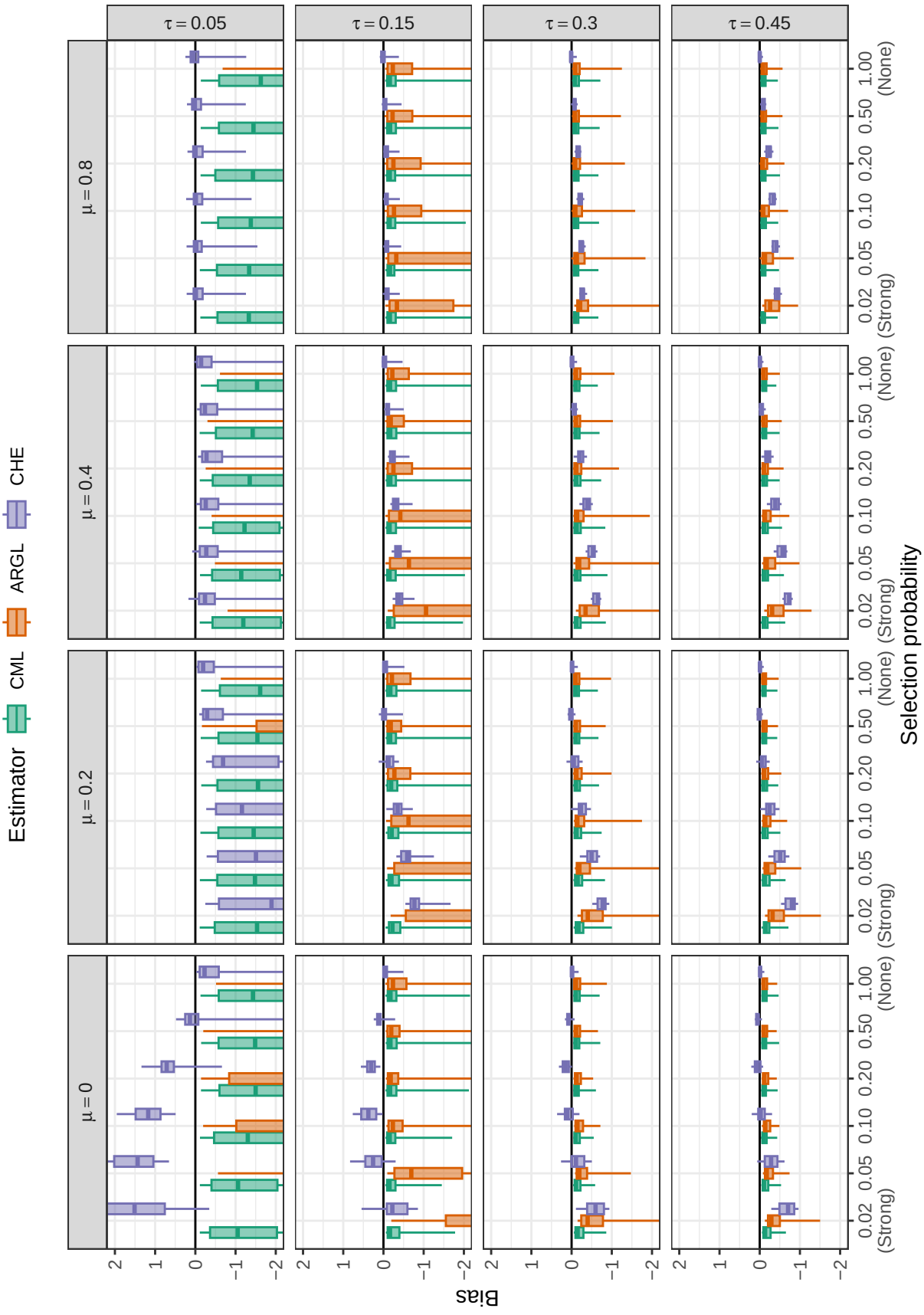

Figure E1

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

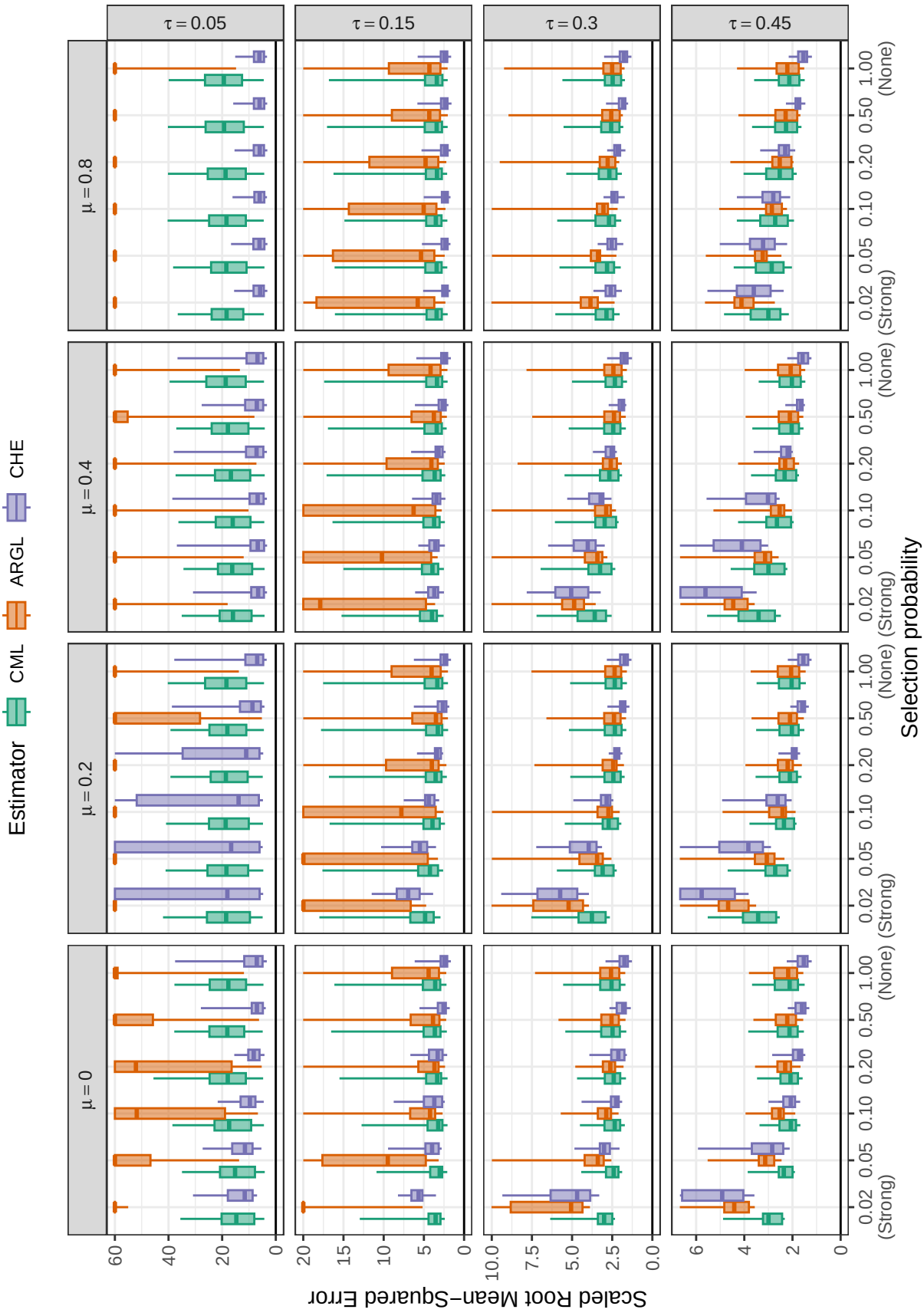

Figure E2

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

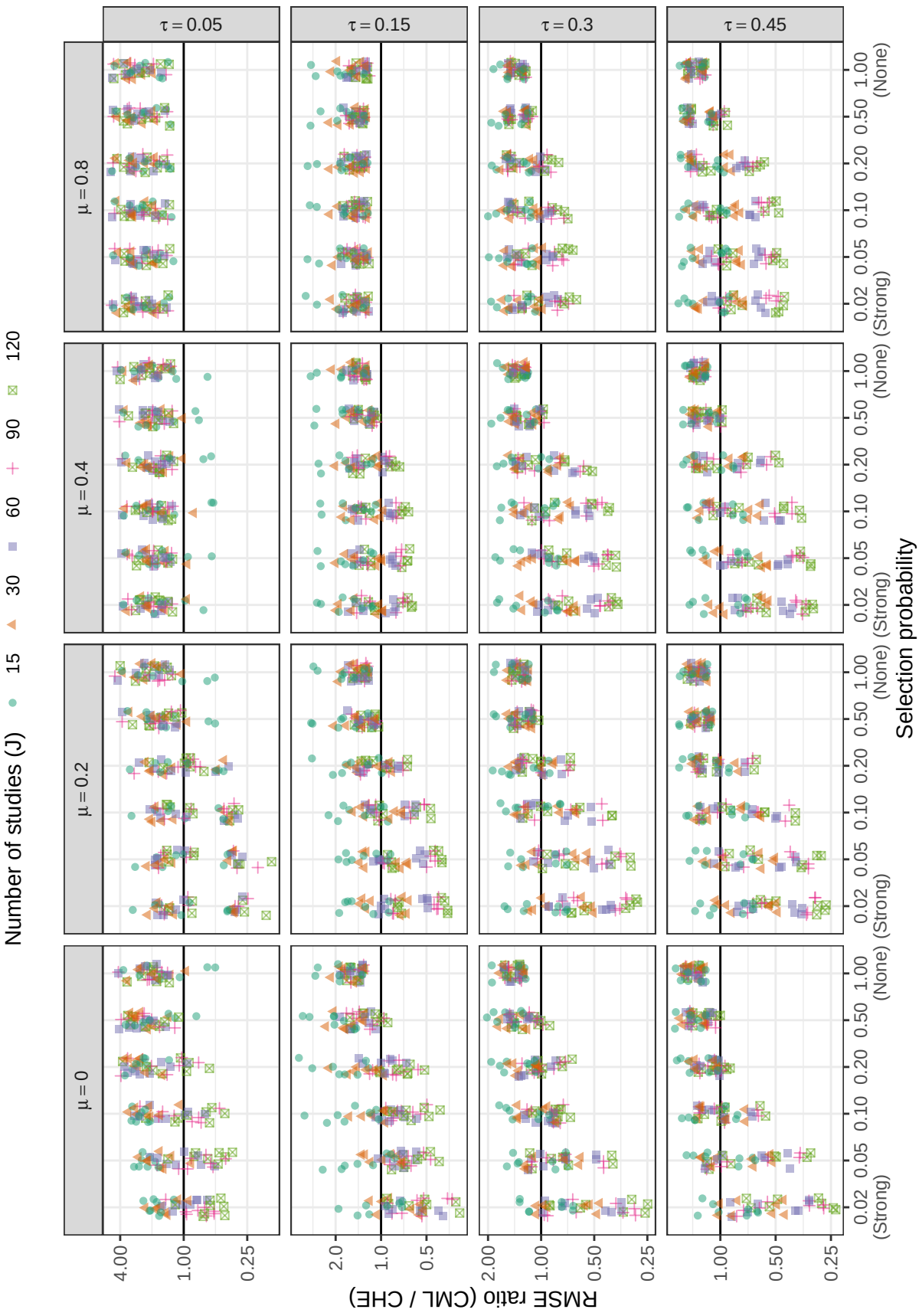

Figure E3

Ratio of root mean-squared error for CML heterogeneity estimator to root mean-squared error of CHE heterogeneity estimator by selection probability, number of studies, average SMD, and between-study heterogeneity

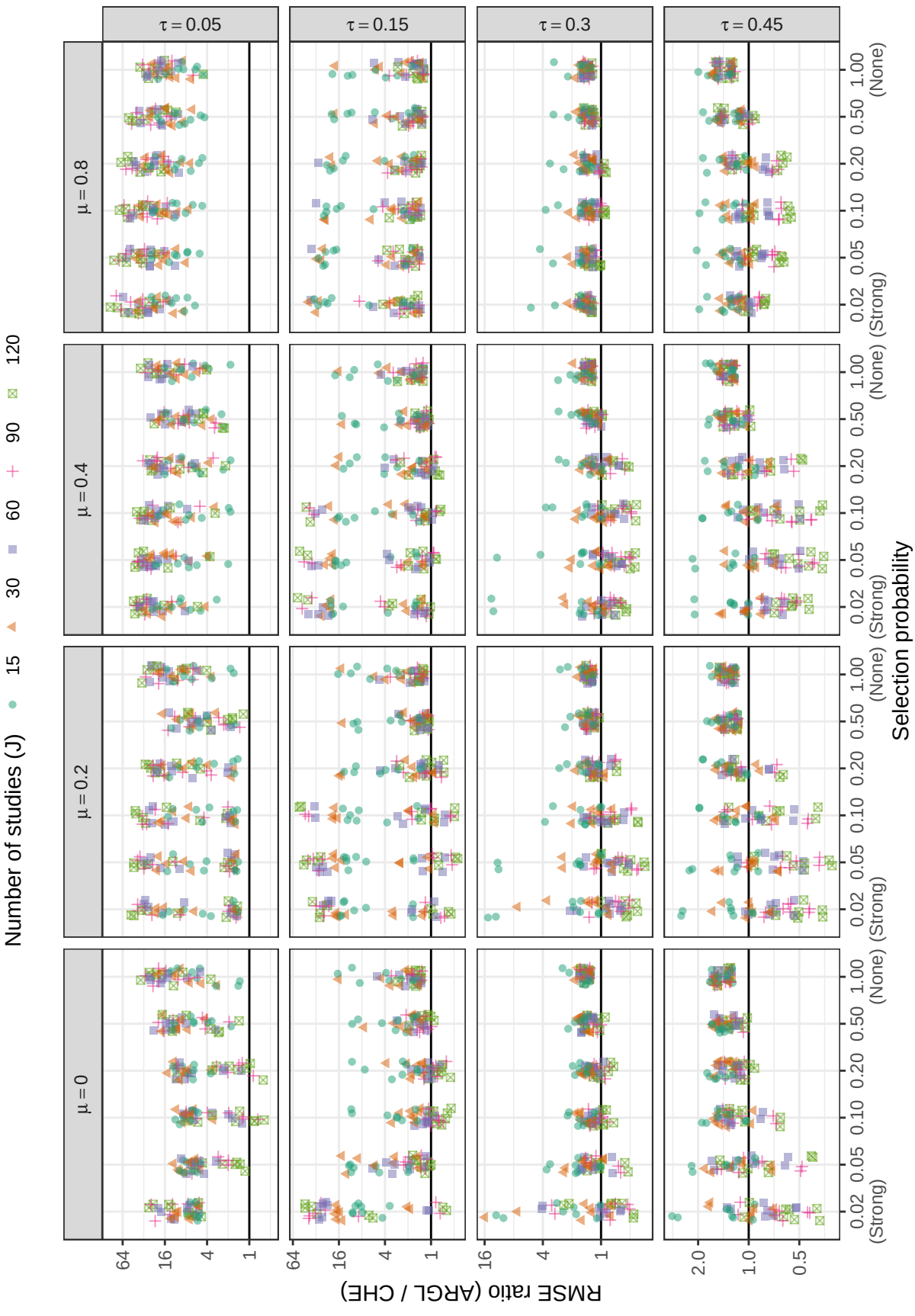

Figure E4

Ratio of root mean-squared error for ARGL heterogeneity estimator to root mean-squared error of CHE heterogeneity estimator by selection probability, number of studies, average SMD, and between-study heterogeneity

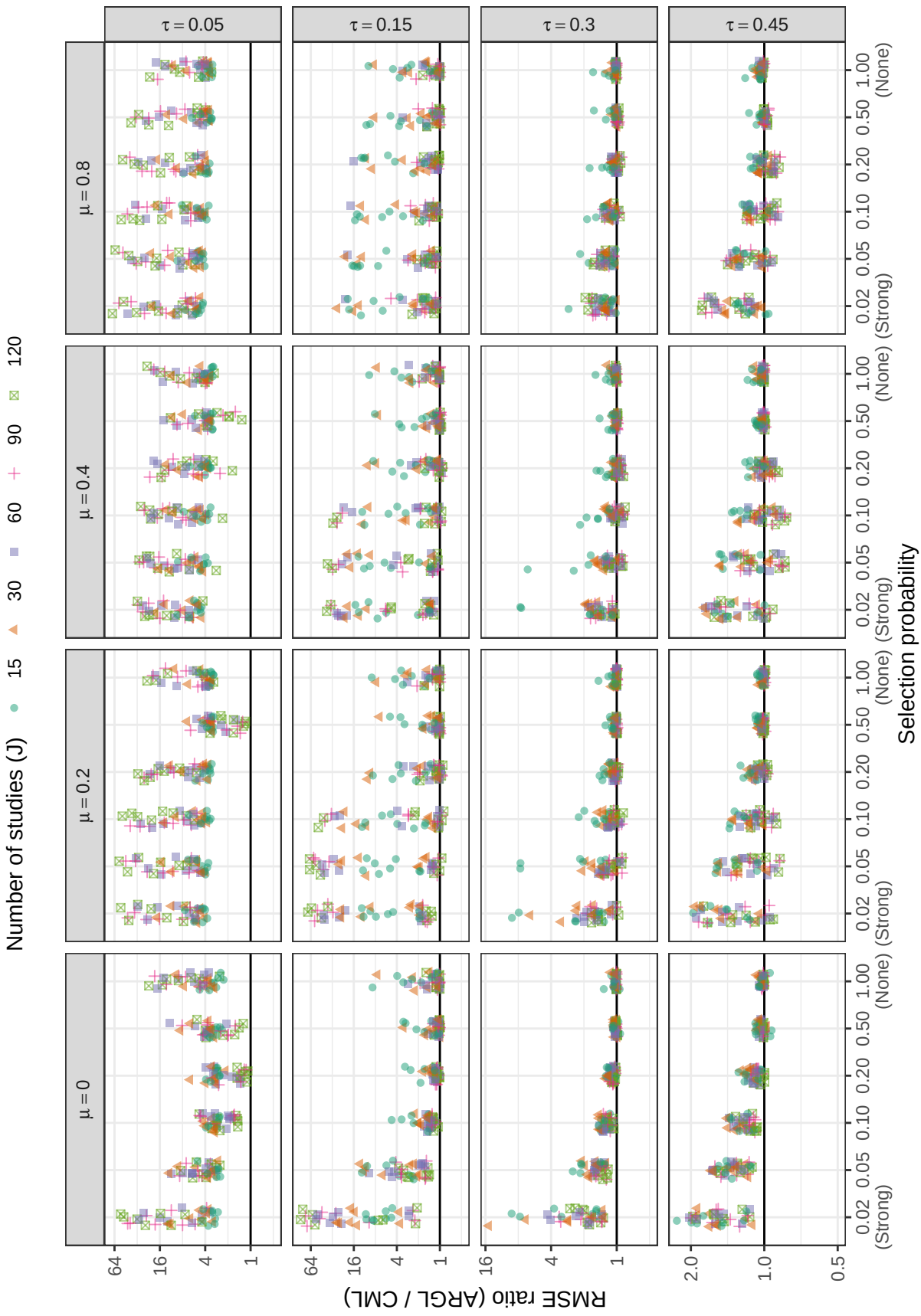

Figure E5

Ratio of root mean-squared error for ARGL heterogeneity estimator to root mean-squared error of CML heterogeneity estimator by selection probability, number of studies, average SMD, and between-study heterogeneity

Appendix F

Additional simulation results for estimators of selection parameter (ζ_1)

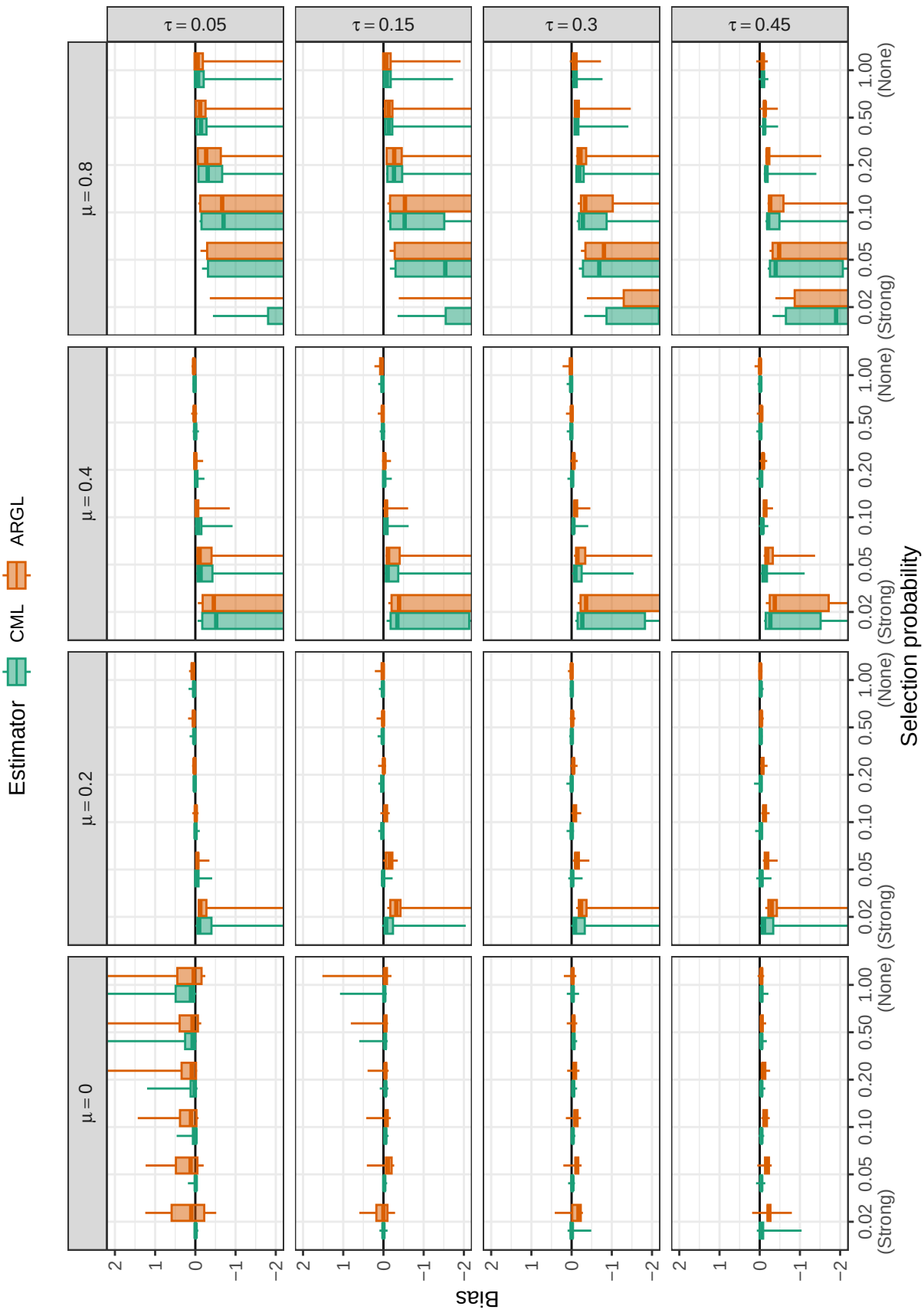

Figure F1

Bias for estimators of log-selection parameter by selection probability, average SMD, and between-study heterogeneity

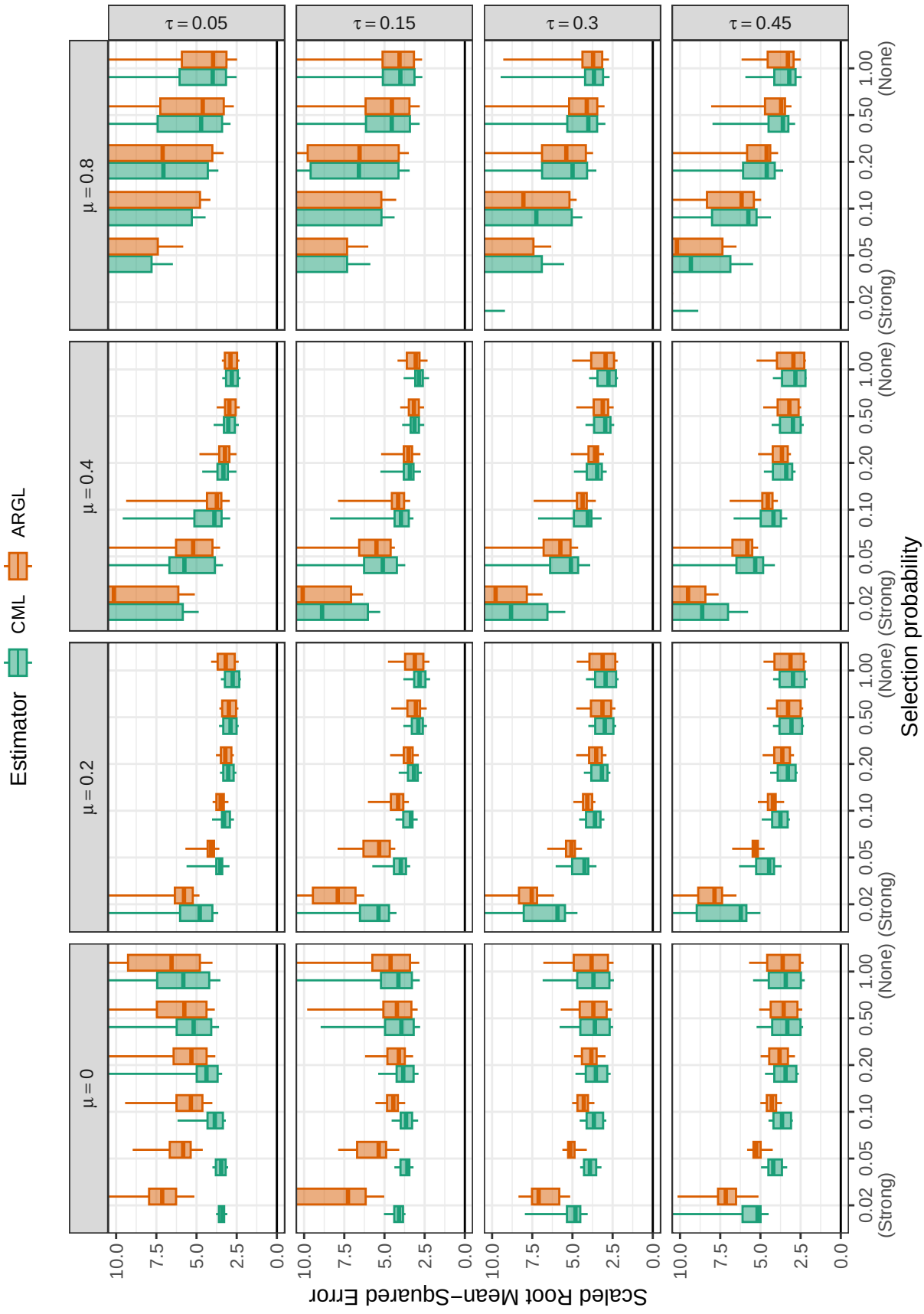

Figure F2

Scaled root mean-squared error for log-selection parameter estimators by selection probability, average SMD, and between-study heterogeneity

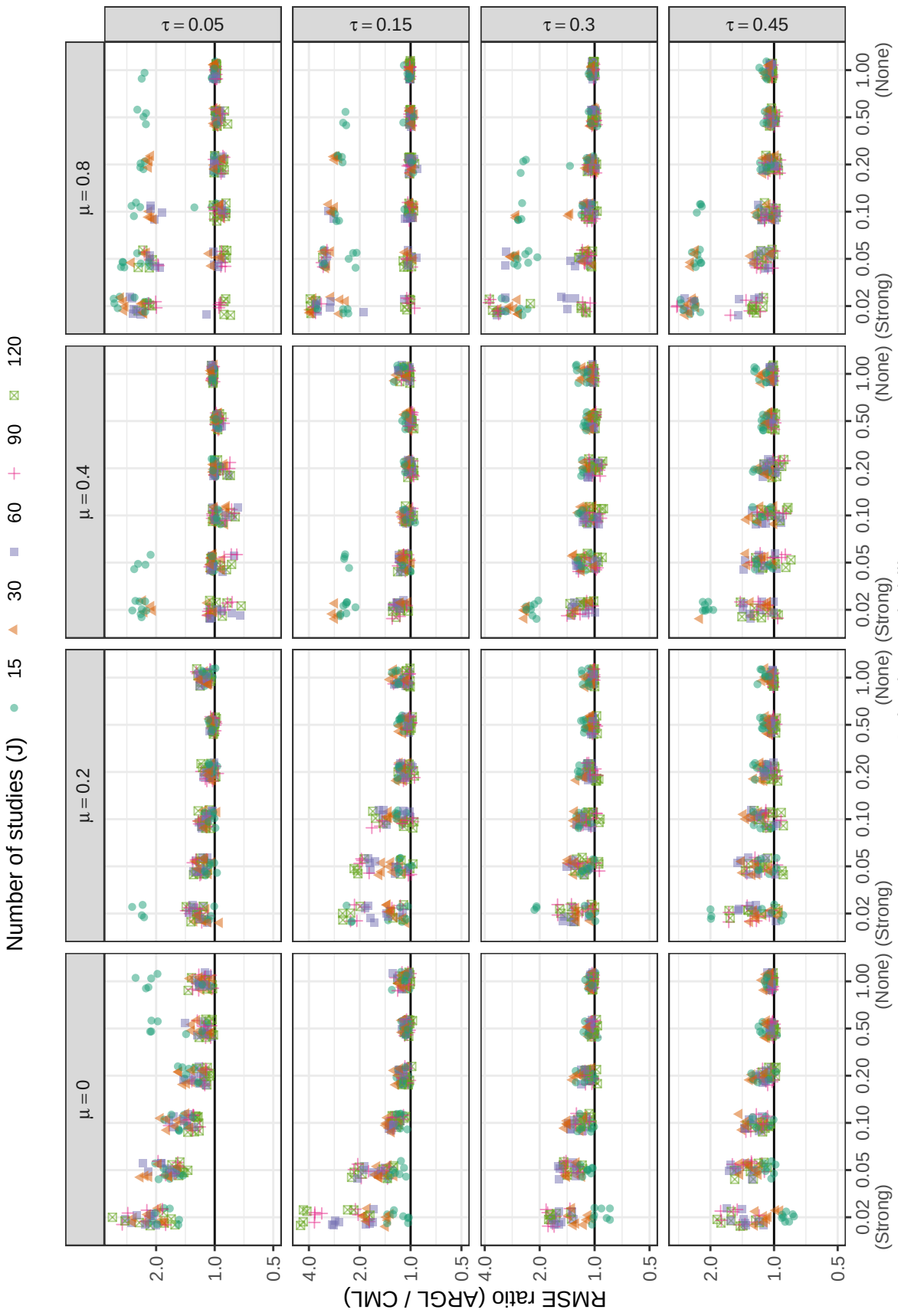

Figure F3

Ratio of root mean-squared error for ARGL log-selection parameter estimator to root mean-squared error of CML log-selection parameter estimator by selection probability, number of studies, average SMD, and between-study heterogeneity

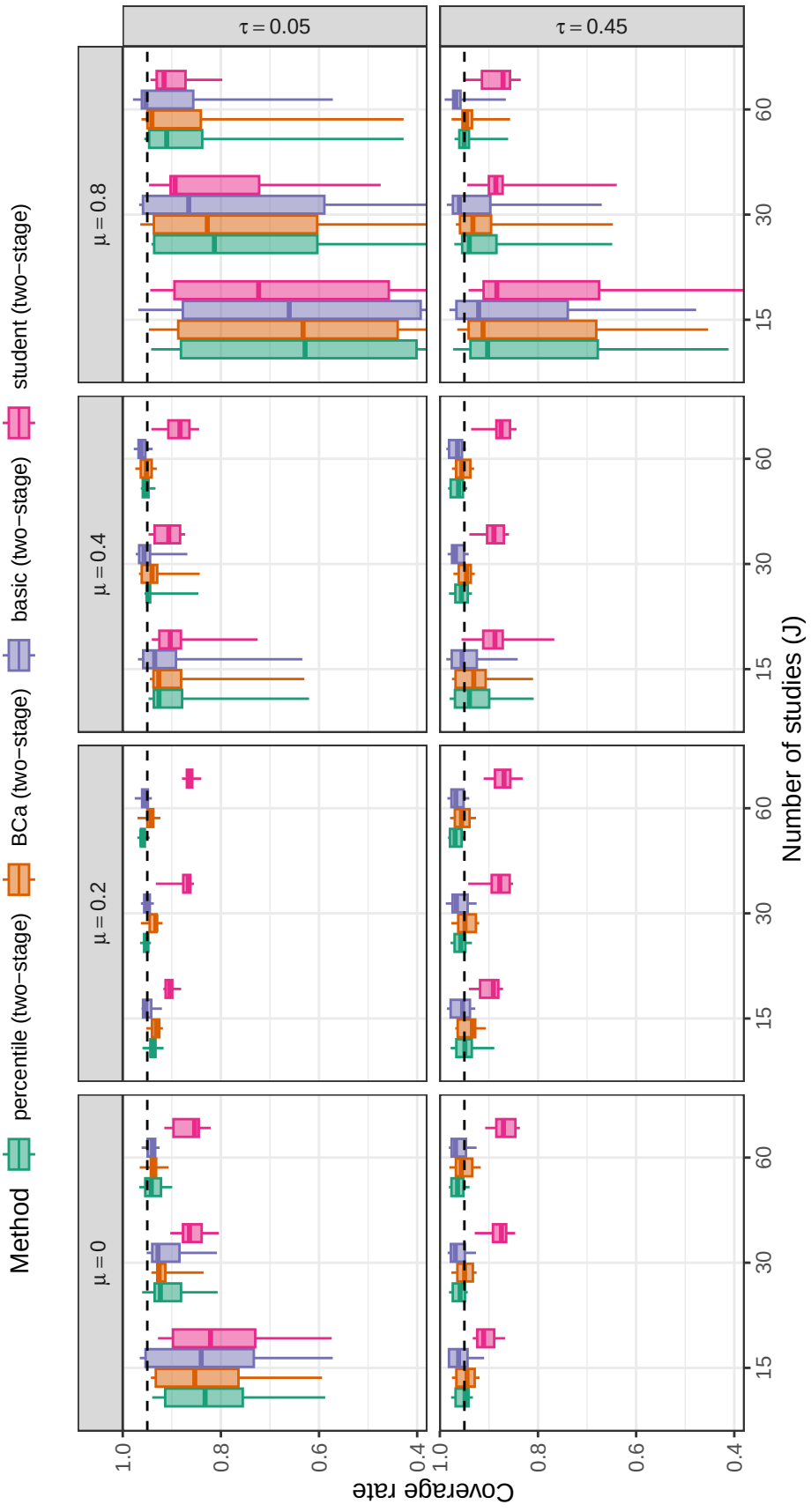

Figure F4
Coverage levels of two-stage bootstrap confidence intervals based on the CML estimator of log-selection parameter by number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95.

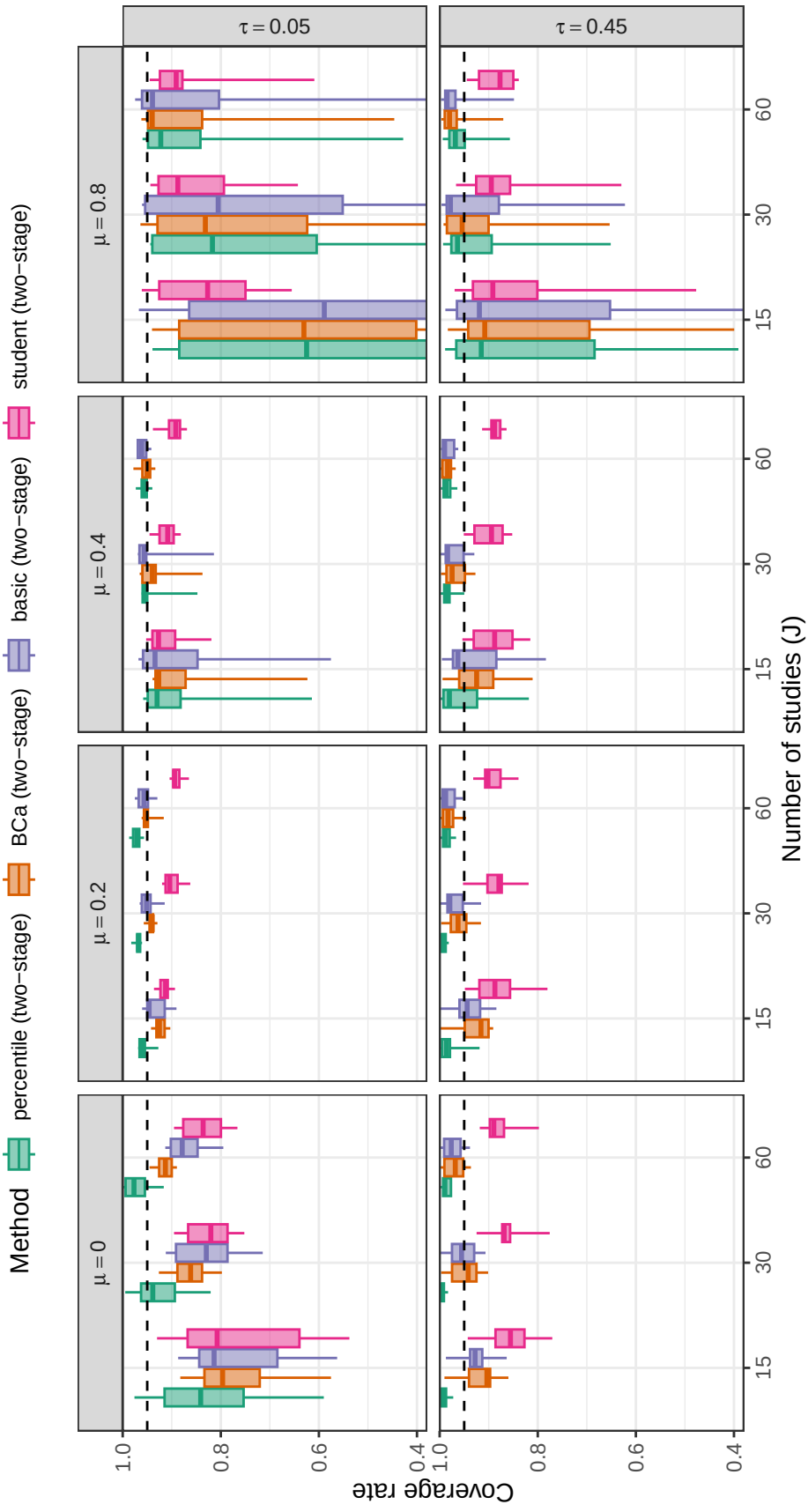

Figure F5

Coverage levels of two-stage bootstrap confidence intervals based on the ARGL estimator of log-selection parameter by number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95.