# Estimating beta-function selection models in meta-analysis with dependent effects

Martyna Citkowicz[1] & James E. Pustejovsky[2]

[1] American Institutes for Research

[2] University of Wisconsin-Madison

August 01, 2025

## Author Note

Correspondence concerning this article should be addressed to Martyna Citkowicz, Address. E-mail: mcitkowicz@air.org

## Highlights

**What is already known**

**What is new**

**Potential impact for RSM readers**

# 1 Estimating beta-function selection models in meta-analysis with dependent effects

Meta-analysis is a critical tool for synthesizing evidence across studies to draw generalizable conclusions in fields such as education, psychology, and medicine. However, the validity of meta-analytic findings depends on the completeness and representativeness of the available data. A persistent threat to validity is selective reporting, where effect size estimates are more likely to be published, and thus included in a meta-analysis, if they are statistically significant or consistent with researchers' hypotheses[1]. Extensive evidence from diverse domains suggests that such reporting biases are widespread[e.g., 2,3–7], resulting in distorted estimates of intervention effects and potentially misleading conclusions.

To address selective reporting bias, a variety of statistical approaches have been developed. Among the most promising are $p$-value selection models, which assume that the probability of an effect being reported depends on its statistical significance. These models are appealing because they are based on explicit assumptions about the selection mechanism and can be integrated into conventional meta-analytic frameworks (e.g., meta-regression). The step-function selection model, originally developed by[8] and later extended by[9], assumes a piece-wise constant selection probability across different $p$-value thresholds (e.g., $\alpha = .05$). This approach captures plausible patterns of selective reporting and has been shown to outperform simpler diagnostics or regression-based adjustments, particularly when effect sizes are heterogeneous[1,10].

Most existing methods for assessing and/or correcting for selective reporting—including $p$-value selection models—have been developed under the assumption that each study contributes a single, independent effect size. However, this assumption is increasingly unrealistic. Many meta-analyses now include multiple, dependent effect sizes per study, such as estimates derived from different outcomes, time points, or treatment comparisons. Our recent work[11] addressed this gap by integrating the step-function selection

model with robust variance estimation (RVE) and bootstrap methods to account for dependent effect sizes. Through simulation studies, we demonstrated that this approach reduces bias in the estimate of the overall effect size, but that there is a bias variance trade-off relative to the unadjusted meta-analytic model. Moreover, cluster bootstrapping leads to confidence intervals with coverage rates that are close to the nominal level of 0.95.

While the step-function selection model provides a structured and intuitive way to characterize selective reporting, it relies on the meta-analyst to specify "psychologically salient" $p$-value thresholds, such as 0.05 or 0.01. In practice, the true pattern of selection may not conform neatly to such step-wise forms. To address this limitation, the present paper introduces a beta-function selection model that allows the selection probability to vary smoothly as a function of the $p$-value by using the beta density to model the selection process. This model builds on earlier work by[12], extending it to the context of dependent effect sizes by incorporating RVE and bootstrap methods. The beta-function selection model both offers greater flexibility in capturing diverse forms of selection in meta-analyses with dependent effect sizes and allows meta-analysts to assess whether the form of selection matters.

We begin the paper by formally describing the beta-function selection model and outlining our proposed estimation and inference procedures. We then illustrate the model using a previously published meta-analysis, highlighting how it can reveal patterns of selective reporting not captured by the step-function selection model. Next, we report findings from an extensive simulation study that evaluates the model's performance under a variety of conditions and that investigates whether the form of selection matters. We conclude with summary findings, limitations, directions for future research, and implications for practice.

## 2   Models and Estimation Methods

Selection models comprises two components. The first component, herein after termed the *evidence-generating process*, models the distribution of effect sizes before

selection, typically using a conventional random-effects model or meta-regression model. The second component, hereinafter termed the *selection process*, identifies how the distribution is changed based on the likelihood of an effect size being reported. The combined model provides parameter estimates that define the selection process, along with meta-analytic estimates that are adjusted for selective reporting.

Following the approach outlined in our previous paper[11], we model the *marginal* distribution of effect size estimates rather than the joint distribution within studies. To account for dependence among effect sizes, we use cluster-robust variance estimation or clustered bootstrap methods, which accommodate within-study correlation without requiring explicit modeling of the dependence structure. While this strategy limits interpretation to the marginal distribution and does not distinguish between study-level and outcome-level selection, it remains a practical and plausible framework for modeling selective reporting based on the significance of individual estimates.

We use the following notation to describe the model and estimation procedures. Consider a meta-analytic dataset comprising $J$ studies, where study $j$ reports $k_j$ effect size estimates. Let $y_{ij}$ denote the $i$th effect size estimate from study $j$, with associated standard error $\sigma_{ij}$ and one-sided $p$-value $p_{ij}$. The one-sided $p$-value is defined relative to the null hypothesis that the true effect is less than or equal to zero. Let $\mathbf{x}_{ij}$ be a $1 \times x$ row vector of predictors representing characteristics of the effect size, sample, or study procedures. We use $\Phi()$ to denote the standard normal cumulative distribution function and $\phi()$ to denote the standard normal density function.

## 2.1 Evidence-generating process

We assume an evidence-generating process based on a standard random-effects meta-regression model. Let $Y^*$ denote a potentially reported effect size estimate, with standard error $\sigma^*$, one-sided $p$-value $p^*$, and predictor vector $\mathbf{x}^*$. Then the evidence-generating process is defined 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 and $\tau^2$ is the marginal variance of the effect size distribution. This model treats effect sizes as independent and characterizes *total* heterogeneity without decomposing within- and between-study variation.

## 2.2 Selection process

A $p$-value selection process is defined by a selection function that specifies the probability that an effect size is reported, conditional on its $p$-value. Let $O$ indicate whether $Y^*$ is observed. The process implies that

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

where $w\left(.; \boldsymbol{\lambda}\right)$ is a known, strictly positive function on the interval $[0, 1]$ with an unknown $h \times 1$ parameter vector $\boldsymbol{\lambda}$.

[12] defined the selection function using a truncated beta density with two parameters, offering flexibility to capture diverse selection patterns more parsimoniously than the step functions developed by[8] and[9]. Since the beta density can be unbounded near 0 and 1, they proposed truncating it to make the model computationally tractable, assuming constant selection probabilities for $p$-values in the range $[0, \alpha_1]$ and $[\alpha_2, 1]$. Given these pre-specified thresholds $\alpha_1$ and $\alpha_2$ and selection parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$, the beta density selection function is expressed by

$$w(p_i^*, \boldsymbol{\lambda}) = \begin{cases} \alpha_1^{\lambda_1-1}(1-\alpha_1)^{\lambda_2-1} & \text{if} \quad p_i^* \leq \alpha_1 \\ (p_i^*)^{\lambda_1-1}(1-p_i^*)^{\lambda_2-1} & \text{if} \quad \alpha_1 < p_i^* < \alpha_2 \\ \alpha_2^{\lambda_1-1}(1-\alpha_2)^{\lambda_2-1} & \text{if} \quad \alpha_2 \leq p_i^*. \end{cases} \tag{3}$$
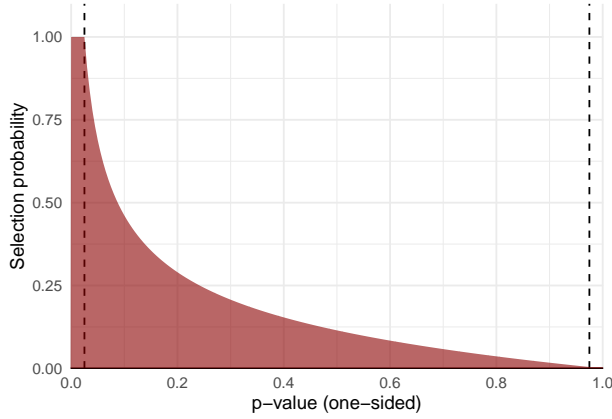
(a) *Selection:* $\lambda_1 = 0.5, \lambda_2 = 2.0$          (b) *No selection:* $\lambda_1 = 1, \lambda_2 = 1$
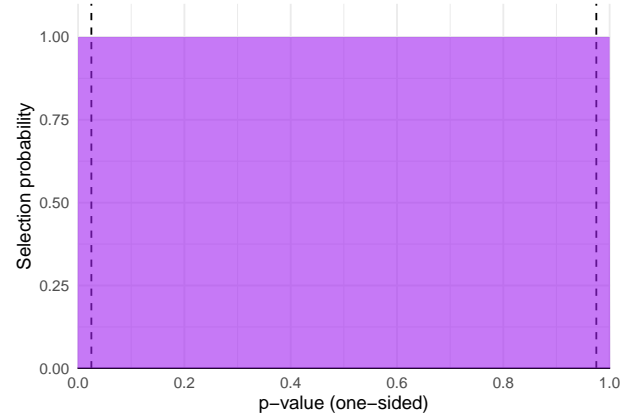
**Figure 1**

*Examples of beta density functions*

Equation (3) can be written equivalently as

$$
w(Y_i^*/\sigma_i^*, \boldsymbol{\lambda}) = \begin{cases} \alpha_1^{\lambda_1-1}(1-\alpha_1)^{\lambda_2-1} & \text{if} \quad \sigma_i^*\Phi^{-1}(1-\alpha_1) \leq Y_i^* \\[2mm] \left[\Phi\left(-Y_i^*/\sigma_i^*\right)\right]^{\lambda_1-1}\left[\Phi\left(Y_i^*/\sigma_i^*\right)\right]^{\lambda_2-1} & \text{if} \quad \sigma_i^*\Phi^{-1}(1-\alpha_2) < Y_i^* < \sigma_i^*\Phi^{-1}(1-\alpha_1) \\[2mm] \alpha_2^{\lambda_1-1}(1-\alpha_2)^{\lambda_2-1} & \text{if} \quad Y_i^* \leq \sigma_i^*\Phi^{-1}(1-\alpha_2). \end{cases}
$$

(4)

[12] used extreme truncation points ($\alpha_1 = 10^{-5}$, $\alpha_2 = 1 - 10^{-5}$), but such choices can make the model overly sensitive to rare, extreme $p$-values, potentially producing implausible estimates[13]. Using more moderate, psychologically salient thresholds such as $\alpha_1 = .025$ and $\alpha_2 = .975$ could potentially reduce this sensitivity and yield more plausible selection patterns.

1 shows several shapes the beta density can assume. 1a. presents a typical situation in which strong preference is shown for highly significant one-sided $p$-values, where $\lambda_1 = 0.5$ and $\lambda_2 = 2.0$. 1b. presents $\lambda_1 = 1$ and $\lambda_2 = 1$, which signifies no selection.

## 2.3 Distribution of observed effect size estimates

The combined model for the marginal density of an observed effect size estimate $Y$ with standard error $\sigma$ has the form

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

where

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

If $w(y, \sigma; \boldsymbol{\lambda}) = 1$, then $A(\mathbf{x}, \sigma; \boldsymbol{\beta}, \tau^2, \boldsymbol{\lambda}) = 1$ and there is no selective reporting, as depicted in 1b. The density then reduces to the unweighted density of the evidence-generating process and the $\boldsymbol{\beta}$ estimates from the adjusted beta function selection model will approximate those of the standard meta-analytic model.

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

$$A_{ij} = A(\mathbf{x}_{ij}, \sigma_{ij}; \boldsymbol{\beta}, \tau^2, \boldsymbol{\lambda}) = \alpha_1^{\lambda_1}(1 - \alpha_1)^{\lambda_2} B_{0ij} + E_Y(1|\lambda_1, \lambda_2) + \alpha_2^{\lambda_1}(1 - \alpha_2)^{\lambda_2} B_{2ij} \qquad (7)$$

where

$$E_Y[f(Y)|\lambda_1, \lambda_2] = \int_{\sigma_{ij}\Phi^{-1}(1-\alpha_2)}^{\sigma_{ij}\Phi^{-1}(1-\alpha_1)} f(Y) \left[\Phi(-Y/\sigma_{ij})\right]^{\lambda_1} \left[\Phi(Y/\sigma_{ij})\right]^{\lambda_2} \frac{1}{\sqrt{\eta_{ij}}} \phi\left(\frac{Y - \mu_{ij}}{\sqrt{\eta_{ij}}}\right) dY, \quad (8)$$

$c_{hij} = \left(\sigma_{ij}\Phi^{-1}(1 - \alpha_h) - \mathbf{x}_{ij}\boldsymbol{\beta}\right)/\sqrt{\tau^2 + \sigma_{ij}^2}$ for $h = 1, 2$, $B_{0ij} = 1 - \Phi(c_{1ij})$, and $B_{2ij} = \Phi(c_{2ij})^{12}$.

## 2.4 Estimation Method

We estimate model parameters using maximum composite marginal likelihood (CML), which treats each observed effect size estimate as if it were mutually independent, following established composite likelihood approaches[e.g., 14,15,16]. Estimation proceeds by maximizing a

weighted log-likelihood function defined over the marginal contributions of each observation, using reparameterizations of the variance and selection parameters. Confidence intervals are constructed using robust (sandwich-type) variance estimators based on study-level score contributions. A detailed explanation of CML methods is provided in our previous paper[11], and the exact expressions used for estimating the beta-function selection model are presented in APPENDIX.

## 2.5 Bootstrap inference

To improve inference accuracy with a limited number of studies, we also implement bootstrap procedures, which generate pseudo-samples through random resampling or reweighting of the original data. We consider both the non-parametric clustered bootstrap and the fractional random weight bootstrap[17], which differ in how they preserve the dependence structure across clusters. Confidence intervals are then computed using standard bootstrap-based methods such as the percentile, basic, studentized, and bias-corrected-and-accelerated intervals[18,19]. These resampling-based procedures are particularly useful in small-sample contexts where sandwich estimators may perform poorly. APPENDIX provides further details about the bootstrap CI calculations.

## 3 Empirical Example

To illustrate the new beta-function selection model in an empirical application, we reanalyzed the data from a meta-analysis conducted by BSCS Science Learning. BSCS Science Learning conducted a meta-analysis of science interventions effects on K–12 student science achievement[20]. The meta-analytic sample included findings from 96 studies with 292 effects. Approximately half of the studies contributed multiple effects per study, due to either multiple samples or multiple outcomes, leading to dependent effect sizes. Effect sizes were measured as standardized mean differences representing the effects of the interventions on student achievement outcomes. We conducted the analyses using R Version 4.4.3[21].

We used the `rma.uni()` function from the `metafor` package to fit a standard

random-effects model based on an independent effects working model with random effects for each study[22]. To adjust for selective reporting, we used the `selection_model()` function from the `metaselection` package to fit the new beta-function selection model along with our recently developed step-function selection model[23]. Both selection models are based on an independent effects working model with random effects for each study and account for dependent effects using RVE. For the step-function model, we estimated both a single-step model with a threshold at $\alpha_1 = 0.025$ and a two-step model with thresholds at $\alpha_1 = 0.025$ and $\alpha_2 = 0.5$.

The overall estimate of the average effect using the unadjusted model is 0.44, 95% CI [0.38, 0.50], which is significantly different from zero ($p = 0$). The estimates from the selection models are also all positive and significant, but they vary in magnitude and thus their adjustment for selective reporting. Compared to the unadjusted model, the average adjusted effect is 25% small when using the beta-function selection model ($ES = 0.33$), 32% larger when using the single-step model ($ES = 0.58$), and 12% larger when using the two-step selection model ($ES = 0.49$).

A key advantage of selection models over other methods for diagnosing and adjusting for selective reporting bias is their ability to include both discrete and continuous moderators. This allows meta-analysts to distinguish selective reporting bias from systematic variation in effect sizes that can be explained by primary study characteristics. The BSCS Science Learning meta-analytic data includes a number of potential moderators of effect size. Among these is an indicator for assessment author type, defined as 1 for assessment developer is the author or researcher and 0 otherwise. We reran all four models using assessment author type as a moderator[20].

When the assessment developer is the author or researcher, the unadjusted estimate is 0.09, 95% CI [-0.04, 0.21], which is not significantly different from zero ($p = 0.18$). The estimates from the selection models are also all nonsignificant and ranged from 0.10 to 0.14.

However, when the assessment developer is not the author or researcher, the models lead to different conclusions. The estimate from the unadjusted model is 0.38 and significant (95% CI [0.27, 0.48], $p = 0$). The estimate from the beta-function selection model is 0.21 and nonsignificant ($p = 0.39$). The estimates from the step-function selection models are both significant, with estimates of 0.51 for the single-step selection model and 0.40 for the two-step selection model.

The model estimates reveal some key differences between estimation methods. In this dataset, the beta-function selection model adjusts the estimates downward—so much so in the moderator analysis that neither estimate remains statistically significant. The step-function selection model, on the other hand, adjusts the estimates upward, though the adjustment is minimal when using the two-step model. These differences suggest that the models may behave differently in different conditions. To investigate this, we conducted simulations across a wide range of conditions. The simulations will allow us to draw firmer conclusions about the comparative performance of these methods, including their robustness to misspecification of the selection function.

## 4   Simulation Methods

We conducted Monte Carlo simulation studies to assess the performance of the new beta-function selection model. The simulations covered a wide range of conditions in which primary studies contributed multiple, statistically dependent effect size estimates. We compared the new model to three exiting methods: (1) a new version of the correlated hierarchical effects model with inverse sampling-covariance weights (CHE-ISCW), which accounts for dependency but not selective reporting[24]; (2) the PET-PEESE method, which addresses selective reporting and has been adapted to handle dependent data structures[25]; and (3) our recently developed step-function selection model that accounts for both selective reporting and dependent effects[11]. We evaluated the model estimates based on convergence rates, bias, accuracy, and confidence interval coverage for estimating the average effect size

from the unselected distribution. Bootstrap confidence intervals were assessed under a narrower set of conditions to limit computational burden. Simulations were conducted in R Version 4.X.X [21] using the high-throughput computing cluster at the University of Wisconsin–Madison[26]. The code relied on several R packages, including metafor[22], clubSandwich[27], simhelpers[28], optimx[29], nleqslv[30], and tidyverse[31].

## 4.1 Data generation

We generated simulated data using an approach similar to[11], with the key difference that effect size estimates were selected for inclusion based on the beta-function selection model. For each simulated meta-analysis, we generated a pool of primary studies using a CHE model, with effect size estimates selected for inclusion according to probabilities defined by the beta-function selection model. Each study followed a two-group design, with sample sizes and numbers of effect sizes per study drawn from an empirical distribution based on the What Works Clearinghouse database. We generated outcome correlations across studies by sampling from a beta distribution with mean $\rho$ and standard deviation 0.05, but assumed a constant correlation between pairs of outcomes within a study.

Within each study, we simulated a study-level average effect size $\delta_j$, and then generated individual effect size parameters from a normal distribution centered at $\delta_j$ with variance $\omega^2$. Using these parameters, we drew multivariate normal outcomes for participants equally divided into treatment and control groups and computed standardized mean differences with Hedges's $g$ small sample bias correction. One-sided $p$-values were computed for each effect size, and weights from the beta-function selection model were applied to determine the probability of selection. We repeated this process until the simulated meta-analytic dataset included at total of $J$ studies with at least one observed result. See[11] for details.

### 4.2  Estimation methods

We estimated the beta-function selection model using the CML approach described in Section 4.2. 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 confidence intervals based on the two-stage bootstrap.* To maintain computational feasibility, we used $B = 399$ bootstrap replications of each estimator.

We compared the performance of the beta-function selection model to three other methods. First, we estimated a summary meta-analysis model using the CHE-ISCW approach proposed by[24], which accounts for effect size dependence but does not adjust for selective reporting. This method fits a CHE working model, but it allocates more weight to studies with smaller sampling variances by using generalized least squares with weighting matrices that are the inverse of the variance-covariance matrix of the sampling errors only. We assumed a correlation of 0.80, which allows for some misspecification when the average correlation used in the data-generating process differs from 0.80. Confidence intervals were computed using cluster-robust variance estimation with the CR2 small-sample correction and Satterthwaite degrees of freedom.

Second, we estimated a variation of the PET/PEESE model originally proposed by[25], adapted to handle dependent effect sizes. The PET model regresses effect size estimates on their standard errors, while the PEESE model uses sampling variances instead. Both models assume normally distributed errors with a correlation of 0.80 and were estimated using the same procedure as the CHE-ISCW model, including using CR2 cluster-robust standard errors. Following[25], we used the PET estimate if it was not statistically distinguishable from zero at an $\alpha$-level of 0.10; otherwise, we used the PEESE estimate.

---

* In our recent work[11], we also evaluated confidence intervals using the non-parametric clustered bootstrap and the fractional random weight bootstrap. The two-stage bootstrap consistently outperformed the alternatives, so we focus exclusively on this approach in the present paper.

Third, we estimated the step-function selection model using the CML approach described in[11]. We estimated two step-function selection models: (1) a three-parameter selection model (3PSM) with a single step at $\alpha_1 = 0.025$, and (2) a four-parameter selection model (4PSM) with steps at $\alpha_1 = 0.025$ and $\alpha_2 = 0.500$. Like the beta-function selection model, the 3PSM and 4PSM are $p$-value selection models designed to address selective reporting. The new models account for effect size dependency using cluster-robust standard errors, modeling the marginal rather than joint distribution of estimates within studies. The main distinction between the step-function and beta-function selection models lies in their assumptions about the selection mechanism. By fitting the 3PSM and 4PSM to data simulated under a beta-function selection process, we can assess how robust these models are to misspecification of the selection function and whether the form of selection affects their performance.

## 4.3 Experimental design

We examined performance across a range of simulation conditions, summarized in Table 1. Manipulated parameters included overall average standardized mean difference ($\mu$), between-study heterogeneity ($\tau$), ratio of within- to between-study heterogeneity ($\omega^2/\tau^2$), average correlation between outcomes ($\rho$), probability of selection for non-affirmative results ($\lambda_1, \lambda_2$), and number of observed studies ($J$). The full simulation crossed all parameter values for a total of $4 \times 4 \times 2 \times 2 \times 5 \times 4 = 1,280$ conditions. For the more computationally intensive bootstrap simulations, we limited the design to a smaller subset ($4 \times 3 \times 2 \times 1 \times 3 \times 3 = 216$ conditions), focusing on smaller meta-analyses and reducing values for factors where results were stable (e.g., $\tau = 0.30$). For each condition, we generated 2,000 replications.

In the full simulation, we varied $\mu$ from 0.0 to 0.80, reflecting the range of effects observed in a large-scale review of education randomized controlled trials[32], and $\tau$ from 0.05 (minimal heterogeneity) to 0.45 (substantial heterogeneity). Within-study heterogeneity was specified relative to between-study heterogeneity using a ratio of $\omega^2/\tau^2$ equal to 0 (no

**Table 1**

*Parameter values examined in the simulation study*

| Parameter | Full Simulation | Bo |
|---|---|---|
| Overall average SMD ($\mu$) | 0.0, 0.2, 0.4, 0.8 | 0.0 |
| Between-study heterogeneity ($\tau$) | 0.05, 0.15, 0.30, 0.45 | 0.0 |
| Heterogeneity ratio ($\omega^2/\tau^2$) | 0.0, 0.5 | 0.0 |
| Average correlation between outcomes ($\rho$) | 0.40, 0.80 | 0.8 |
| Probability of selection for non-affirmative effects ($\lambda_1, \lambda_2$) | (0.01, 0.90), (0.20, 0.90), (0.50, 0.90), (0.80, 0.90), (1.00, 1.00) | (0. |
| Number of observed studies ($J$) | 15, 30, 60, 90 | 15 |

within-study heterogeneity) or 0.5.

To assess the impact of working model misspecification, we manipulated the average within-study correlation $\rho$ across two levels: 0.80 (the default used in RVE software and correctly specified when $\rho = 0.80$) and 0.40 (representing misspecification).

Selective reporting was modeled as the probability of selecting non-affirmative results based on the selection parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ fed into the beta-function selection model. We considered five levels of selective reporting, including no selection ($\lambda_1 = 1.00, \lambda_2 = 1.00$), weak selection ($\lambda_1 = 0.80, \lambda_2 = 0.90$), moderate selection ($\lambda_1 = 0.50, \lambda_2 = 0.90$), strong selection ($\lambda_1 = 0.20, \lambda_2 = 0.90$), and very strong selection ($\lambda_1 = 0.01, \lambda_2 = 0.90$).

The number of observed studies ($J$) ranged from 15 to 90, covering the typical size of meta-analyses in education and psychology[33].

Primary study sample sizes were drawn from an empirical distribution in the What Works Clearinghouse database. The sample sizes in the database ranged from 37 to 2,295 with a median of 211, and the number of effect sizes ranged from 1 to 48 with a median of 3.[*]

———

[*] In our previous paper, we included a condition in which the study sizes were divided by three to represent smaller studies, such as those in psychology lab settings. However, the simulation results for this condition were similar to those from the empirical distribution condition, so we have omitted it from the current

### 4.4   Performance criteria

We evaluated each method's performance in terms of convergence rates, bias, scaled root mean-squared error (RMSE), and 95% confidence interval coverage for the overall effect size $\mu$. Bias reflects systematic deviation from the true parameter value, while RMSE captures both bias and sampling variability. To account for expected reductions in RMSE with more studies, we scaled RMSE by $\sqrt{J}$. Bias and scaled RMSE were calculated after winsorizing to limit the influence of extreme outliers, using fences set at 2.5 times the interquartile range beyond the 25th and 75th percentiles.

For confidence intervals based on cluster-robust variance estimation, we defined coverage as the proportion of intervals that contained the true parameter value. For bootstrap intervals, we used $B = 399$ replicates per simulation due to computational limits—fewer than ideal for applied use. To estimate practical coverage rates, we followed an approach similar to[34]: we calculated coverage for smaller subsamples ($B = 49, 99, 199, 299$) randomly selected without replacement from the full set of $B = 399$ bootstraps, fit a regression of coverage on $1/B$, and used the intercept to extrapolate expected coverage for $B = 1999$.

## 5   Simulation Results

We organize our simulation results into two parts. First, we compare the performance of the beta-function selection model to the CHE-ISCW and PET/PEESE approaches. Specifically, we compare the bias and accuracy of the average effect size, the coverage rates of 95% confidence intervals of the average effect size, and the estimation of the marginal variance of the effect size distribution. Second, we compare the performance of the beta-function selection model to the 3PSM and 4PSM step-function models to assess the robustness of these models to misspecification of the selection function.

———

simulations.

## 5.1   Beta-Function Selection Model Compared to CHE-ISCW and PET/PEESE

### 5.1.1   Average Effect Size

The CHE-ISCW and PET/PEESE approaches had perfect convergence rates, producing results for every replication in every condition. The beta-function selection model exhibited very high convergence rates, with an average rate of 99.60%; the lowest convergence rates occurred under conditions with the lowest degree of heterogeneity $\tau = 0.05$. Supplementary Figure **??** depicts the range of convergence rates of the beta-function selection model.

**5.1.1.1   Bias.**   Figure**ref?**(fig:mu-bias-main) shows the bias for each method of estimating the average effect size (vertical axis) as a function of selective reporting strength (horizontal axis), average effect size (grid column), and between-study heterogeneity ($\tau$, grid row). Each box plot summarizes variation in bias across the remaining simulation factors: the heterogeneity ratio, the correlation between effect size estimates, and the number of observed studies. Note that the vertical axis scale differs by grid row, reflecting how some methods' bias is more sensitive to the level of heterogeneity.

The beta-function selection model has negligible to small bias across all conditions, ranging from -0.06 to 0.09. Its bias was essentially zero for all conditions when the average effect size is non-zero or when selective reporting is weak or absent. Bias was largest when selective reporting is very strong, average effect size is zero, and heterogeneity is large.

In contrast, bias for the comparison methods ranged from 0 to 0.41 for CHE-ISCW and -0.38 to 0.37 for PET/PEESE. Both methods are generally biased when selective reporting is not absent. For CHE-ISCW, which does not directly adjust for selective reporting, bias was closest to zero when average effect size is large ($\mu = 0.8$) and heterogeneity is low ($\tau \leq 0.15$). For PET/PEESE, which uses a regression adjustment to account for possible selective reporting, bias was closest to zero when average effect size is moderate ($\mu = 0.4$) and heterogeneity is low ($\tau = 0.15$). For both comparison methods, bias
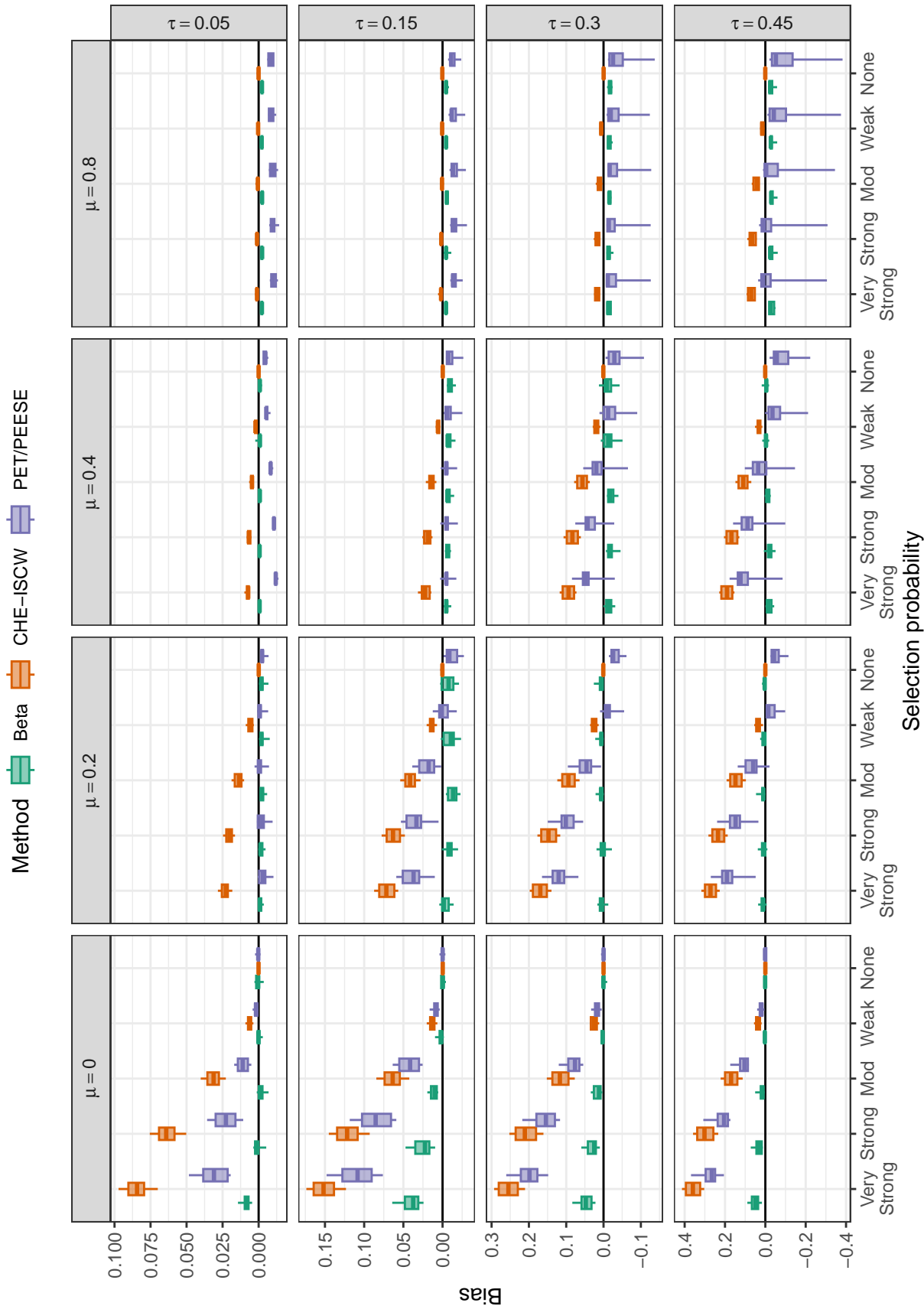
**Figure 2**
*Bias of the average effect size by method, selection probability, average SMD, and between-study heterogeneity*

grows stronger when selection is stronger, when average effect size is smaller, and when heterogeneity is larger; however, bias is generally less pronounced for PET/PEESE than CHE-ISCW.

**5.1.1.2  *Scaled RMSE.*** Scaled RMSE captures both bias and variability, providing an overall measure of inaccuracy. Figure**ref?**(fig:mu-rmse-main), constructed in the same format as Figure**ref?**(fig:mu-bias-main), shows the scaled RMSE for each method of estimating the average effect size. Additional detail is provided in Figures**ref?**(fig:rmse-CHE-Beta-main) and**ref?**(fig:rmse-PET-Beta-main) in Appendix**ref?**(mu-simulation-results-main), which plot the ratio of RMSEs for each pair of methods to compare their relative accuracy.

Taken together, the figures indicate that no single method achieves the lowest RMSE across all conditions. Instead, each method reflects different bias–variance trade-offs. The beta-function selection model generally outperforms the others—achieving lower RMSE—when selective reporting is strong or very strong, average effect size is small ($\mu \leq 0.2$), and heterogeneity is moderate to large ($\tau \geq 0.3$). CHE-ISCW consistently has the lowest RMSE when selective reporting is weak or absent. PET/PEESE performs best when selective reporting is strong to very strong, average effect is small ($\mu \leq 0.2$), and heterogeneity is low ($\tau \leq 0.15$).

These bias–variance trade-offs stem from the fact that CHE-ISCW, which does not explicitly adjust for selective reporting, is more prone to bias when selection is present. In contrast, the beta-function selection model tends to produce smaller biases under such conditions. However, when selective reporting is weak or absent, CHE-ISCW is more precise than the methods that adjust for selective reporting (i.e., the beta-function selection model and PET/PEESE). Under those conditions, the added variability introduced by estimating a selection model or PET/PEESE adjustment outweighs the minimal reduction in bias they provide.
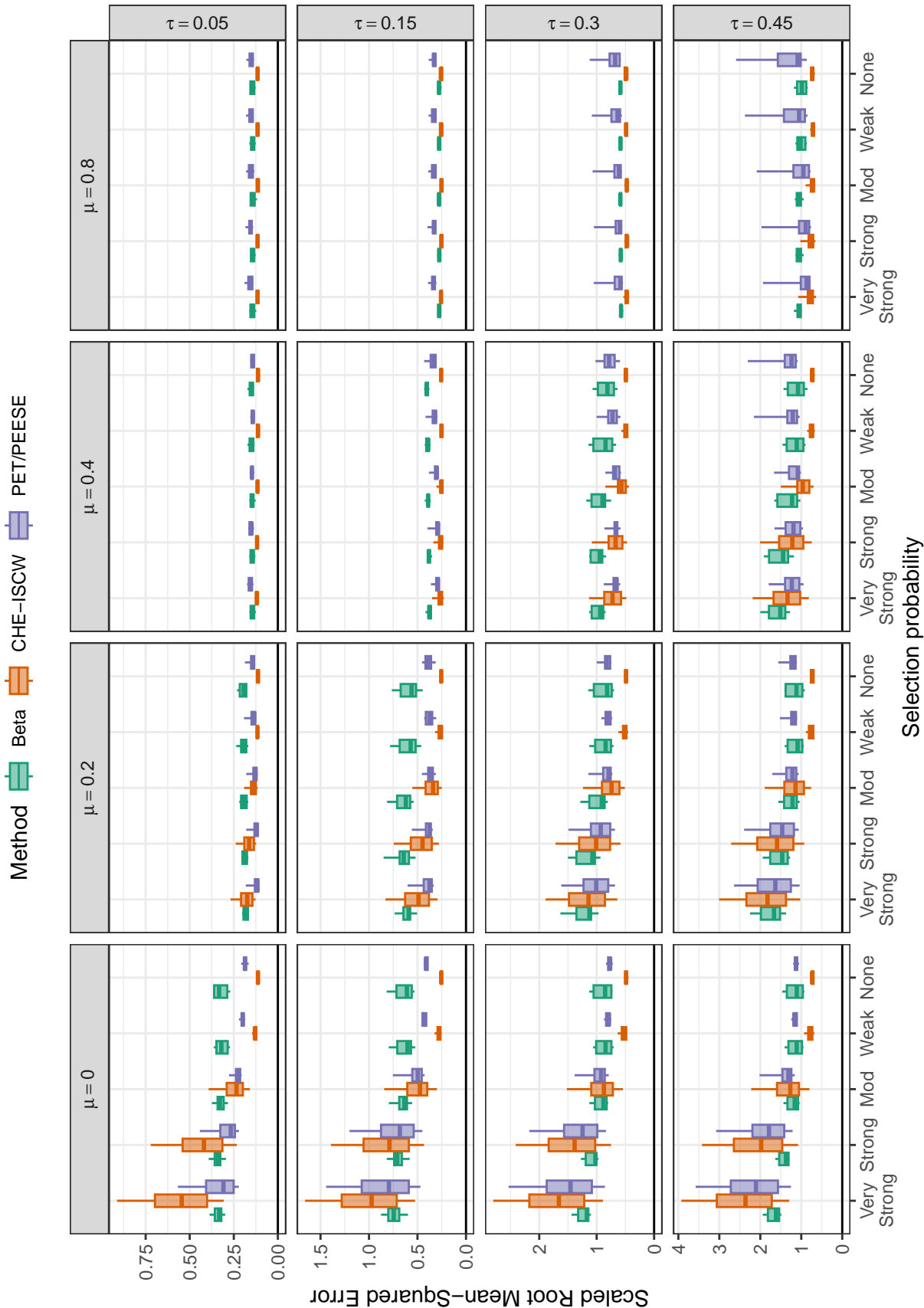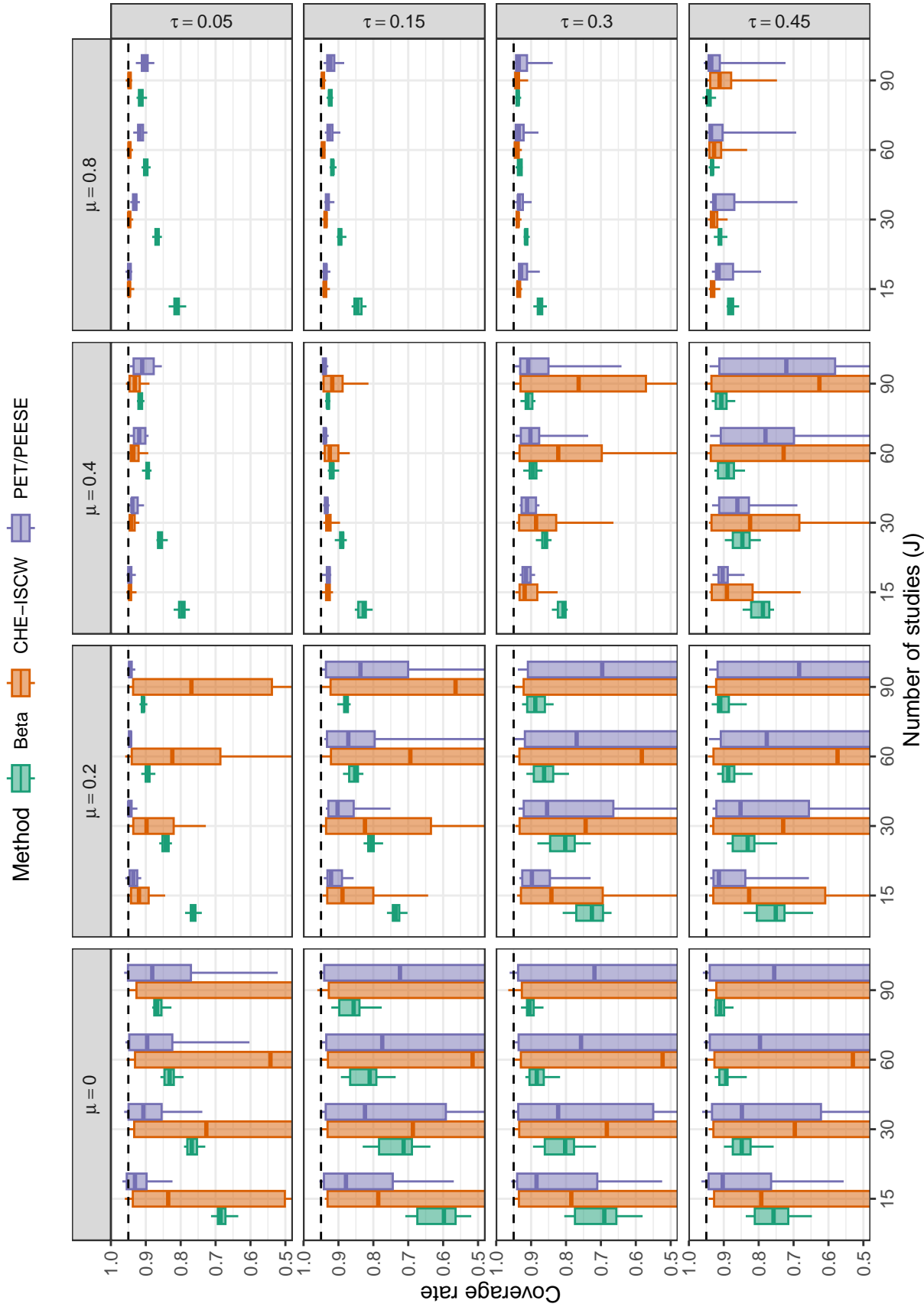
**Figure 3**

*Scaled root mean-squared error of the average effect size by method, selection probability, average SMD, and between-study heterogeneity*

**Figure 4**

*Coverage levels of confidence intervals for the average effect size based on cluster-robust variance approximations, by method, 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*

**5.1.1.3    *Confidence Interval Coverage.*** Figure 4 shows the coverage rates of 95% confidence intervals based on large-sample cluster-robust variance approximations for the three methods.* Across most conditions, coverage rates fall below the nominal rate of 0.95 for all methods. However, the beta-function selection model generally achieves higher coverage than the comparison methods, particularly when heterogeneity is moderate to large ($\tau \geq 0.3$), or when heterogeneity is small ($\tau \leq 0.15$), average effect size is small ($\mu \leq 0.2$), and number of studies ($J$) is 60 or more.

In contrast, the confidence intervals produced by the comparison methods are often severely miscalibrated. When CHE-ISCW and PET/PEESE are biased due to selective reporting, their intervals tend to be centered away from the true parameter. As a result, as the number of studies increases, the standard errors of the average effect size estimate——and thus the interval widths——shrink, causing coverage rates to decline sharply, in some cases approaching zero.

Figure 5 depicts the coverage rates of 95% confidence intervals for the average effect size estimated using the beta-function selection model, comparing intervals based on the large-sample cluster-robust variance method to four two-stage bootstrapping methods (percentile, basic, studentized, and bias-corrected-and-accelerated). Due to the computational demands of bootstrapping, we evaluated the bootstrap confidence intervals under a more limited range of data-generating conditions, including a maximum sample size of $J = 60$. Although no method achieves exact nominal coverage across all conditions, the percentile, studentized, and bias-corrected-and-accelerated bootstrap intervals consistently yield coverage rates comparable to——or better than——those based on the large-sample cluster-robust variance method. Among these, the percentile bootstrap intervals performed best, achieving coverage above 90% in nearly all conditions examined.

————

* Note that the vertical axis of Figure 4 is restricted to the range [0.5, 1.0], and coverage rates of the intervals based on CHE-ISCW and PET/PEESE are not depicted when they fall below 0.5. Supplementary Figure **??** depicts the full range of coverage rates.
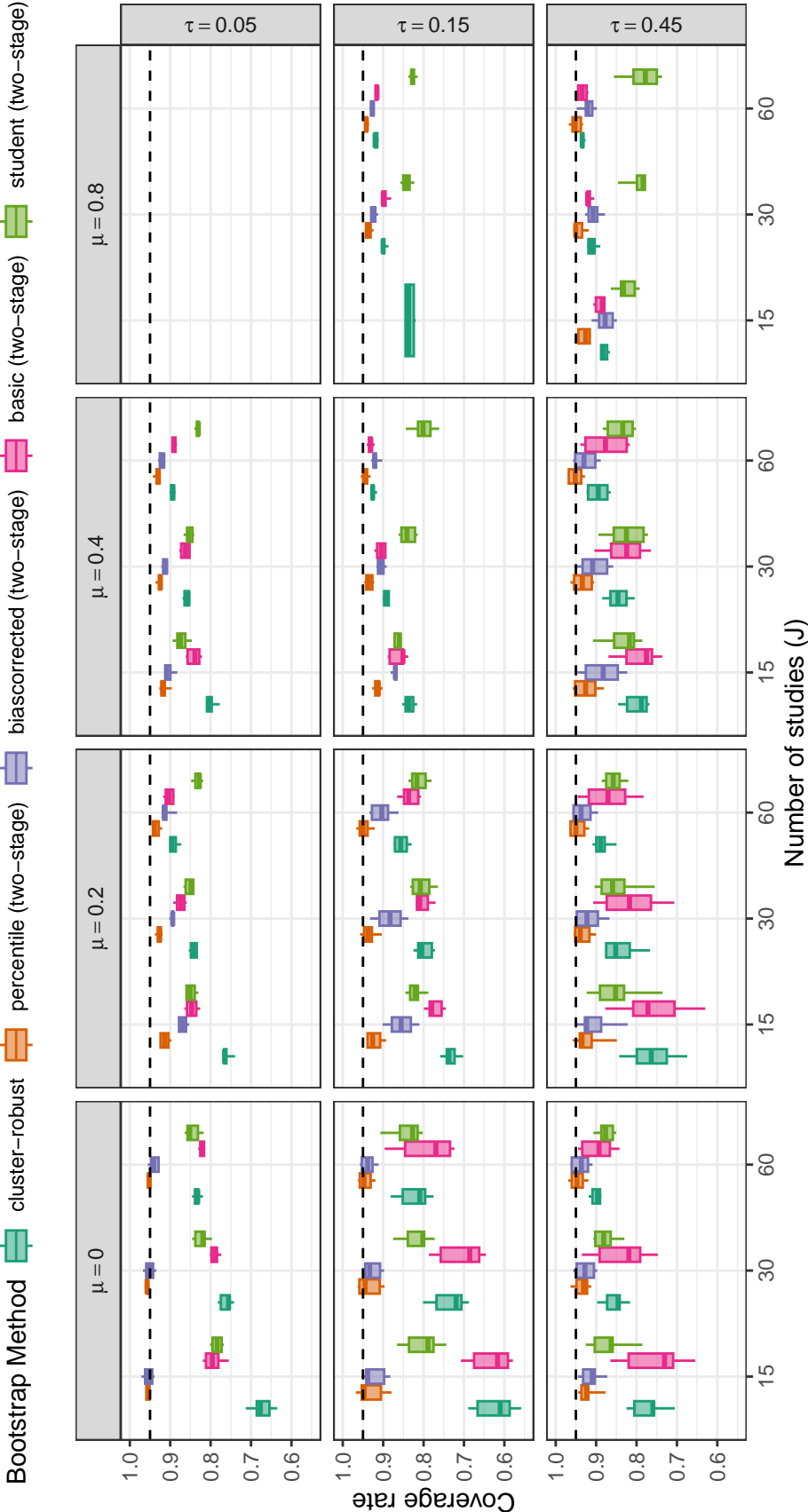
**Figure 5**

*Coverage levels of confidence intervals for the average effect size estimated using the beta-function selection model, by bootstrap method, number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95.*

## 5.2 Beta-Function Selection Model Compared to 3PSM and 4PSM Step-Function Selection Models

### 5.2.1 Average Effect Size

Convergence rates were higher for the step-function selection models than for the beta-function selection model. Both the 3PSM and 4PSM models had convergence rates of 99% and above across all 1,280 conditions, while the beta-function selection model had convergence rates below 99% for 40 conditions.

**5.2.1.1 Bias.** Figure**ref?**(fig:mu-bias-miss) displays the bias for the three selection models. Across most conditions, bias is consistently closer to zero when the average effect size is estimated using the beta-function selection model compared to the step-function selection models. Among the step-function selection models, 4PSM generally outperforms 3PSM, though both are more prone to bias when the selection process is misspecified. The main exception occurs when average effect size is moderate to large ($\mu \geq 0.4$) and heterogeneity is low ($\tau \leq 0.15$), in which case all three models exhibit essentially zero bias. This pattern is consistent with the fact that the data were generated under a beta-function selection process and suggests that the step-function selection models, particularly 3PSM, are not robust to misspecification of the selection mechanism——especially when selective reporting is moderate to strong or when average effect is not large ($\mu < 0.8$).

**5.2.1.2 Scaled RMSE.** Figure**ref?**(fig:mu-rmse-miss) presents the scaled RMSE for the three selection models and highlights a clear bias–variance trade-off. When average effect size is zero and selective reporting is moderate to very strong, the results mirror the bias results pattern: the beta-function selection model outperforms the step-function selection models, and the 4PSM performs better than the 3PSM. However, the relative performance shifts under other conditions. When average effect size is $\mu = 0.2$ and selective reporting is strong to very strong, 4PSM yields the lowest RMSE. In contrast, when average effect size is moderate or large ($\mu \geq 0.2$), or when $\mu = 0.2$ and selective reporting is absent to moderate, 3PSM outperforms both alternatives. RMSE is similar across all models when
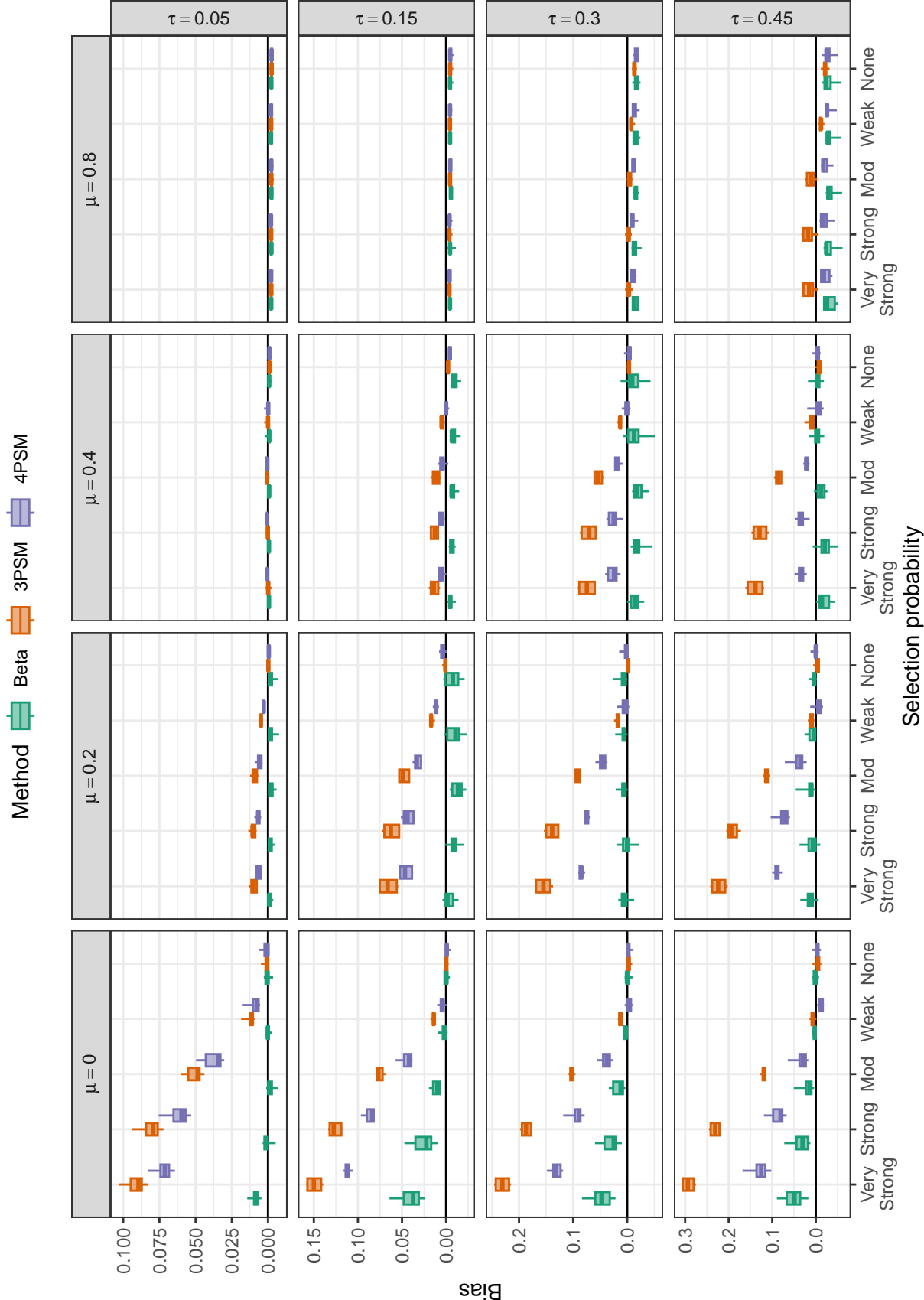
**Figure 6**

*Bias of the average effect size by method, selection probability, and between-study heterogeneity*
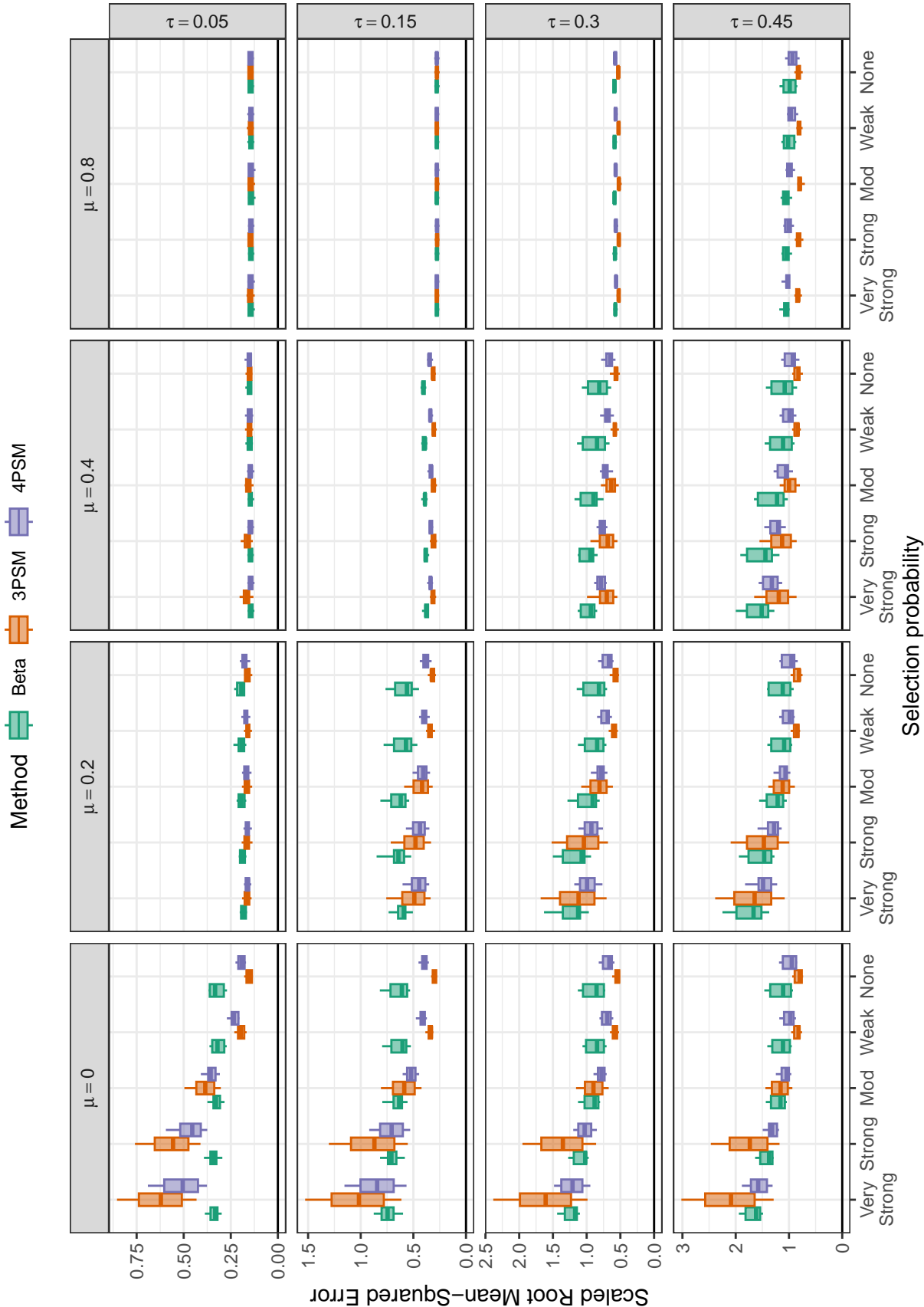
**Figure 7**
*Scaled root mean-squared error of the average effect size by method, selection probability, average SMD, and between-study heterogeneity*

average effect size is moderate to large ($\mu \geq 0.4$) and heterogeneity is low ($\tau \leq 0.15$).

**5.2.1.3 *Confidence Interval Coverage.*** Figure 8 shows the coverage rates of 95% confidence intervals based on large-sample cluster-robust variance approximations for the three models* Coverage rates fall below the nominal 0.95 level for all three selection models across most conditions. However, the beta-function selection model generally achieves higher coverage than the step-function selection models, particularly when heterogeneity is moderate to large ($\tau \geq 0.3$), or when heterogeneity is low ($\tau \leq 0.15$), average effect size is small ($\mu \leq 0.2$), and number of studies ($J$) is 60 or more. As with CHE-ISCW and PET/PEESE, the confidence intervals produced by the step-function selection models are often miscalibrated, due in part to underestimated standard errors, which result in overly narrow intervals. Among the step-function selection models, coverage is generally higher for 4PSM than for 3PSM.

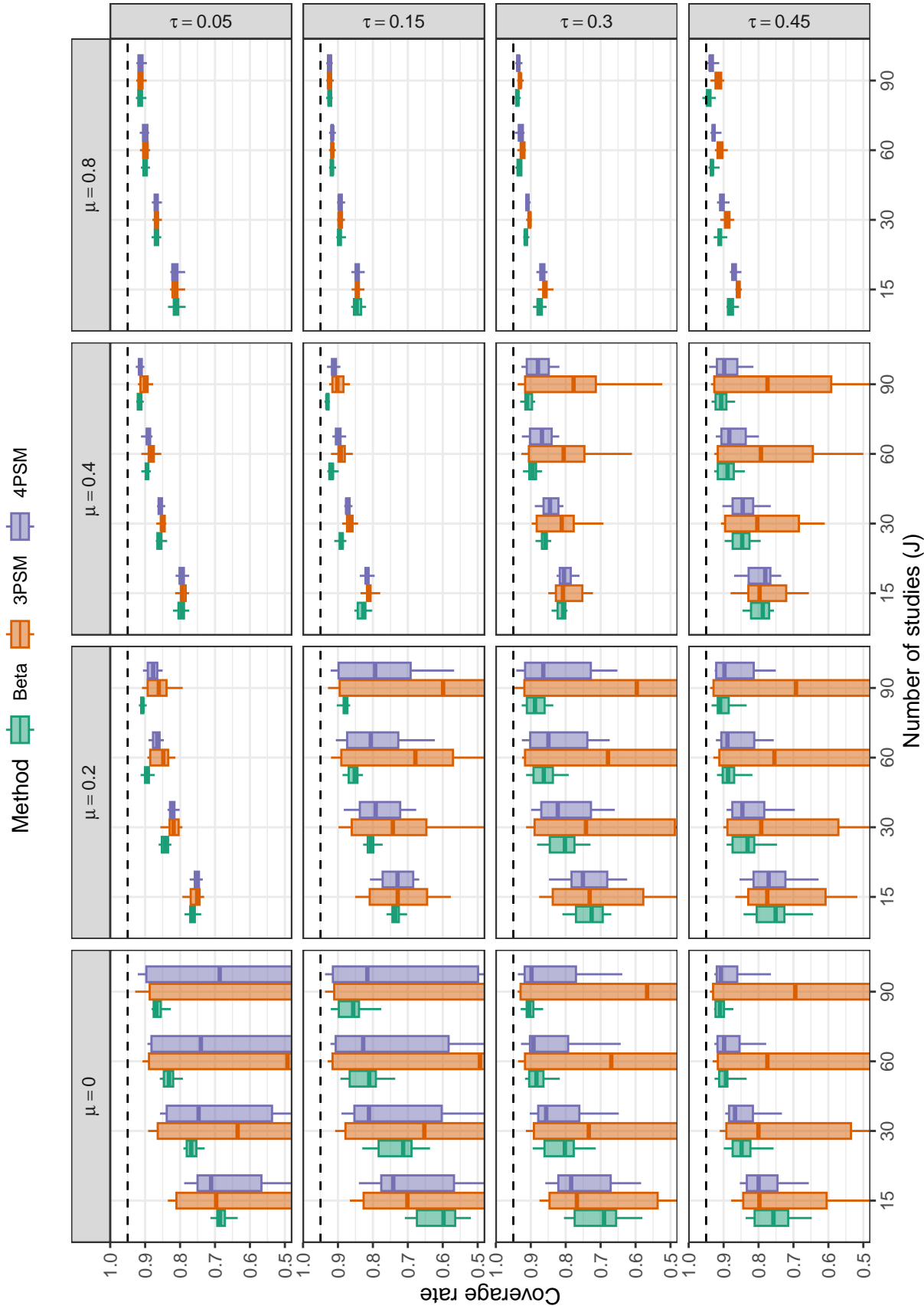### 5.2.2 *Effect Size Variance*

## 6 Discussion

### Author Contributions

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

### Funding

---

* Once again, the vertical axis of Figure 8 is restricted to the range [0.5, 1.0], and coverage rates of the intervals based on 3PSM and 4PSM are not depicted when they fall below 0.5. Supplementary Figure **??** depicts the full range of coverage rates.

**Figure 8**

*Coverage levels of confidence intervals for the average effect size based on cluster-robust variance approximations, by method, number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95. Coverage rates of the 3PSM and 4PSM intervals are not depicted when they fall below 0.5*

## Acknowledgements

### Data and Replication Materials

### Conflict of Interest Statement

### References

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

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

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

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

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

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

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

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

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

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

11. Pustejovsky JE, Citkowicz M, Joshi M. Estimation and inference for step-function selection models in meta-analysis with dependent effects. *Journal Name.* Published online 2025.

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

13. Hedges LV. Plausibility and influence in selection models: A comment on Citkowicz and Vevea (2017). *Psychological Methods.* 2017;22(1):42-46. doi:10.1037/met0000108

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

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

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

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

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

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

20. Taylor JA, Kowalski SM, Polanin JR, et al. Investigating science education effect sizes: Implications for power analyses and programmatic decisions. *AERA Open.* 2018;4(3):1-19. doi:10.1177/2332858418791991

21. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing; 2023. https://www.R-project.org/

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

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

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

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

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

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

28. Joshi M, Pustejovsky JE. *Simhelpers: Helper Functions for Simulation Studies.*; 2024. https://meghapsimatrix.github.io/simhelpers/

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

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

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

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

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

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