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 06, 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 (Carter, Schönbrodt, Gervais, & Hilgard, 2019). Extensive evidence from diverse domains suggests that such reporting biases are widespread (Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2004; Franco, Malhotra, & Simonovits, 2016; John, Loewenstein, & Prelec, 2012; Lancee, Lemmens, Kahn, Vinkers, & Luykx, 2017; e.g., O'Boyle Jr, Banks, & Gonzalez-Mulé, 2017; Pigott, Valentine, Polanin, Williams, & Canada, 2013), 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 Hedges (1992) and later extended by Vevea and Hedges (1995), 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 (Carter et al., 2019; Terrin, Schmid, Lau, & Olkin, 2003).

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 (Pustejovsky, Citkowicz, & Joshi, 2025) 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-density 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 Citkowicz and Vevea (2017), extending it to the context of dependent effect sizes by incorporating RVE and bootstrap methods. The beta-density 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-density 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, hereinafter 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 work on step-function models (Pustejovsky, Citkowicz, et al., 2025), 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 ith effect size estimate from study j, with associated standard error σ_{ij} and one-sided p-value p_{ij}. The one-sided p-value is defined relative to the null hypothesis that the true effect size is less than or equal to zero, with alternative hypothesis

that the effect size is positive. 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 σ^*, 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 τ^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 $\boldsymbol{\lambda}$.

Citkowicz and Vevea (2017) defined the selection function using a truncated beta density with two parameters, offering flexibility to capture diverse selection patterns more parsimoniously than the step function model. Because 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 α_1 and α_2 and selection parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$, the beta density selection function is expressed by

$$w(p_i^*, \boldsymbol{\lambda}) = \begin{cases} \alpha_1^{\lambda_1 - 1}(1 - \alpha_1)^{\lambda_2 - 1} & \text{if} \quad p_i^* \leq \alpha_1 \\ (p_i^*)^{\lambda_1 - 1}\,(1 - p_i^*)^{\lambda_2 - 1} & \text{if} \quad \alpha_1 < p_i^* < \alpha_2 \\ \alpha_2^{\lambda_1 - 1}(1 - \alpha_2)^{\lambda_2 - 1} & \text{if} \quad \alpha_2 \leq p_i^*. \end{cases} \tag{3}$$

Equation (3) can be written equivalently as

$$w(Y_i^*/\sigma_i^*, \boldsymbol{\lambda}) = \begin{cases} \alpha_1^{\lambda_1 - 1}(1 - \alpha_1)^{\lambda_2 - 1} & \text{if} \quad \sigma_i^* \Phi^{-1}(1 - \alpha_1) \leq Y_i^* \\ \left[\Phi\left(-Y_i^*/\sigma_i^*\right)\right]^{\lambda_1 - 1}\left[\Phi\left(Y_i^*/\sigma_i^*\right)\right]^{\lambda_2 - 1} & \text{if} \quad \sigma_i^* \Phi^{-1}(1 - \alpha_2) < Y_i^* < \sigma_i^* \Phi^{-1}(1 - \alpha_1) \\ \alpha_2^{\lambda_1 - 1}(1 - \alpha_2)^{\lambda_2 - 1} & \text{if} \quad Y_i^* \leq \sigma_i^* \Phi^{-1}(1 - \alpha_2). \end{cases} \tag{4}$$

When $\lambda_1 = \lambda_2 = 1$, the selection function is flat, the probability of selection does not depend on the p-values, and selective reporting is absent.

Citkowicz and Vevea (Citkowicz & Vevea, 2017) 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 (Hedges, 2017). 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.

Figure 1 depicts several shapes that the beta density can assume. Figure 1a. presents a curve in which there is preference for highly significant effects and the probability of selection is strictly decreasing in the one-sided p-value, where $\lambda_1 = 0.5$ and $\lambda_2 = 2.0$, and thresholds of $\alpha_1 = .025, \alpha_2 = .975$. Figure 1b. depicts a curve with $\lambda_1 = 0.5$ and $\lambda_2 = 0.6$, a

(a) *Decreasing selection:* $\lambda_1 = 0.5, \lambda_2 = 2.0$ (b) *Complex selection:* $\lambda_1 = 0.5, \lambda_2 = 0.6$

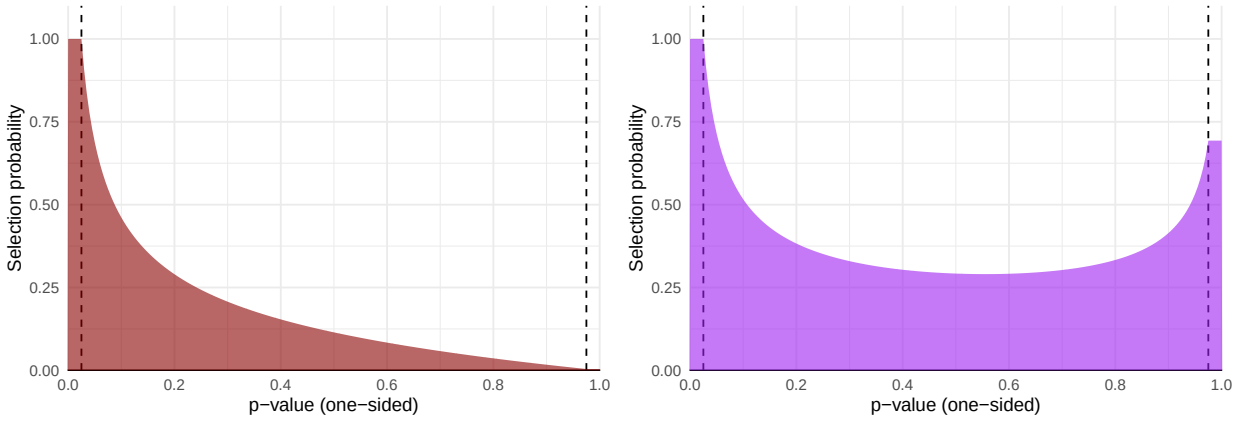

Figure 1
Examples of beta density functions

more complex form of selection in which both significantly positive and significantly negative effect sizes are more likely to be reported than null effects.

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 σ has the form

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

where

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

For the beta-density selection process, the $A(\mathbf{x}, \sigma; \boldsymbol{\beta}, \tau^2, \boldsymbol{\lambda})$ term in the beta-density composite likelihood must be computed using numerical integration. 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. 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.

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., Cox & Reid, 2004; Lindsay, 1988; Varin, 2008). 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 (Pustejovsky, Citkowicz, et al., 2025), and the exact expressions used for estimating the beta-density 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 (Xu, Gotwalt, Hong, King, & Meeker, 2020), 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 (Davison & Hinkley, 1997; Efron, 1987). 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 demonstrate the interpretation of the beta-function selection model in practice, we reanalyzed the data from a meta-analysis of the effects of science interventions on K–12 student science achievement, conducted by BSCS Science Learning (Taylor et al., 2018). 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, with positive effects corresponding to improvements in student achievement.

We conducted the analyses using R Version 4.4.3 (R Core Team, 2023). We first 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 (Viechtbauer, 2010), with standard errors clustered by study. The overall estimate of the average effect using the unadjusted model is 0.44, cluster-robust 95% CI [0.34, 0.54], which is significantly different from zero ($p < 0.001$). The unadjusted estimate of heterogeneity is $\tau = 0.38$.

To adjust for selective reporting, we used the `selection_model()` function from the `metaselection` package to fit the beta-function selection model (with default truncation thresholds of $\alpha_1 = .025, \alpha_2 = .975$) along with two step-function selection models (Pustejovsky, Joshi, & Citkowicz, 2025). Both selection models are based on an independent effects working model with random effects for each study and account for dependent effects using robust variance estimation, clustering by study. For the step-function models, we estimated both a single-step model with a threshold at $\alpha_1 = 0.025$ and a two-step model with thresholds at $\alpha_1 = .025$ and $\alpha_2 = .5$.

The upper panel of Table 1 reports parameter estimates from all three models, along with cluster-robust standard errors and 95% confidence intervals. Across the three selection models, the estimated average effect sizes are all positive and significant, but they vary in magnitude and interpretation. Compared to the unadjusted model, the average adjusted effect is 25% smaller when using the beta-density selection model ($ES = 0.33$), versus 32% larger when using the single-step model ($ES = 0.58$) and 12% larger when using the

Table 1
Selection model parameter estimates fit to science intervention effects from Taylor et al. (2018)

	Beta density		One-step		Two-step	
Parameter	Est. (SE)	95% CI	Est. (SE)	95% CI	Est. (SE)	95% CI
Summary meta-analysis						
β_0	0.33 (0.12)	[0.10, 0.56]	0.58 (0.06)	[0.46, 0.71]	0.49 (0.07)	[0.35, 0.64]
τ	0.46 (0.17)	[0.39, 0.54]	0.40 (0.21)	[0.32, 0.49]	0.43 (0.19)	[0.36, 0.52]
λ_1	1.10 (0.14)	[0.83, 1.46]	2.21 (0.26)	[1.34, 3.65]	1.93 (0.26)	[1.16, 3.22]
λ_2	1.55 (0.17)	[1.11, 2.17]			1.08 (0.43)	[0.47, 2.49]
Moderator analysis						
β_0	0.21 (0.24)	[-0.27, 0.69]	0.51 (0.11)	[0.29, 0.74]	0.40 (0.15)	[0.10, 0.70]
β_1	0.14 (0.18)	[-0.22, 0.50]	0.10 (0.13)	[-0.16, 0.35]	0.12 (0.15)	[-0.18, 0.42]
τ	0.46 (0.17)	[0.40, 0.55]	0.40 (0.21)	[0.32, 0.49]	0.43 (0.18)	[0.36, 0.52]
λ_1	1.09 (0.15)	[0.82, 1.46]	2.22 (0.26)	[1.34, 3.67]	1.93 (0.26)	[1.16, 3.22]
λ_2	1.58 (0.18)	[1.11, 2.24]			1.05 (0.43)	[0.45, 2.47]

Note:
Est. = estimate; SE = cluster-robust standard error; CI = cluster-robust confidence interval.

two-step selection model ($ES = 0.49$). These differences are due to differences in the form of the selection curve estimated from each model specification, as depicted in figure 2. The beta-density model estimates of λ_1 and λ_2 imply that positive but non-significant results are slightly more likely to be reported compared to affirmative results, but that negative effects are increasingly less likely to be be reported. On balance, this leads to an adjusted average effect size estimate that is smaller than under the unadjusted random effects model. In contrast, the step-function models both estimate that non-significant results are more likely to be reported compared to affirmative results, leading to upward adjustment in the average effect size.

A key advantage of selection models over other methods for diagnosing and adjusting for selective reporting bias is that they allow for inclusion of potential moderator variables as predictors of average effect size. This feature allows meta-analysts to distinguish selective reporting bias from systematic variation in effect sizes that could be explained by primary

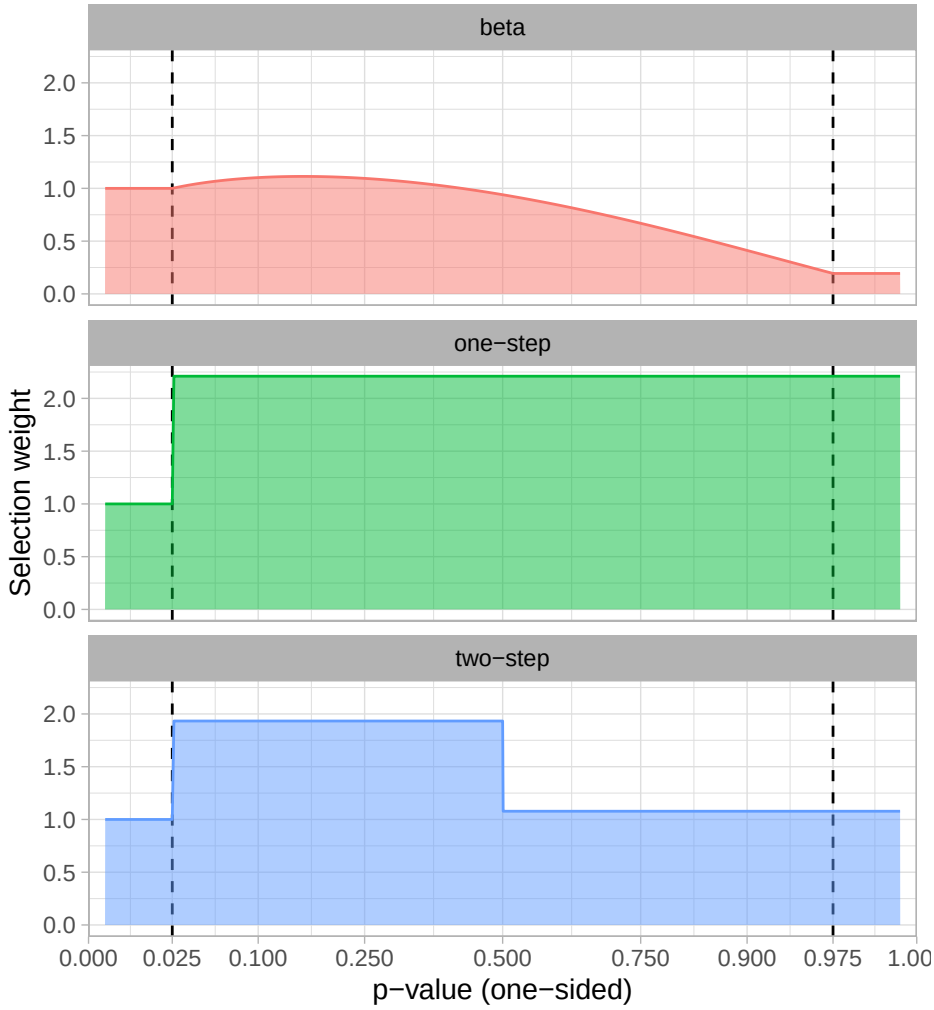

Figure 2

Estimated selection functions based on beta density, one-step, and two-step selection models

study characteristics. Past work has demonstrated that evaluations of educational interventions tend to produce larger effects on assessments developed by study authors or developers of the intervention (i.e., non-independent groups) than on assessments developed by groups independent of the evaluation (Wolf & Harbatkin, 2023). We investigated whether this pattern holds in the BSCS Science Learning meta-analysis data, which includes an indicator for assessment type, equal to one for assessments developed by non-independent groups and equal to zero otherwise.

We reran all four models using assessment type as a moderator (Taylor et al., 2018),

with assessments developed by independent groups as the reference category. Based on a random effects meta-regression without adjustment for selective reporting, assessments developed by independent groups produced an average effect size of 0.38, 95% CI [0.14, 0.62]. The average effect on assessments developed by non-independent groups was larger by 0.09 SD, 95% CI [-0.17, 0.35], which was not statistically distinct from zero ($p = 0.495$). The unadjusted estimate of heterogeneity is $\tau = 0.38$, just as in the summary meta-analysis.

The lower panel of Table 1 reports parameter estimates from the corresponding selection models that include assessment type as a predictor. The β_1 coefficient estimates represent differences in average effects between non-independent and independent assessment developers, which were all non-significant and similar in magnitude to the difference based on the unadjusted model. However, the selection models lead to different about the average effects on independently developed assessments. The estimate from the beta-density model is not statistically distinct from zero and is smaller than the estimate from the unadjusted random effects meta-regression. In contrast, the estimated average effects from the step-function selection models are both statistically significant and larger than the estimate from the unadjusted model. Selection parameter estimates are very similar to the estimates from the corresponding summary meta-analysis models.

The model estimates reveal some key differences between the selection models. 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 marginal beta-density 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 (Chen & Pustejovsky, 2024); (2) the PET-PEESE method, which addresses selective reporting and has been adapted to handle dependent data structures (Stanley & Doucouliagos, 2014); and (3) our recently developed step-function selection model that accounts for both selective reporting and dependent effects (Pustejovsky, Citkowicz, et al., 2025). 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 4.4.0 (R Core Team, 2023) using the high-throughput computing cluster at the University of Wisconsin–Madison (Center for High Throughput Computing, 2006). The code relied on several R packages, including metafor (Viechtbauer, 2010), clubSandwich (Pustejovsky, 2024), simhelpers (Joshi & Pustejovsky, 2024), optimx (Nash & Varadhan, 2011), and tidyverse (Wickham et al., 2019).

4.1 Data generation

We generated simulated data using an approach similar to Pustejovsky, Citkowicz, et al. (2025), with the key difference that effect size estimates were selected for inclusion based on the beta-density 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-density 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 ρ and standard deviation 0.05, then assumed a constant correlation between pairs of outcomes within a study.

Within each study, we simulated a study-level average effect size δ_j from a normal distribution with mean μ and variance τ_B^2, then generated individual effect size parameters from a normal distribution centered at δ_j with variance ω^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 individual effect sizes were then selected for inclusion in the dataset, with selection probabilities determined by the beta-density selection model with parameters λ_1, λ_2. We repeated this process until the simulated meta-analytic dataset included at total of J studies with at least one observed result.

4.2 Estimation methods

We estimated the beta-density 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.[1] To maintain computational feasibility, we used $B = 399$ bootstrap replications of each estimator.

We compared the performance of the beta-density selection model to three other methods. First, we estimated a summary meta-analysis model using the CHE-ISCW approach proposed by Chen and Pustejovsky (2024), which accounts for effect size dependence but does not adjust for selective reporting. This method fits a CHE working

[1] In our recent work (Pustejovsky, Citkowicz, et al., 2025), 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.

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 leads to a degree of misspecification when the average correlation used in the data-generating process differs from 0.80. We computed confidence intervals for the average effect size 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 Stanley and Doucouliagos (2014), 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 Stanley and Doucouliagos (2014), we used the PET estimate if it was not statistically distinguishable from zero at an α-level of 0.10; otherwise, we used the PEESE estimate.

Third, we estimated the step-function selection model using the CML approach described in Pustejovsky, Citkowicz, et al. (2025). 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-density selection model, the 3PSM and 4PSM are p-value selection models designed to address selective reporting, they are models for the marginal rather than joint distribution of estimates within studies, and they use cluster-robust variance estimation or bootstrapping to account for dependency. The main distinction between the step-function and beta-density selection models lies in the form of the selection mechanism. By fitting the 3PSM and 4PSM to data simulated under a beta-density 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 2. Manipulated parameters included overall average standardized mean difference (μ), between-study heterogeneity (τ_B), ratio of within- to between-study heterogeneity (ω^2/τ^2), average correlation between outcomes (ρ), probability of selection for non-affirmative results (λ_1, λ_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 of $4 \times 3 \times 2 \times 1 \times 3 \times 3 = 216$ 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.

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	0.05, 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, λ_2)	(0.01, 0.90), (0.20, 0.90), (0.50, 0.90), (0.80, 0.90), (1.00, 1.00)	(0.20, 0.90), (0.50, 0.90), (1.00, 1.00)
Number of observed studies (J)	15, 30, 60, 90	15, 30, 60

In the full simulation, we varied μ from 0.0 to 0.80, reflecting the range of effects observed in a large-scale review of education randomized controlled trials (Kraft, 2020), and τ from 0.05 (minimal heterogeneity) to 0.45 (substantial heterogeneity). Within-study heterogeneity was specified relative to between-study heterogeneity using a ratio of ω^2/τ^2 equal to 0 (no within-study heterogeneity) or 0.5.

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

Selective reporting was modeled as the probability of selecting non-affirmative results based on the parameters of the beta-density. 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 ($\lambda_1 = 0.01, \lambda_2 = 0.90$).

We varied the number of observed studies (J) from 15 to 90, covering the typical size of meta-analyses in education and psychology (Tipton, Pustejovsky, & Ahmadi, 2019). We simulated primary study sample sizes based on the 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.[2]

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 μ. 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

[2] Pustejovsky, Citkowicz, et al. (2025) included a condition in which the primary study sample 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 simulations.

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 Boos and Zhang (2000): 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-density 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-density selection model to the 3PSM and 4PSM step-function models to assess the robustness of these models to misspecification of the selection function.

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

The CHE-ISCW and PET/PEESE approaches had perfect convergence rates, producing results for every replication in every condition. The beta-density 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 A1 depicts the range of convergence rates of the beta-density selection model.

5.1.1 Bias

Figure 3 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 (τ, grid row). Each box plot summarizes variation in bias across the remaining simulation factors: the heterogeneity ratio, the

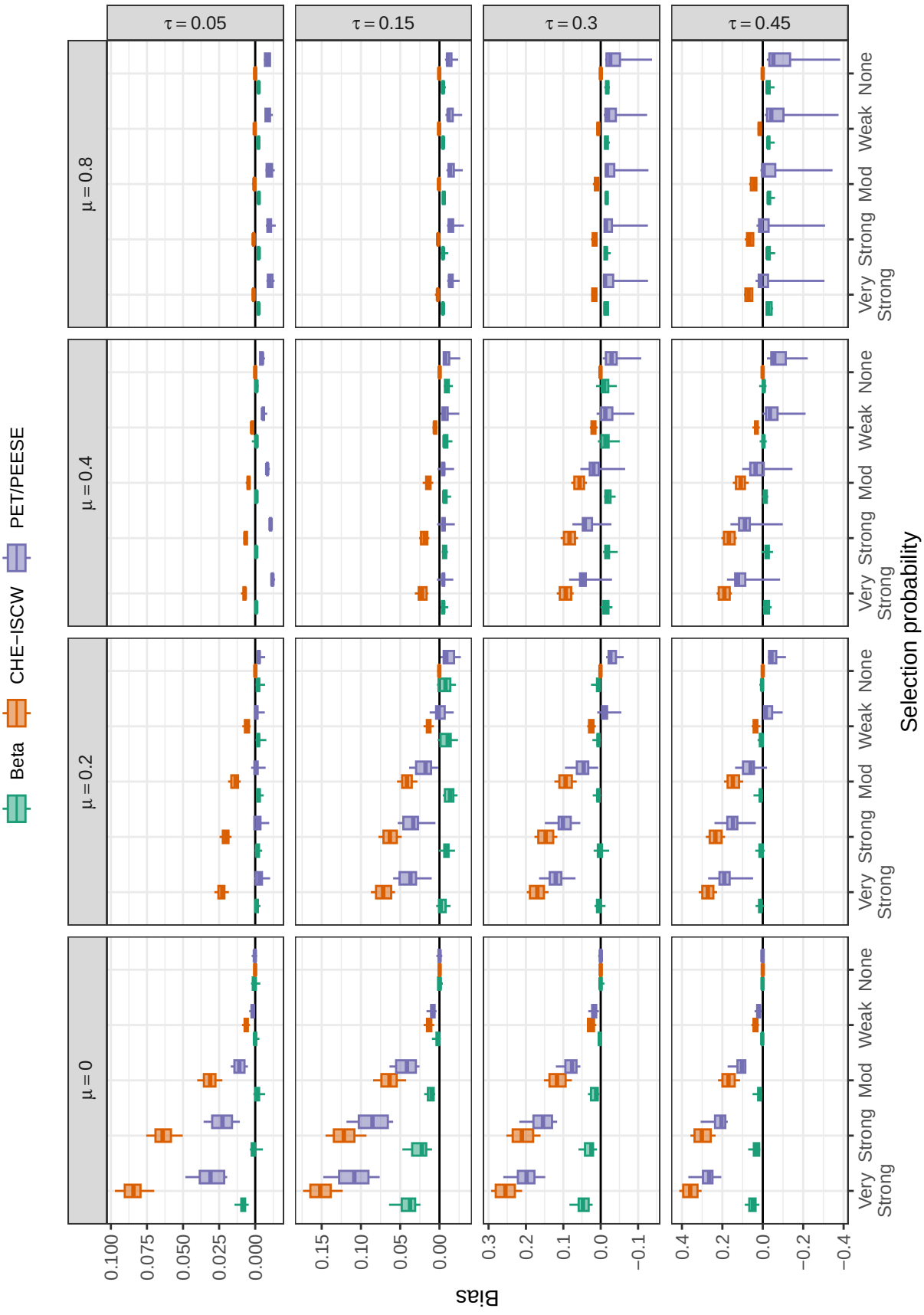

Figure 3

Bias of the average effect size by method, selection probability, average SMD, and between-study heterogeneity

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

5.1.2 Scaled RMSE

Scaled RMSE captures both bias and variability, providing an overall measure of inaccuracy. Figure 4, constructed in the same format as Figure 3, shows the scaled RMSE for each method of estimating the average effect size. Additional detail is provided in Figures A2 and A3 in Appendix 6.1, 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-density selection model generally outperforms the others—achieving lower

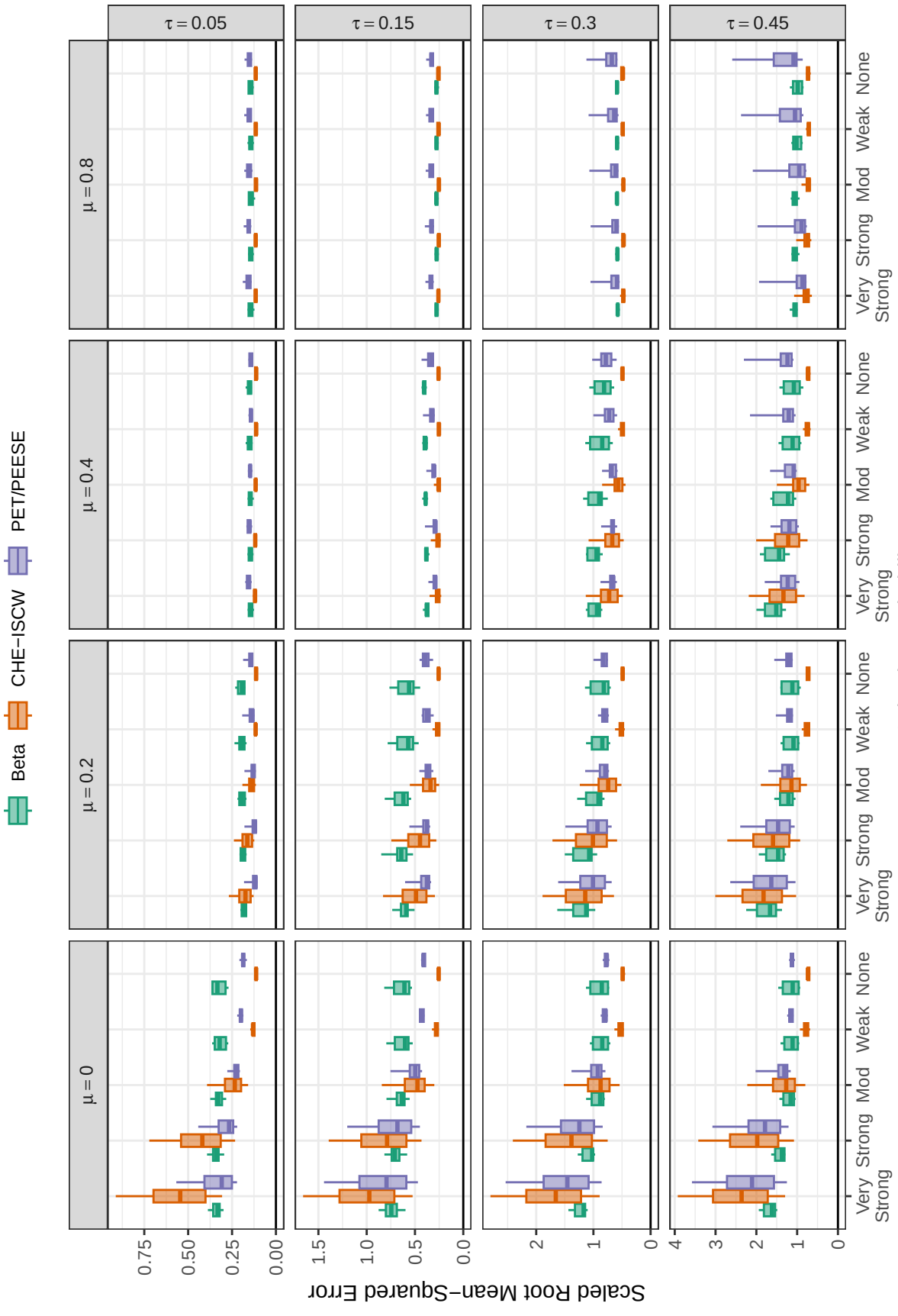

Figure 4

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

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-density 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-density 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.

5.1.3 Confidence Interval Coverage

Figure 5 shows the coverage rates of 95% confidence intervals based on large-sample cluster-robust variance approximations for the three methods.[3] Across most conditions, coverage rates fall below the nominal rate of 0.95 for all methods. However, the beta-density 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

[3] Note that the vertical axis of Figure 5 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 A4 depicts the full range of coverage rates.

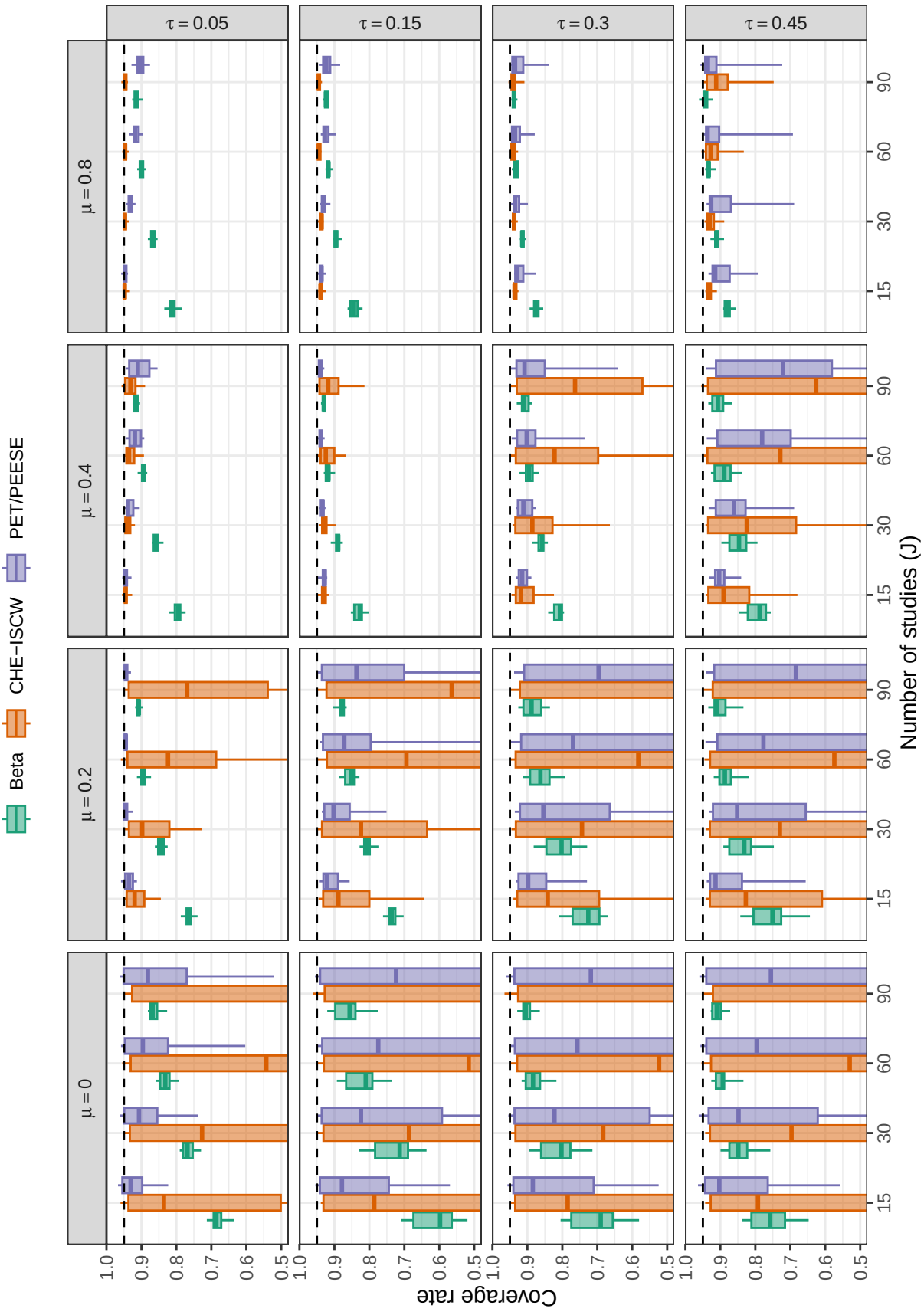

Figure 5

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

thus the interval widths——shrink, causing coverage rates to decline sharply, in some cases approaching zero.

Figure 6 depicts the coverage rates of 95% confidence intervals for the average effect size estimated using the beta-density 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.

5.2 Beta-Density Compared to One-Step and Two-Step Selection Models

Convergence rates were higher for the step-function selection models than for the beta-density selection model. Both the 3PSM and 4PSM models had convergence rates of 99% and above across all 1,280 conditions, while the beta-density selection model had convergence rates below 99% for 0 conditions.

5.2.1 Bias

Figure 7 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-density 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

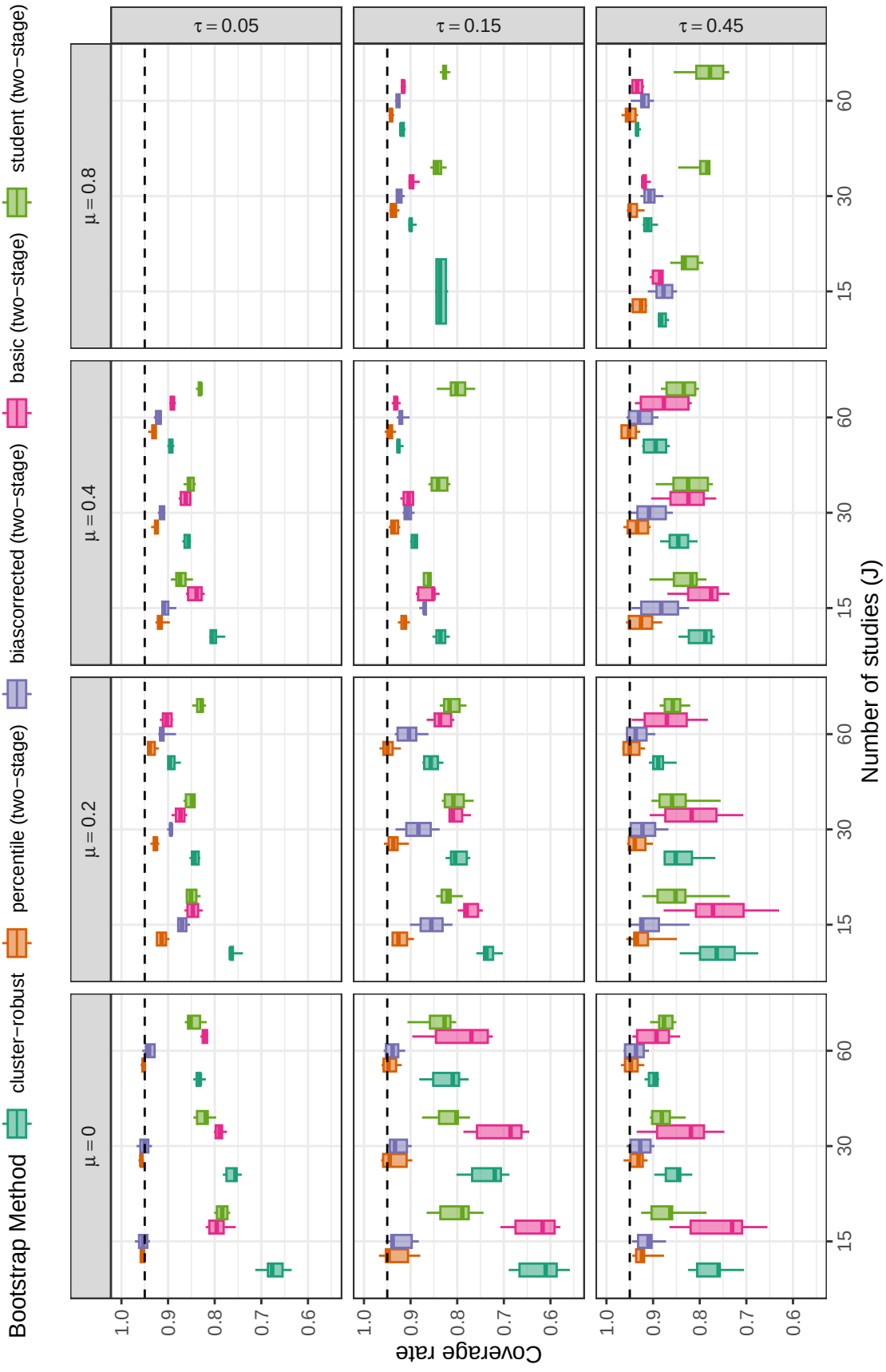

Figure 6

Coverage levels of confidence intervals for the average effect size estimated using the beta-density 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.

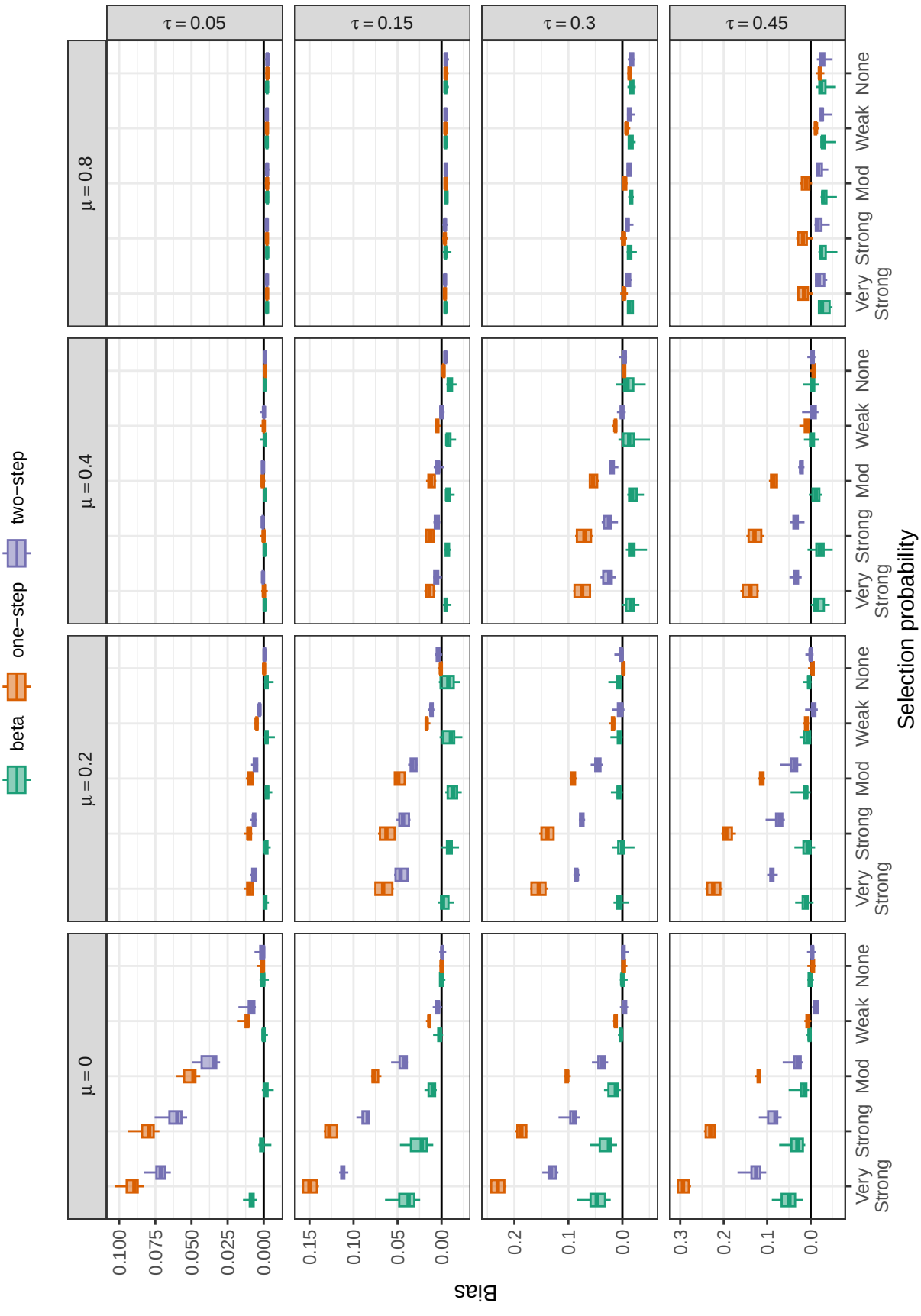

Figure 7

Bias of the average effect size by method, selection probability, average SMD, and between-study heterogeneity

data were generated under a beta-density 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 ($\mu < 0.8$).

5.2.2 Scaled RMSE

Figure 8 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-density selection model outperforms the step-function selection models, and the 4PSM performs better than the 3PSM. However, the relative performance shifts under other conditions. For instance, when average effect size is $\mu = 0.2$ and selective reporting is strong to very strong, 4PSM yields the lowest RMSE, even though it is mis-specified. 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, despite being mis-specified. Scaled RMSE is similar across all models when average effect size is moderate to large ($\mu \geq 0.4$) and heterogeneity is low ($\tau \leq 0.15$).

5.2.3 Confidence Interval Coverage

Figure 9 shows the coverage rates of 95% confidence intervals based on large-sample cluster-robust variance approximations for the three models[4] Coverage rates fall below the nominal 0.95 level for all three selection models across most conditions. However, the beta-density 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

[4] Once again, the vertical axis of Figure 9 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 B1 depicts the full range of coverage rates.

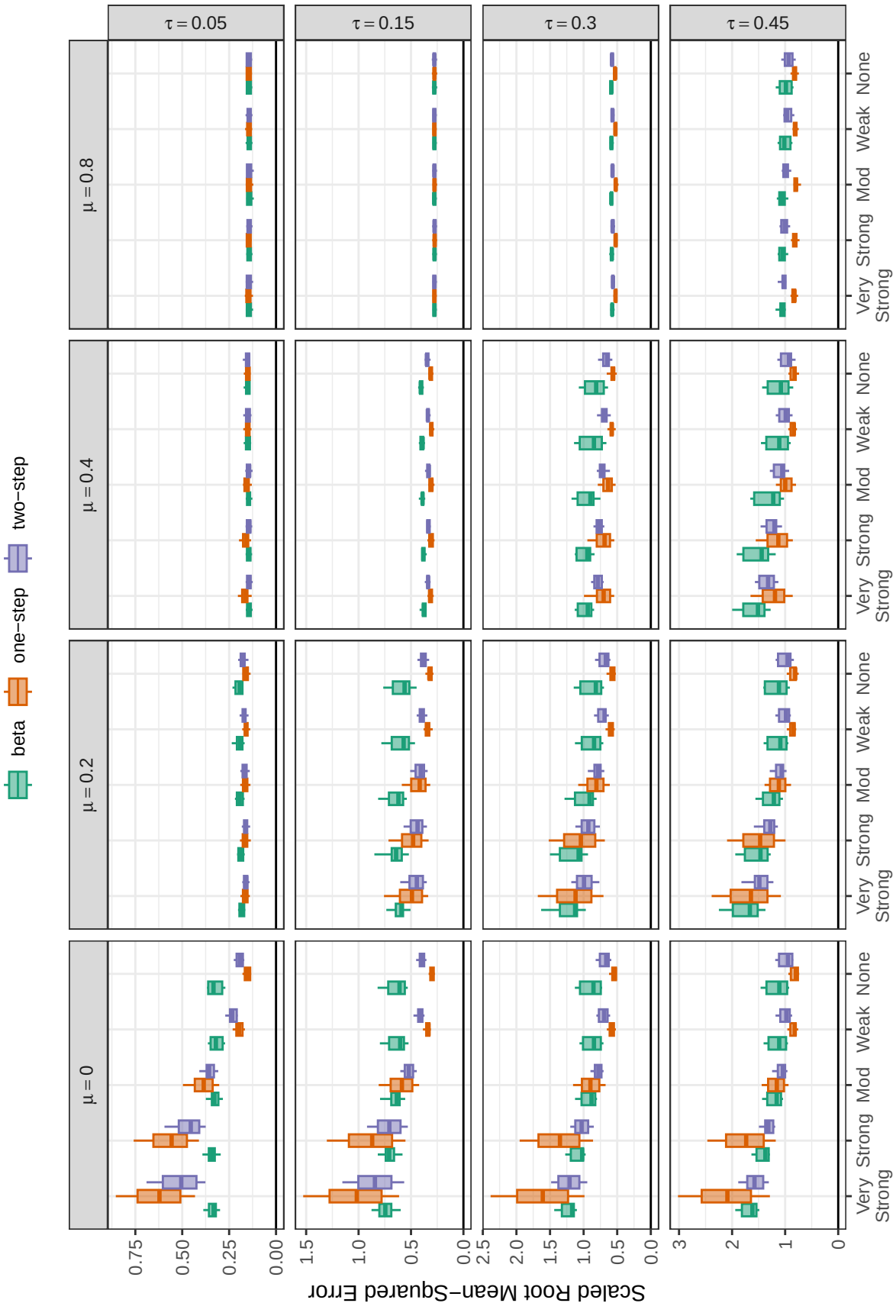

Figure 8

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

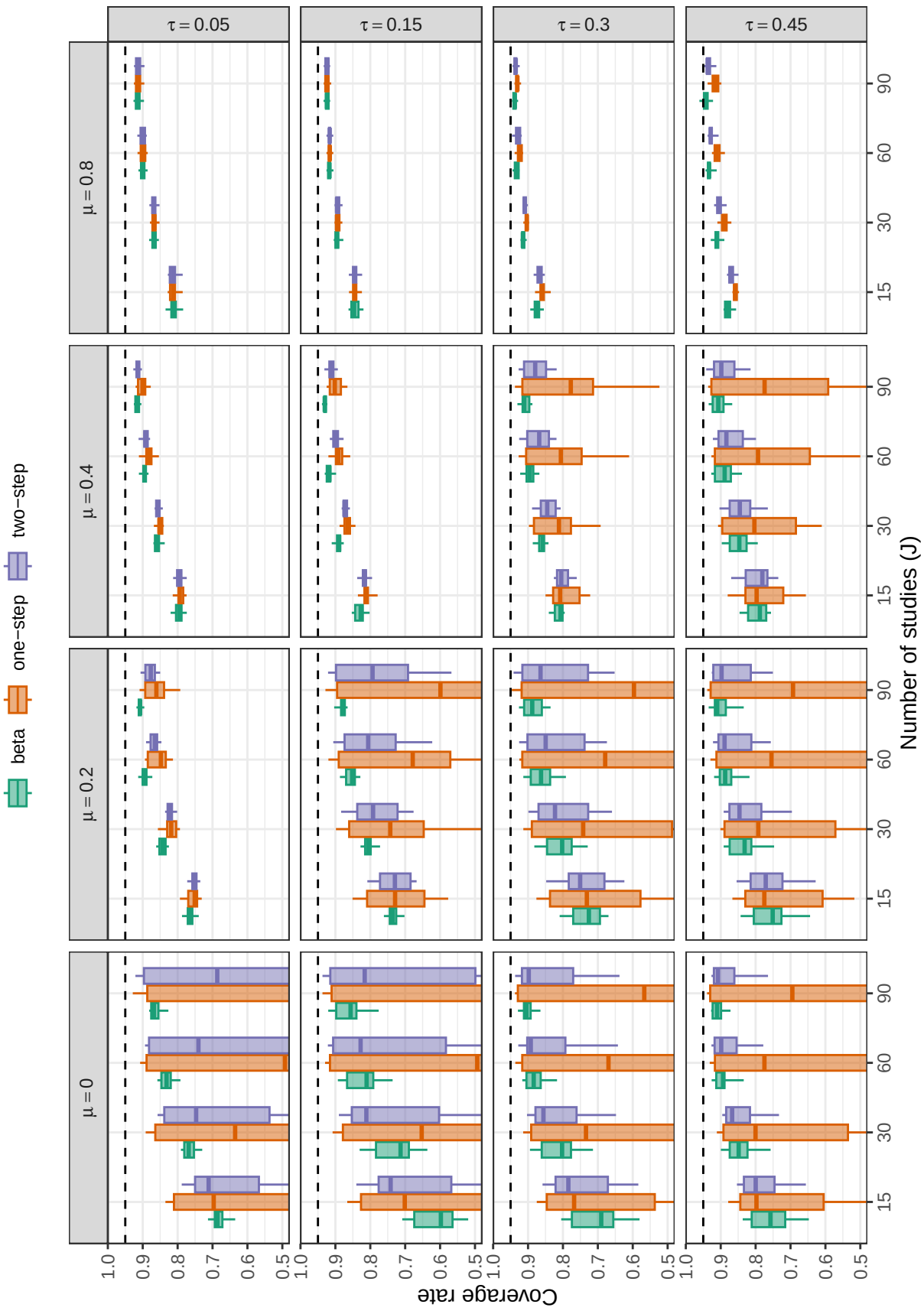

Figure 9

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

standard errors, which result in overly narrow intervals. Among the step-function selection models, coverage is generally higher for 4PSM than for 3PSM.

5.2.4 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

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

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 .

Conflict of Interest Statement

The authors declare no conflicts of interest.

References

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

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

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: Cambridge University Press.

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

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

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

Hedges, L. V. (2017). Plausibility and influence in selection models: A comment on Citkowicz and Vevea (2017). *Psychological Methods*, *22*(1), 42–46. https://doi.org/10.1037/met0000108

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*. Retrieved from https://meghapsimatrix.github.io/simhelpers/

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.

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

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

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.

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.

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

Pustejovsky, J. E., Citkowicz, M., & Joshi, M. (2025). Estimation and inference for step-function selection models in meta-analysis with dependent effects. *Journal Name*.

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*. Retrieved from https://github.com/jepusto/metaselection

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna,

Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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

Taylor, J. A., Kowalski, S. M., Polanin, J. R., Askinas, K., Stuhlsatz, M. A. M., Wilson, C. D., … Wilson, S. J. (2018). Investigating science education effect sizes: Implications for power analyses and programmatic decisions. *AERA Open*, *4*(3), 1–19. https://doi.org/10.1177/2332858418791991

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

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.

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

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., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wolf, B., & Harbatkin, E. (2023). Making sense of effect sizes: Systematic differences in intervention effect sizes by outcome measure type. *Journal of Research on Educational Effectiveness*, *16*(1), 134–161. https://doi.org/10.1080/19345747.2022.2071364

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

Appendix A

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

6.1 Additional simulation results for methods of estimating the average effect size (μ)

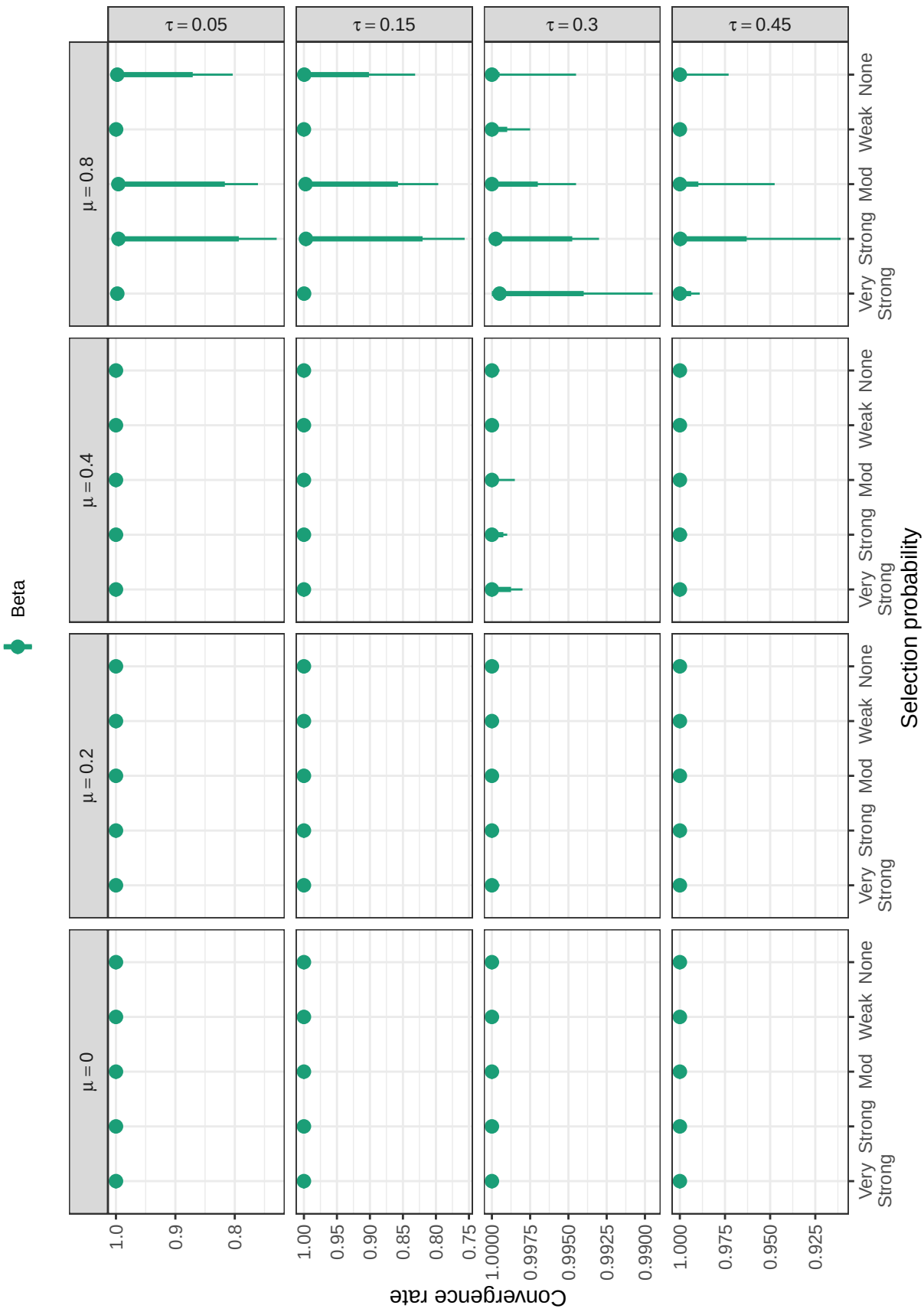

Figure A1

Convergence rates of the beta-function selection model 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.

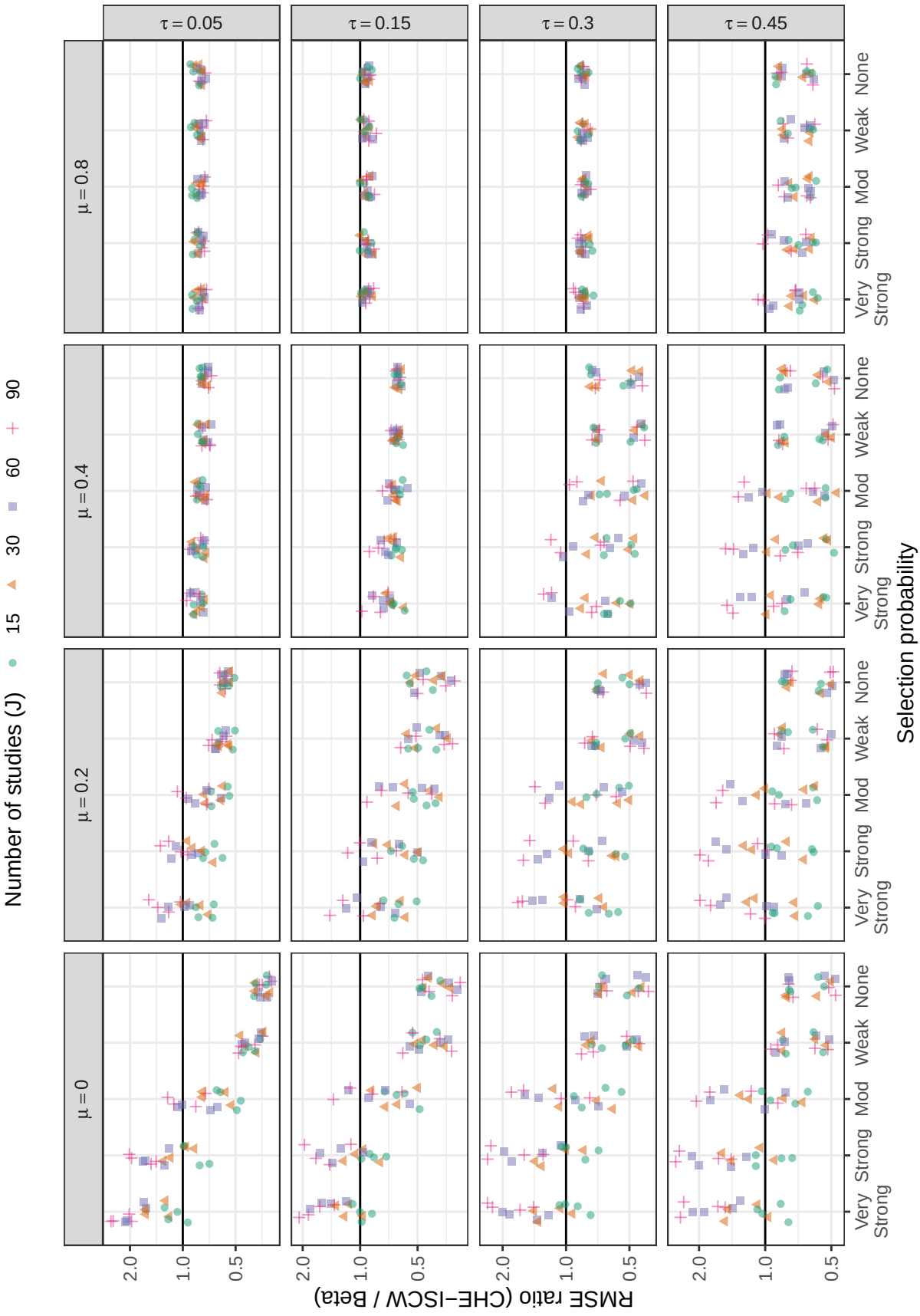

Figure A2

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

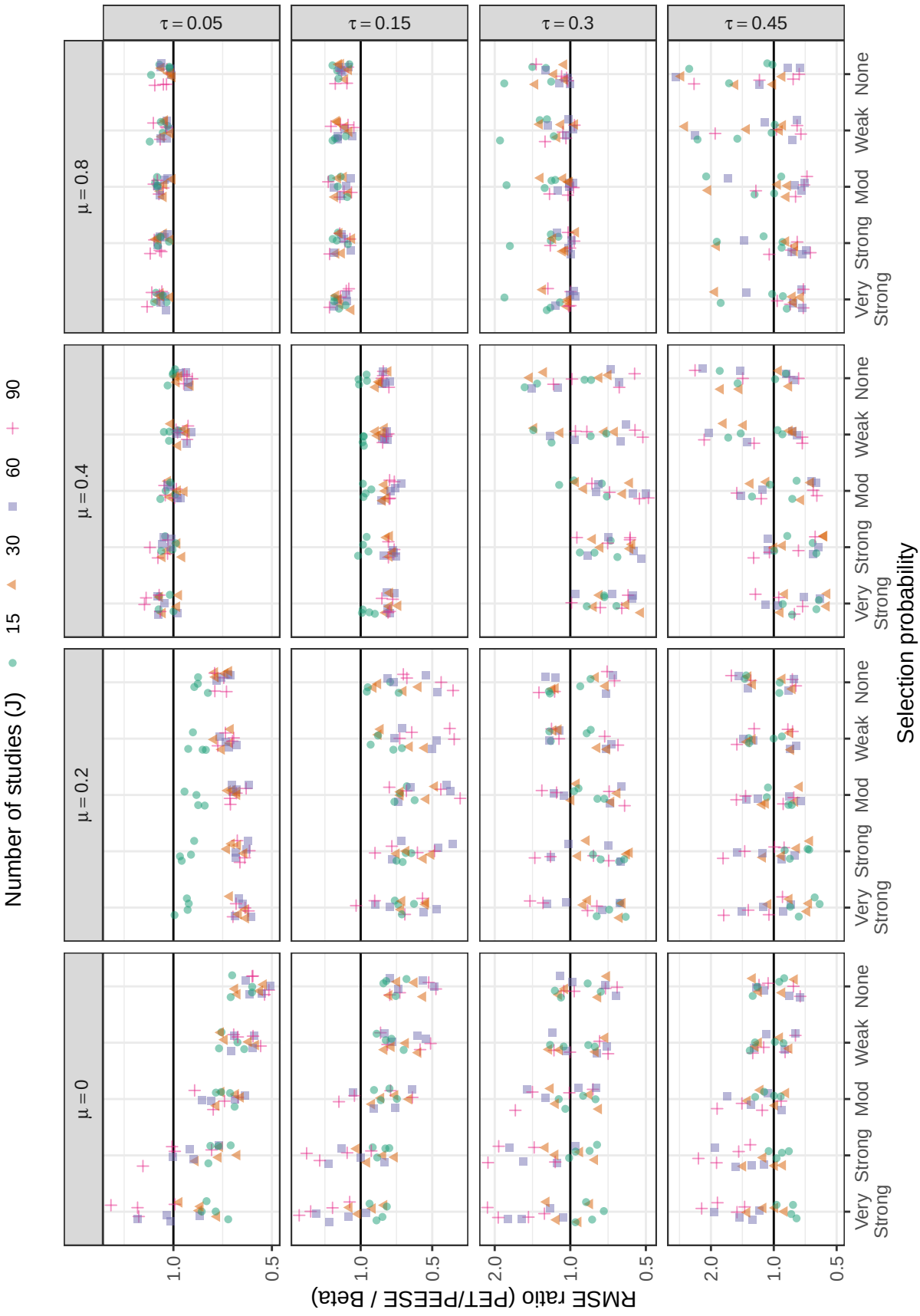

Figure A3

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

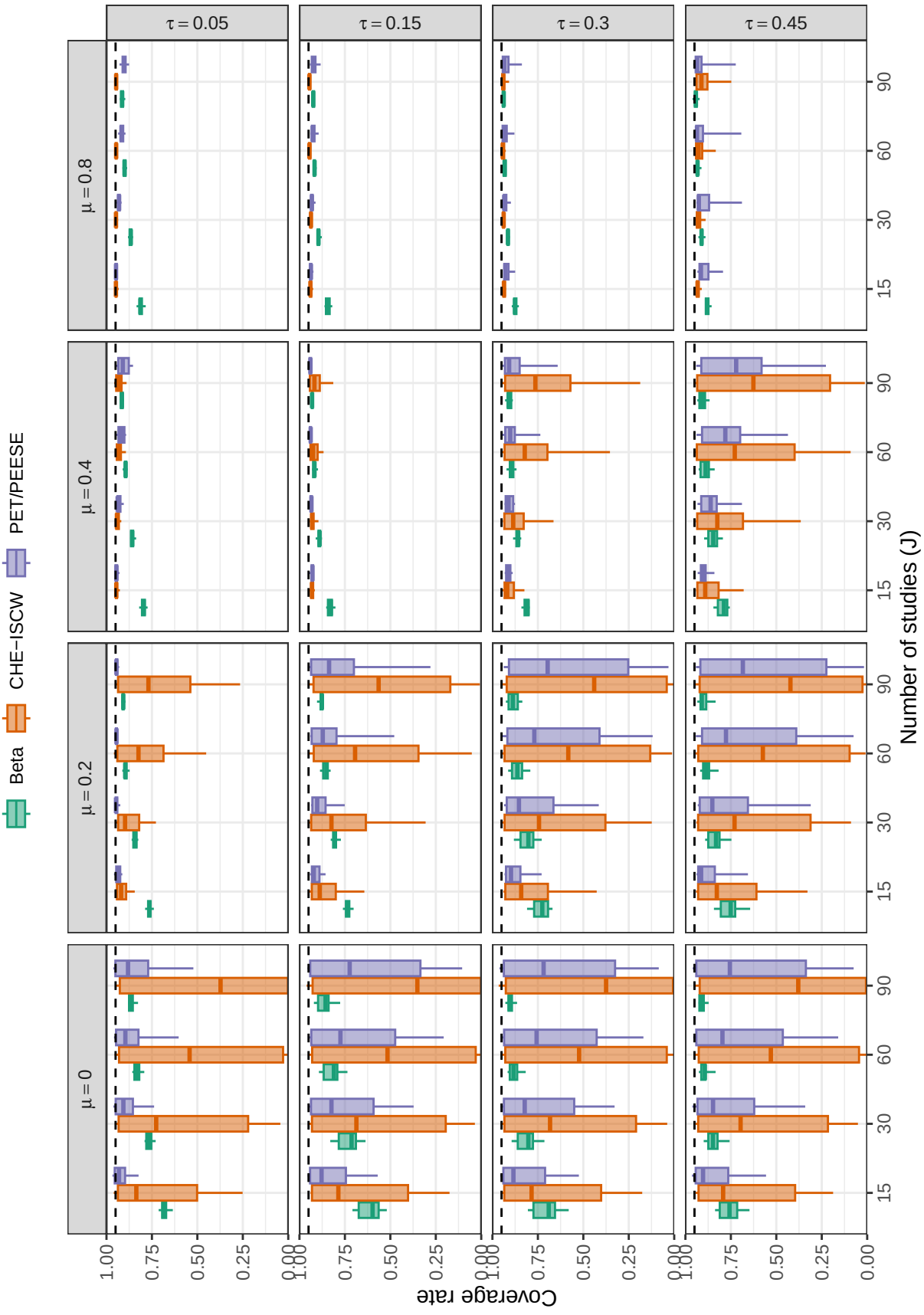

Figure A4

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.

Appendix B

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

6.2 Additional simulation results for methods of estimating the average effect size (μ)

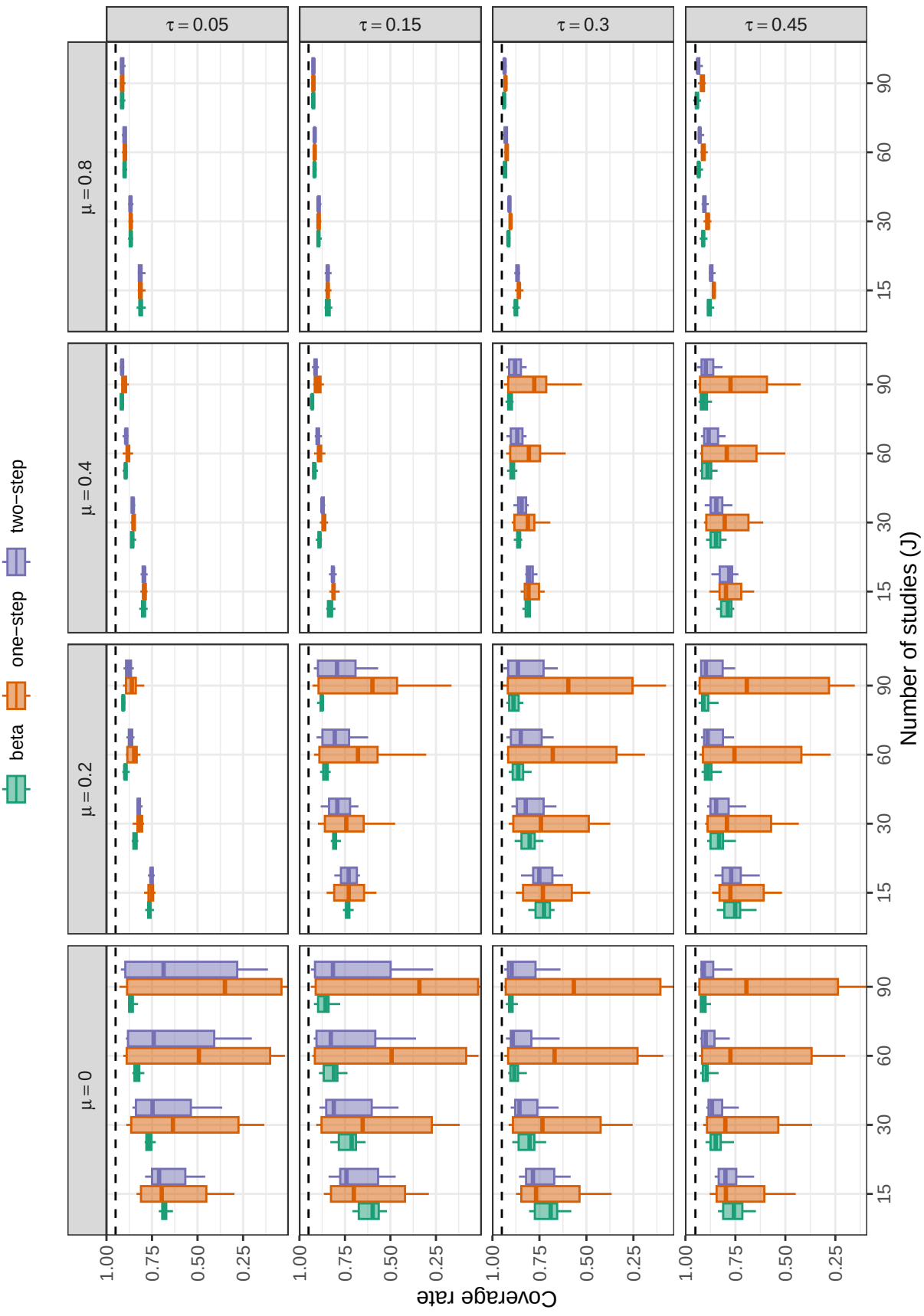

Figure B1
Coverage levels of confidence intervals for the average effect size based on cluster-robust variance approximations, by method, number of studies, average SMD, and between-study heterogeneity. Dashed lines correspond to the nominal confidence level of 0.95.