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 (Chen and Pustejovsky 2024); (2) the PET-PEESE method, which addresses selective reporting and has been adapted to handle dependent data structures (Stanley and Doucouliagos 2014); and (3) our recently developed step-function selection model that accounts for both selective reporting and dependent effects (Pustejovsky, Citkowicz, and Joshi 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.X.X (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 and Pustejovsky 2024), optimx (Nash and Varadhan 2011), nleqslv (Hasselman 2023), and tidyverse (Wickham et al. 2019).

Data generation

We generated simulated data using an approach similar to Pustejovsky, Citkowicz, and Joshi (2025), 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 ρ 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 δ_j, and 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 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 Pustejovsky, Citkowicz, and Joshi (2025) for details.

Estimation methods

We estimated the beta-function selection model using the CML approach described in Section @ref(estimation-methods). 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

[1] In our recent work (Pustejovsky, Citkowicz, and Joshi 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.

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 Chen and Pustejovsky (2024), 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 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, and Joshi (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-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.

Experimental design

We examined performance across a range of simulation conditions, summarized in Table @ref(tab:sim-design). Manipulated parameters included overall average standardized mean difference (μ), between-study heterogeneity (τ), 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 $4\times4\times2\times2\times5\times4 = 1,280$ conditions. For the more computationally intensive bootstrap simulations, we limited the design to a smaller subset ($4\times3\times2\times1\times3\times3 = 216$ 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.

```
library(kableExtra)
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
## Warning: package 'lubridate' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
```

```
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter()     masks stats::filter()
## x dplyr::group_rows() masks kableExtra::group_rows()
## x dplyr::lag()        masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
dat <- tibble(
  Parameter = c(
    "Overall average SMD ($\\mu$)",
    "Between-study heterogeneity ($\\tau$)",
    "Heterogeneity ratio ($\\omega^2 / \\tau^2$)",
    "Average correlation between outcomes ($\\rho$)",
    "Probability of selection for non-affirmative effects ($\\lambda_1, \\lambda_2$)",
    "Number of observed studies ($J$)"
  ),
  `Full Simulation` = c(
    "0.0, 0.2, 0.4, 0.8",
    "0.05, 0.15, 0.30, 0.45",
    "0.0, 0.5",
    "0.40, 0.80",
    #"(d1 = 0.01, d2 = 0.90), (d1 = 0.20, d2 = 0.90), (d1 = 0.50, d2 = 0.90), (d1 = 0.80, d2 = 0.90), (
    "(0.01, 0.90), (0.20, 0.90), (0.50, 0.90), (0.80, 0.90), (1.00, 1.00)",
    "15, 30, 60, 90"
  ),
  `Bootstrap Simulation` = c(
    "0.0, 0.2, 0.4, 0.8",
    "0.05, 0.15, 0.45",
    "0.0, 0.5",
    "0.80",
    #"(d1 = 0.20, d2 = 0.90), (d1 = 0.50, d2 = 0.90), (d1 = 1.00, d2 = 1.00)",
    "(0.20, 0.90), (0.50, 0.90), (1.00, 1.00)",
    "15, 30, 60"
  )
)

kable(
  dat,
  caption = "Parameter values examined in the simulation study",
  booktabs = TRUE,
  escape = FALSE
) |>
  kable_styling() |>
  column_spec(1, width = "2.5in")
```

Table 1: Parameter values examined in the simulation study

Parameter	Full Simulation	Bootstrap Sim
Overall average SMD (μ)	0.0, 0.2, 0.4, 0.8	0.0, 0.2, 0.4, 0
Between-study heterogeneity (τ)	0.05, 0.15, 0.30, 0.45	0.05, 0.15, 0.4
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), (
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 (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 (Tipton, Pustejovsky, and Ahmadi 2019).

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.[2]

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 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 $B = 1999$.

Boos, Dennis D., and Ji Zhang. 2000. "Monte Carlo Evaluation of Resampling-Based Hypothesis Tests." *Journal of the American Statistical Association* 95 (450): 486–92. https://doi.org/10.1080/01621459.2000.10474226.

Center for High Throughput Computing. 2006. "Center for High Throughput Computing." Center for High Throughput Computing. https://doi.org/10.21231/GNT1-HW21.

[2] 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 simulations.

Chen, Man, and James E. Pustejovsky. 2024. "Adapting Methods for Correcting Selective Reporting Bias in Meta-Analysis of Dependent Effect Sizes." https://doi.org/10.31222/osf.io/jq52s.

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

Joshi, Megha, and James E. Pustejovsky. 2024. *Simhelpers: Helper Functions for Simulation Studies*. https://meghapsimatrix.github.io/simhelpers/.

Kraft, Matthew A. 2020. "Interpreting Effect Sizes of Education Interventions." *Educational Researcher* 49 (4): 241–53.

Nash, John C., and Ravi Varadhan. 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.

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

Pustejovsky, James E., Martyna Citkowicz, and Megha Joshi. 2025. "Estimation and Inference for Step-Function Selection Models in Meta-Analysis with Dependent Effects." *Journal Name*.

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

Stanley, Tom D, and Hristos Doucouliagos. 2014. "Meta-Regression Approximations to Reduce Publication Selection Bias." *Research Synthesis Methods* 5 (1): 60–78.

Tipton, Elizabeth, James E Pustejovsky, and Hedyeh Ahmadi. 2019. "Current Practices in Meta-Regression in Psychology, Education, and Medicine." *Research Synthesis Methods* 10 (2): 180–94.

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

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.