Models for Examining Selective Reporting in Meta-Analysis with Dependent Effects

James E. Pustejovsky, Associate Professor, University of Wisconsin-Madison

Martyna Citkowicz, Principal Researcher, AIR

Megha Joshi, Quantitative Researcher, AIR

SREE| September 2024

Acknowledgement

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through grant R305D220026 to the American Institutes for Research. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Selective Reporting

- **Selective reporting** occurs if *affirmative* findings are **more likely to be reported** and available for inclusion in meta-analysis
- Selective reporting can **distort the evidence base** available for systematic review/meta-analysis
  - Inflate average effect size estimates from meta-analysis
  - Bias estimates of heterogeneity (Augusteijn et al., 2019)

Dependent Effect Sizes

- Dependent effect sizes are **ubiquitous** in education and social science meta-analyses
- But most methods to examine selective outcome reporting bias do not account for **effect size dependency**
- Failing to account for dependency can result in **misleading conclusions** like too-narrow confidence intervals, and inflated Type 1 error rates

Treatment	O P Q
Control	O P Q

d_{O1}
d_{P1}
d_{Q1}

Treatment	O_1	O_2	O_3
Control	O_1	O_2	O_3

d_{12}
d_{22}
d_{32}

Treatment T	O
Treatment U	O
Control	O

d_{T3}
d_{U3}

Our Project

- Develop methods for investigating selective reporting in meta-analysis that **account for dependent effect sizes**
- P-value selection models (Hedges, 1992; Vevea & Hedges, 1995)
  - Estimate the model ignoring dependence structure

    Use robust variance estimation or clustered bootstrapping to account for dependence

P-value selection models

- The model for the evidence-generating process (prior to selective reporting):

Effect size *i* in study *j*

$$T_{ij} \sim N\left(x_{ij}\beta, \tau^2 + \sigma_{ij}^2\right)$$

Sampling variance of effect size *i* in study *j*

Heterogeneity variance of effect sizes

Meta-regression model for average effect size

- A step-function model for the selective reporting process:

$$\Pr\left(T_{ij} \text{ is observed}\right) \propto \begin{cases} 1 & \text{if} \quad p_{ij} < \alpha_1 \\ \lambda_1 & \text{if} \quad \alpha_1 \leq p_{ij} < \alpha_2 \\ \lambda_2 & \text{if} \quad \alpha_2 \leq p_{ij} \end{cases}$$

- https://jepusto.com/posts/step-function-selection-models/

Estimation strategies

- Estimate selection models while ***ignoring the dependency structure***
  - Maximum Likelihood (ML) estimation for all parameters.
  - A hybrid approach: Estimate selection parameters with ML score equation, estimate evidence-generating process using inverse-probability of selection weighting.

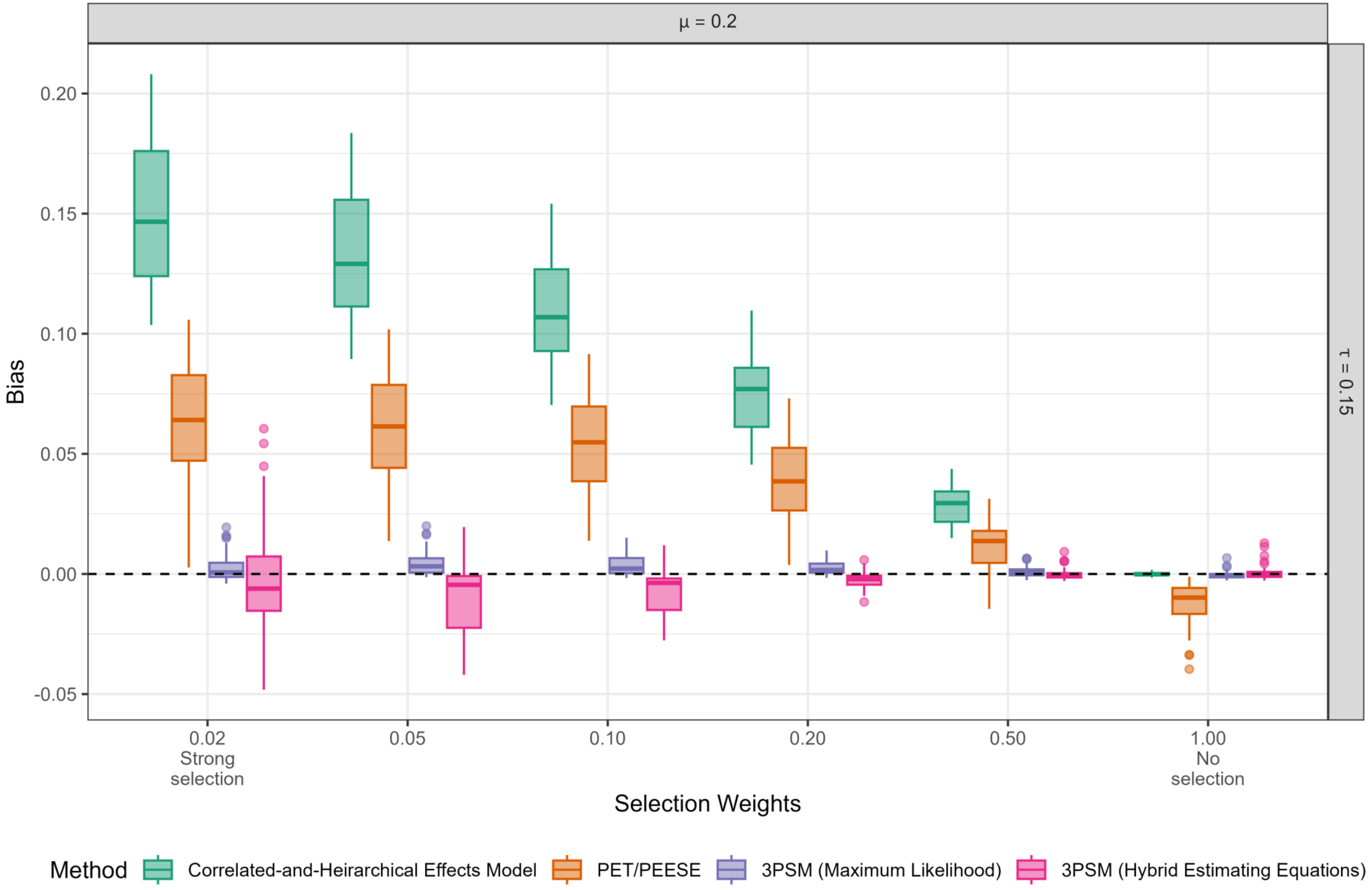
- Account for dependent effect sizes
  - Robust variance estimation ("independent effects" working model)
  - Cluster-wise bootstrapping (multinomial or fractional weighted bootstrap)
    - Percentile, "basic", or studentized confidence intervals

Simulation Study

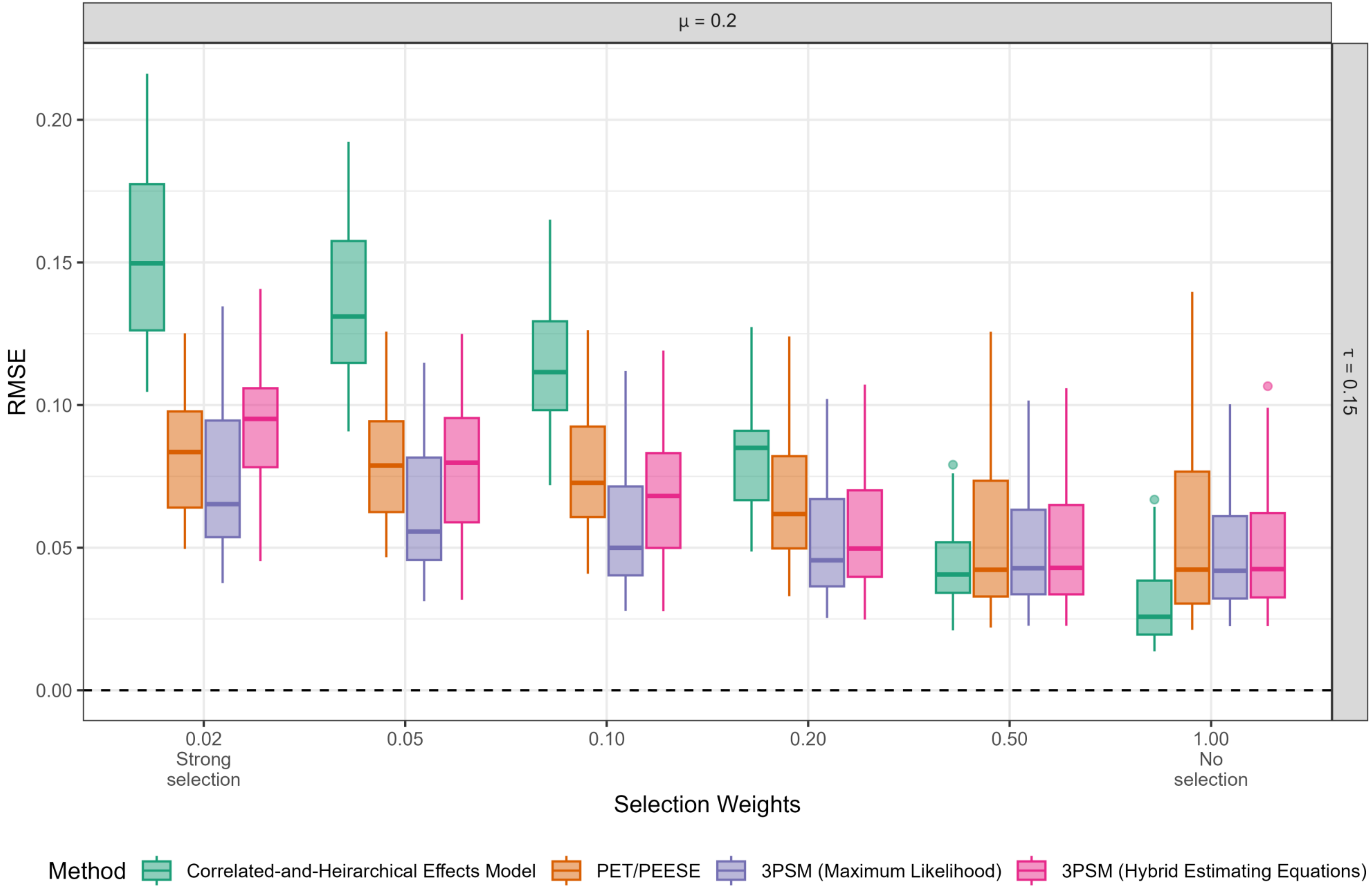
Data Generation

- Generated summary statistics for **correlated outcomes** for two-group comparison designs (equal sample size)
- Generated meta-analytic data
- Censored one-sided p-values > 0.025 with **specified probability of selection**
- Continued sampling until dataset included effect sizes from m studies.

Bias

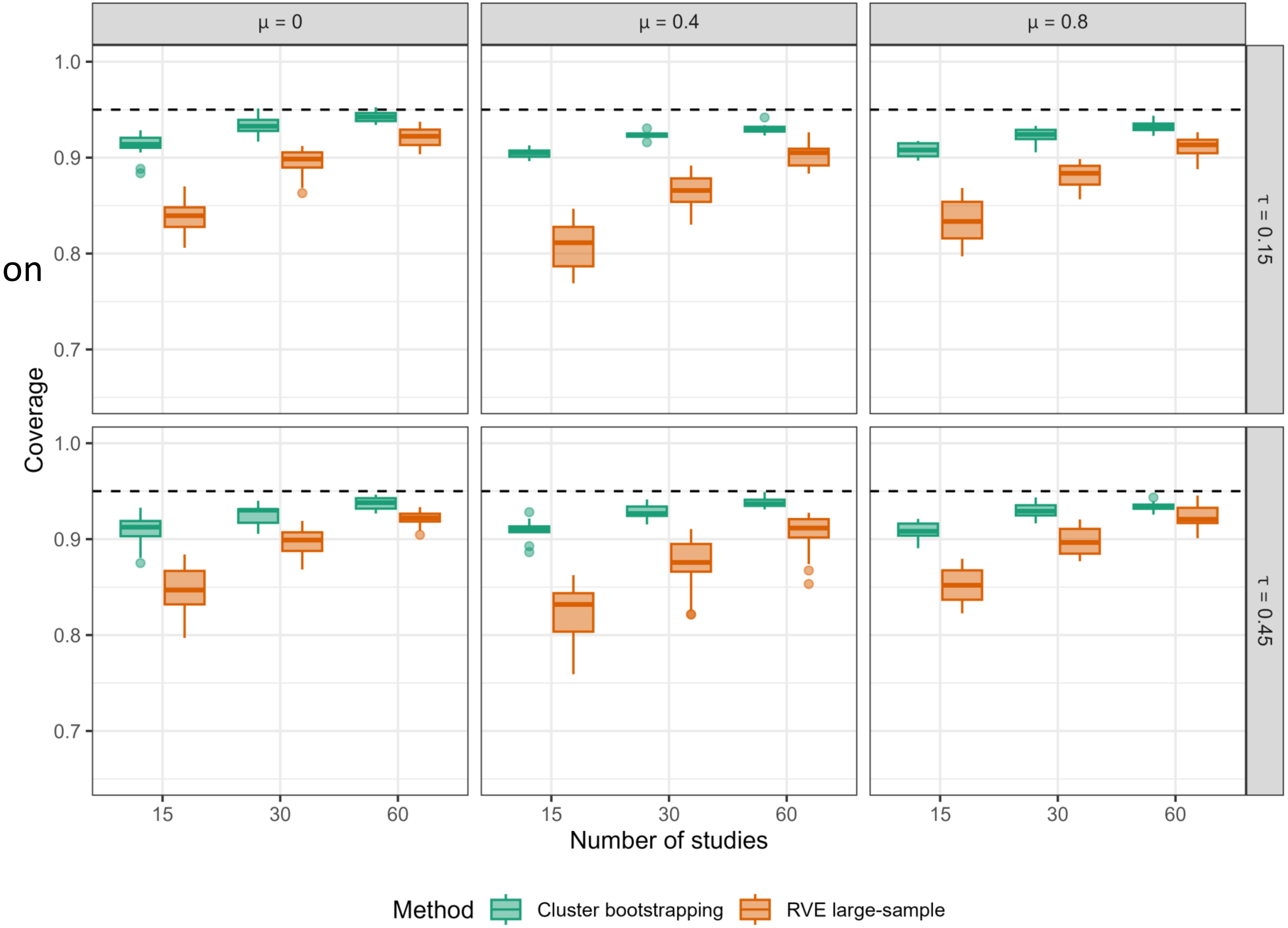

μ = 0.2
τ = 0.15
Bias
0.20
0.15
0.10
0.05
0.00
-0.05
0.02
Strong selection
0.05
0.10
0.20
0.50
1.00
No selection
Selection Weights
Method
Correlated-and-Heirarchical Effects Model
PET/PEESE
3PSM (Maximum Likelihood)
3PSM (Hybrid Estimating Equations)

Root Mean Squared Error



Coverage

- 3-parameter selection model
- Maximum likelihood estimation



AIR MOSAIC Methods of Synthesis and Integration Center

Empirical application

Meta-Analysis of Randomized Trials with Follow-Up

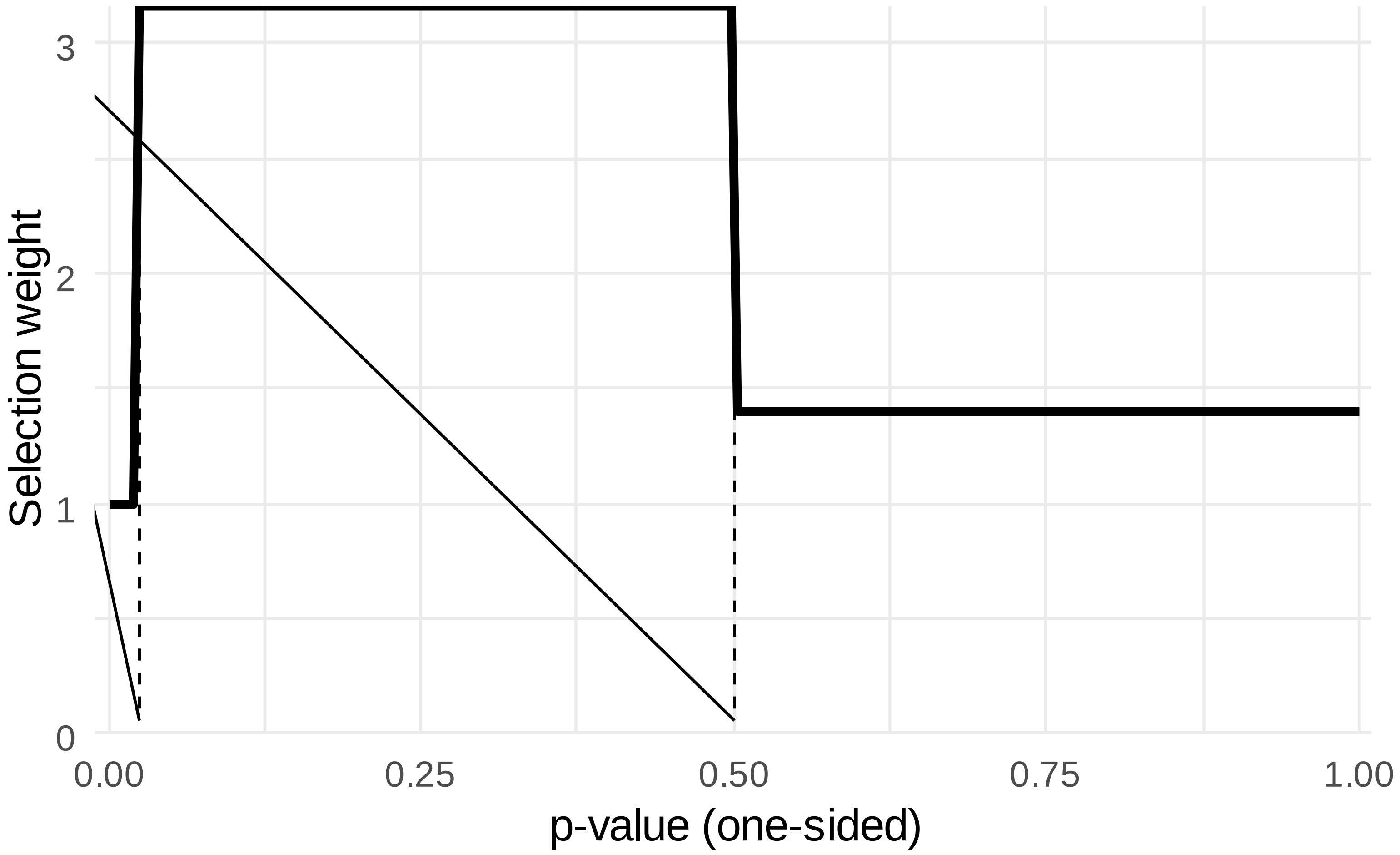
- Meta-analysis of Randomized Control Trials with Follow-up (MERF; Hart et al., 2023)
  - Educational RCTs that include both immediate post-intervention assessments and longer-term follow-up assessments of cognitive or socio-emotional outcomes
  - A total 68 studies with 450 outcomes assessed at multiple time points
- Ran two models and compared the estimates to the original analysis
  - Original model: Correlated and hierarchical effects; no adjustment for selective reporting
  - Random effects model: Independent effects with clustered SEs; no adjustment for selective reporting
  - P-value selection model: Independent effects with clustered SEs; adjusts for selective reporting using a step-function selection model with thresholds at $\alpha = .025$ and $\alpha = .500$

Post-Intervention Effects

Parameter	Original Analysis Est [95%CI]	Random Effects Model Est [95%CI]	P-value Selection Model Est [95%CI]
Average ES: Cognitive	0.341 [0.241, 0.440]	0.360 [0.253, 0.468]	0.479 [0.264, 0.714]
Average ES: Social-Emotional	0.163 [0.085, 0.240]	0.157 [0.071, 0.242]	0.215 [0.022, 0.421]
Heterogeneity SD	0.367 [0.328, 0.412]	0.358 [0.325, 0.395]	0.455 [0.237, 0.610]
Selection weight $.025 < \alpha < .500$			3.164 [1.475, 5.997]
Selection weight $.500 < \alpha < 1.00$			1.395 [0.506, 3.822]

Relative Likelihood of Being Published for Post-Intervention Effects

- Studies with 0.025 < p-value < 0.500 are 3 times as likely to be published than studies with p-values < 0.025
- Studies with p-value > 0.500 are nearly 1.4 times as likely to be published than studies with p-values < 0.025

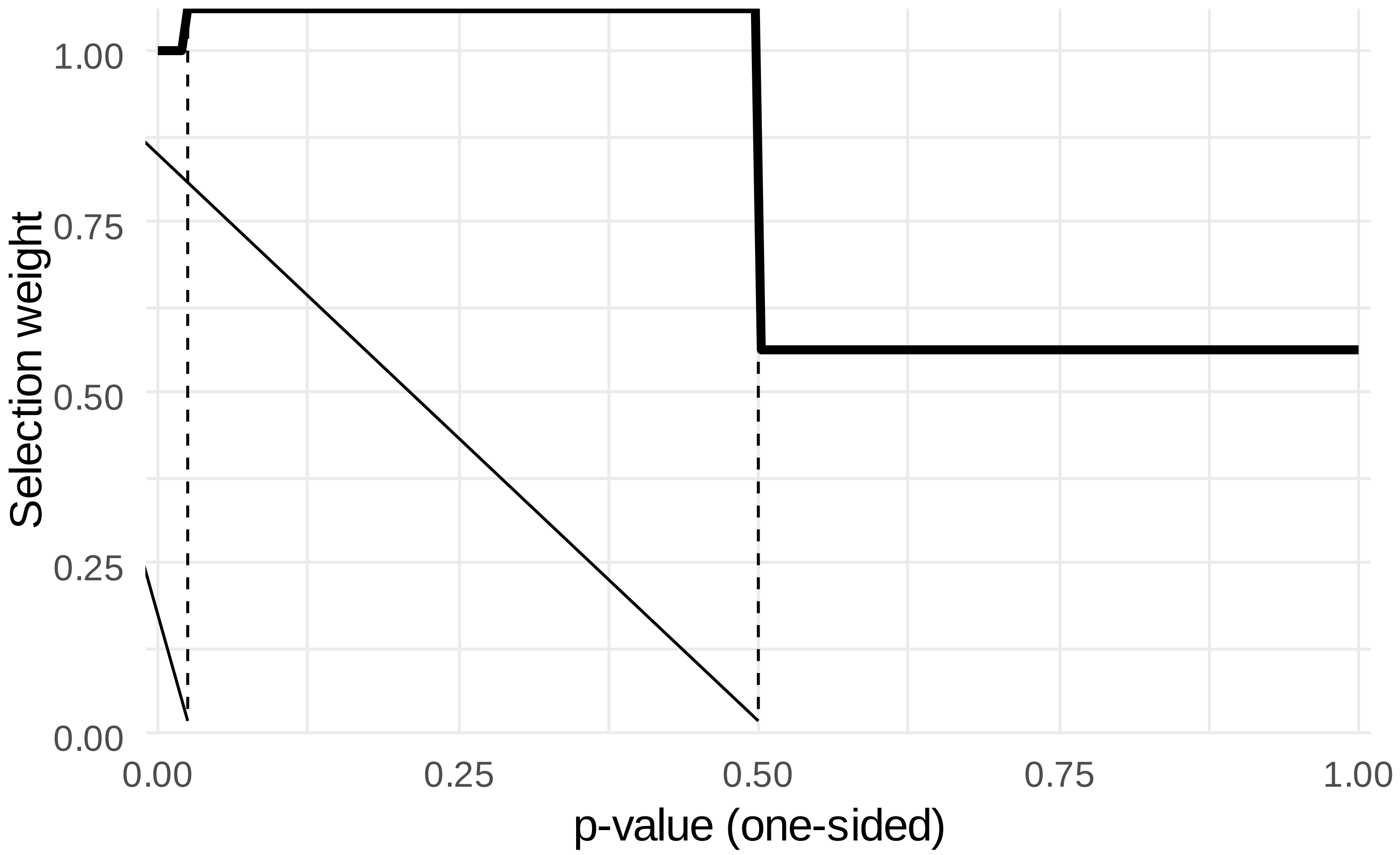


Follow-Up Effects Conditional on Post-Intervention Effects

Parameter	Original Analysis Est [95%CI]	Random Effects Model Est [95%CI]	P-value Selection Model Est [95%CI]
Intercept: Cognitive	-0.003 [-0.059, 0.052]	0.019 [-0.043, 0.081]	-0.012 [-0.056, 0.105]
Intercept: Social-Emotional	0.054 [0.018, 0.090]	0.042 [0.011, 0.072]	0.014 [-0.030, 0.093]
Slope: Cognitive	0.462 [0.333, 0.592]	0.456 [0.383, 0.528]	0.476 [0.385, 0.606]
Slope: Social-Emotional	0.429 [0.258, 0.600]	0.563 [0.298, 0.827]	0.590 [0.302, 0.876]
Heterogeneity SD	0.095 [0.000, 0.123]	0.067 [0.044, 0.093]	0.071 [0.000, 0.200]
Selection weight .025 < α < .500			1.064 [0.622, 2.562]
Selection weight .500 < α < 1.00			0.563 [0.260, 1.630]

Relative Likelihood of Being Published for Follow-Up Effects

- Studies with 0.025 < p-value < 0.500 are 1.1 times as likely to be published than studies with p-values < 0.025
- Studies with p-value > 0.500 are about half as likely to be published than studies with p-values < 0.025



Conclusion

Conclusions & Limitations

- Cluster-bootstrapped selection model
  - Low bias compared to other selective reporting adjustment methods
  - Bias variance trade-off relative to a regular meta-analytic model
  - Tolerable coverage (between 90% to 95%)
- Results are specific to the **selection mechanism** we studied.

Further Directions

- Examining the consequences of selective reporting in education
  - Stratified sample of ~100 published meta-analyses of education research
  - Meta-analytic effects are frequently adjusted upwards
- Developing a second p-value selection model using the beta density
  - Extending the work of Citkowicz and Vevea (2017)

R package `metaselection`

- Currently available on Github at https://github.com/jepusto/metaselection
- Install using

  ```
  remotes::install_github("jepusto/metaselection")
  ```

- Under active development, feedback very welcome!

James Pustejovsky

Associate Professor

pustejovsky@wisc.edu

Martyna Citkowicz

Principal Researcher

mcitkowicz@air.org

Megha Joshi

Quantitative Researcher

mjoshi@air.org

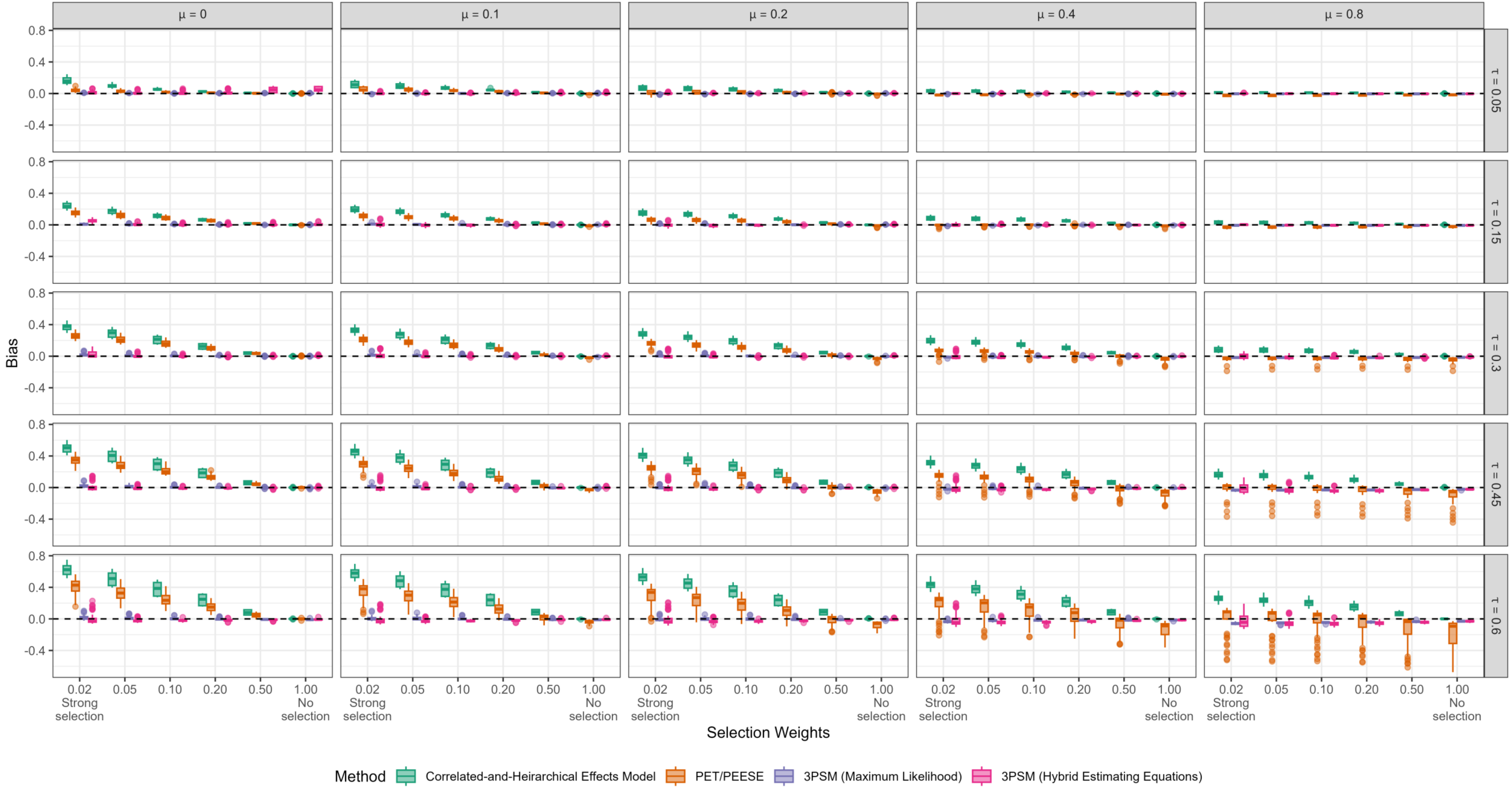
AMERICAN INSTITUTES FOR RESEARCH® | AIR.ORG/MOSAIC

Supplement

Experimental Design

Conditions	Full simulation	Bootstrap CI simulation
Overall average SMD (μ)	0.0, 0.1, 0.2, 0.4, 0.8	0.0, 0.4, 0.8
Between-study heterogeneity (τ)	0.05, 0.15, 0.30, 0.45, 0.60	0.15, 0.45
Heterogeneity ratio (ω^2/τ^2)	0.0, 0.5	0.0, 0.5
Average correlation between outcomes (ρ)	0.4, 0.8	0.8
Weights for censoring (probability of selection for non-significant ES)	0.02, 0.05, 0.1, 0.2, 0.5, 1.0	0.05, 0.2, 1.0
Number of studies (m)	15, 30, 60, 90, 120, 200	15, 30, 60
Primary study sample size	Typical, small	Typical, small

Simulations used 2000 replications for each simulation condition, with 399 bootstraps for each replication.



μ = 0
μ = 0.1
μ = 0.2
μ = 0.4
μ = 0.8
τ = 0.05
τ = 0.15
τ = 0.3
τ = 0.45
τ = 0.6
Bias
0.8
0.4
0.0
-0.4
0.02 Strong selection
0.05
0.10
0.20
0.50
1.00 No selection
Selection Weights
Method
Correlated-and-Heirarchical Effects Model
PET/PEESE
3PSM (Maximum Likelihood)
3PSM (Hybrid Estimating Equations)


μ = 0
μ = 0.1
μ = 0.2
μ = 0.4
μ = 0.8
τ = 0.05
τ = 0.15
τ = 0.3
τ = 0.45
τ = 0.6
RMSE
0.0
0.2
0.4
0.6
0.8
0.02 Strong selection
0.05
0.10
0.20
0.50
1.00 No selection
Selection Weights
Method
Correlated-and-Heirarchical Effects Model
PET/PEESE
3PSM (Maximum Likelihood)
3PSM (Hybrid Estimating Equations)

Coverage

- 3-parameter selection model
- Hybrid estimating equations