Adapting Methods for Correcting Selective Reporting Bias in Meta-Analysis of Dependent Effect Sizes

Man Chen, Assistant Professor, The University of Texas at Austin

February 14, 2025

Background and Motivation

OPEN ACCESS Freely availabl

Systematic R
Publication E
Updated Rev

Kerry Dwan*, Carrol Ga

Department of Biostatistics, University

Abstract

Background: The increas
types of bias that can ari
reporting bias have been
evidence unreliable for d

Methodology/Principal F
assessed study publicatio
which four were newly id
to information regarding
investigated outcome rep
being fully reported comp
to protocols, we found th
We decided not to under

Conclusions: This update
evidence for the existenc
association between sign
be published and outcom
found to be inconsistent
efforts should be concen

Citation: Dwan K, Gamble C, Willi
Bias and Outcome Reporting Bias -

Editor: Isabelle Boutron, University

Received January 25, 2013; **Accep**

Copyright: © 2013 Dwan et al.
unrestricted use, distribution, and r

OPEN ACCESS Freely ava

Publication
Correlation

Anton Kühberger[1,2

1 Department of Psychology, Un
Zentrum für Begabtenförderung

Abstract

Background: The *p* v
the underlying phen
theoretically indepen
publication bias.

Methods: We investi
sampled 1,000 psycho
sizes of all empirical
distribution of *p* valu

Results: We found a
we found an inordina
neither implicit nor e

Conclusion: The nega
pervasive publication

Citation: Kühberger A, Fritz A
ONE 9(9): e105825. doi:10.137

Editor: Daniele Fanelli, Unive

Received April 17, 2014; **Ac**

Copyright: © 2014 Kühberg
unrestricted use, distribution,

Data Availability: The autho
Supporting Information files.

Funding: This research was s
data collection and analysis, c

Competing Interests: The a

* Email: anton.kuehberger@sb

Article

Publication B
Education Me

Nicholas A. Gage[1

Abstract

Publication bias involves t
effects in the published
omnibus effect sizes in r
effects than they actually
fields, research has not
the inclusion of gray lite
bias exists, the relation
differences in effect size
analyses published in sp
analyses included gray l
including gray literature
had larger effect sizes th
research and practice.

J Bus Psychol (2016) 31:323–338
DOI 10.1007/s10869-016-9456-7

CrossMark

EDITORIAL

Editorial: Evidence on Questionable Research Practices: The Good, the Bad, and the Ugly

George C. Banks[1] · Steven G. Rogelberg[1] · Haley M. Woznyj[1] · Ronald S. Landis[2] · Deborah E. Rupp[3,4]

Published online: 25 June 2016

Abstract

Purpose Questionable research or reporting practices (QRPs) contribute to a growing concern regarding the credibility of research in the organizational sciences and related fields. Such practices include design, analytic, or reporting practices that may introduce biased evidence, which can have harmful implications for evidence-based practice, theory development, and perceptions of the rigor of science.

Design/Methodology/Approach To assess the extent to which QRPs are actually a concern, we conducted a systematic review to consider the evidence on QRPs. Using a triangulation approach (e.g., by reviewing data from observations, sensitivity analyses, and surveys), we identified the good, the bad, and the ugly.

Findings Of the 64 studies that fit our criteria, 6 appeared to find little to no evidence of engagement in QRPs and the other 58 found more severe evidence (91 %).

Implications Drawing upon the findings, we provide recommendations for future research related to publication practices and academic training.

Originality/value We report findings from studies that suggest that QRPs are not a problem, that QRPs are used at a suboptimal rate, and that QRPs present a threat to the viability of organizational science research.

Keywords Questionable research practices QRPs · Research methodology · Philosophy of science · Ethics · Research methods

Introduction

Concerns exist regarding the credibility of research in the social and natural sciences (Cortina 2015; Kepes and McDaniel 2013; Nosek et al. 2015; Schmidt and Hunter 2015). These concerns are linked, in part, to the use of questionable research or reporting practices (QRPs). QRPs have been defined as "design, analytic, or reporting practices that have been questioned because of the potential for the practice to be employed with the purpose of presenting biased evidence in favor of an assertion" (Banks et al. 2016, p. 3). Examples of commonly discussed QRPs include selectively reporting hypotheses with a preference for those that are statistically significant, "cherry picking" fit indices in structural equation modeling (SEM), and presenting post hoc hypotheses as if they were developed a

Research Article

Psychological Science
23(5) 524–532

DOI: 10.1177/0956797611430953
http://pss.sagepub.com
SAGE

Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling

Leslie K. John[1], George Loewenstein[2], and Drazen Prelec[3]
[1]Marketing Unit, Harvard Business School; [2]Department of Social & Decision Sciences, Carnegie Mellon University; and [3]Sloan School of Management and Departments of Economics and Brain & Cognitive Sciences, Massachusetts Institute of Technology

Abstract

Cases of clear scientific misconduct have received significant media attention recently, but less flagrantly questionable research practices may be more prevalent and, ultimately, more damaging to the academic enterprise. Using an anonymous elicitation format supplemented by incentives for honest reporting, we surveyed over 2,000 psychologists about their involvement in questionable research practices. The impact of truth-telling incentives on self-admissions of questionable research practices was positive, and this impact was greater for practices that respondents judged to be less defensible. Combining three different estimation methods, we found that the percentage of respondents who have engaged in questionable practices was surprisingly high. This finding suggests that some questionable practices may constitute the prevailing research norm.

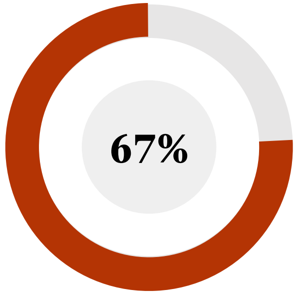
Keywords

professional standards, judgment, disclosure, methodology

Received 5/20/11; Revision accepted 10/20/11

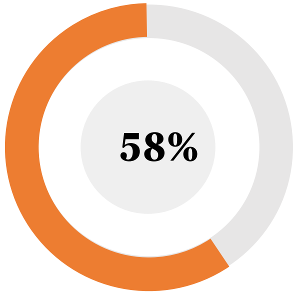
Questionable Research Practices (QRPs)

(John et al., 2012)



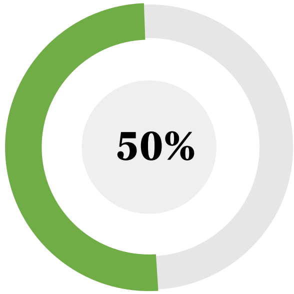
In a paper, failing to report all of a study's dependent measures.

Selective reporting



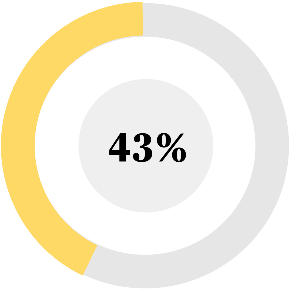
Collecting more data after seeing whether results were significant.

P-hacking



In a paper, selectively reporting studies that “worked”.

Selective reporting



Deciding whether to exclude data after looking at the impact of doing so on the results.

P-hacking

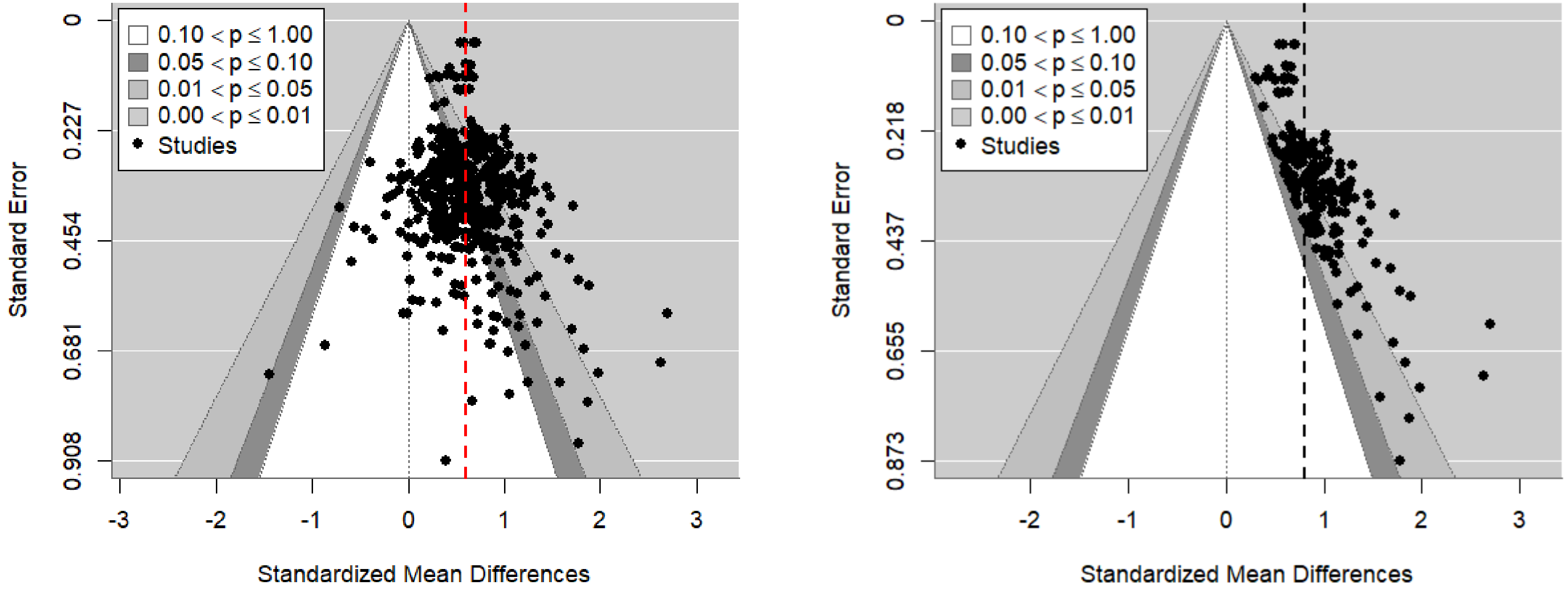
Questionable Research Practices (QRPs)

- A variety of problematic behaviors in research design, analysis, interpretation, and reporting that produce favorable results but undermine the credibility and rigor of scientific research (Banks et al., 2016; Friese & Frankenbach, 2020).

- Common QRPs
  - **Selective reporting of positive findings**
  - P-hacking or fishing for statistical significance
  - HARKing: Hypothesizing after the results are known

Selective Reporting

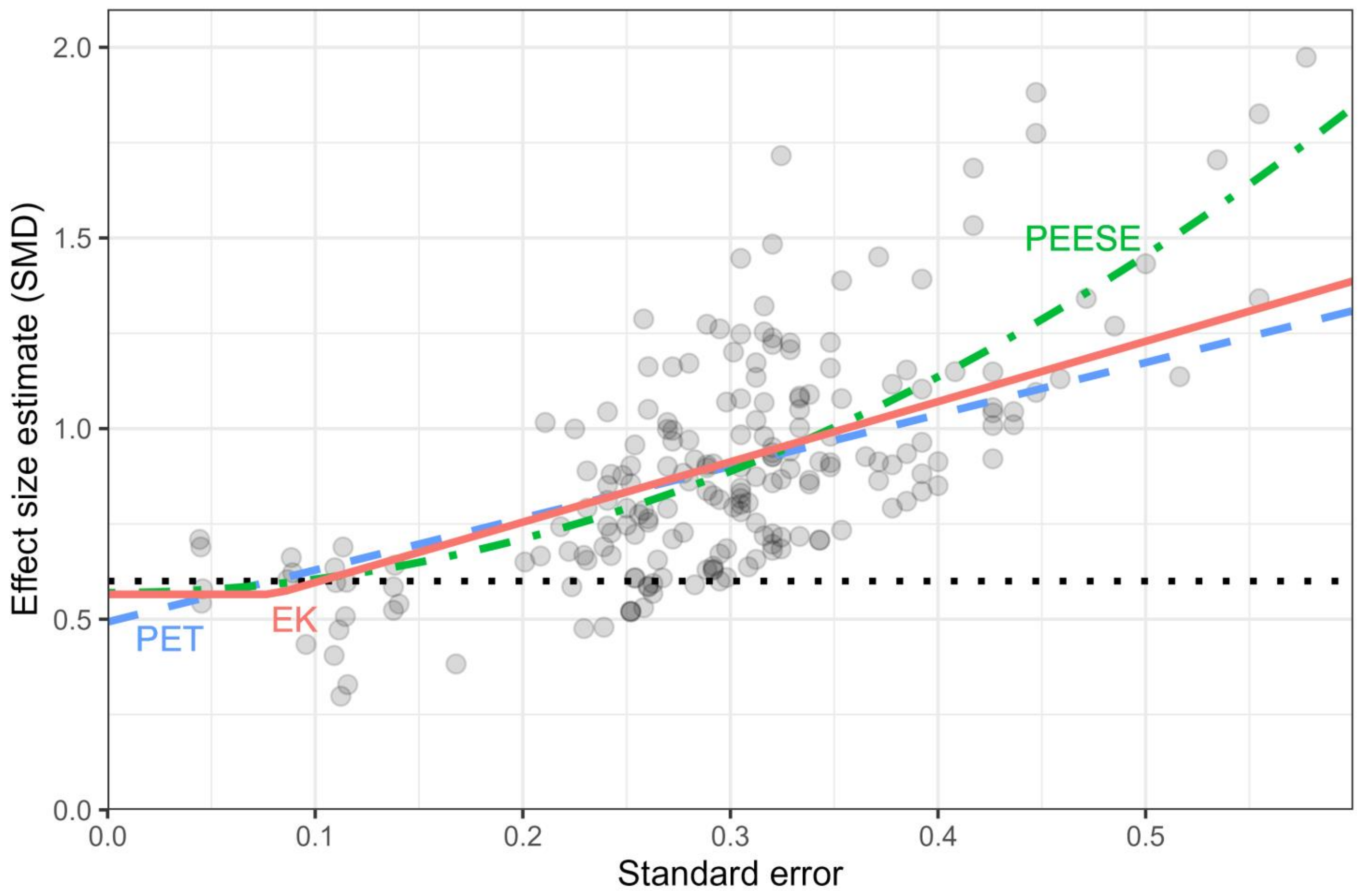
- **Selective reporting** occurs if *affirmative* results within a study or the entire study are preferentially reported and more likely to be included in meta-analysis compared to *non-affirmative* results.
- Selective reporting can result in **over-estimated average effect sizes**, inflated Type I error rates, and inappropriate inferences about intervention effects (Carter et al., 2019)



Existing Methods

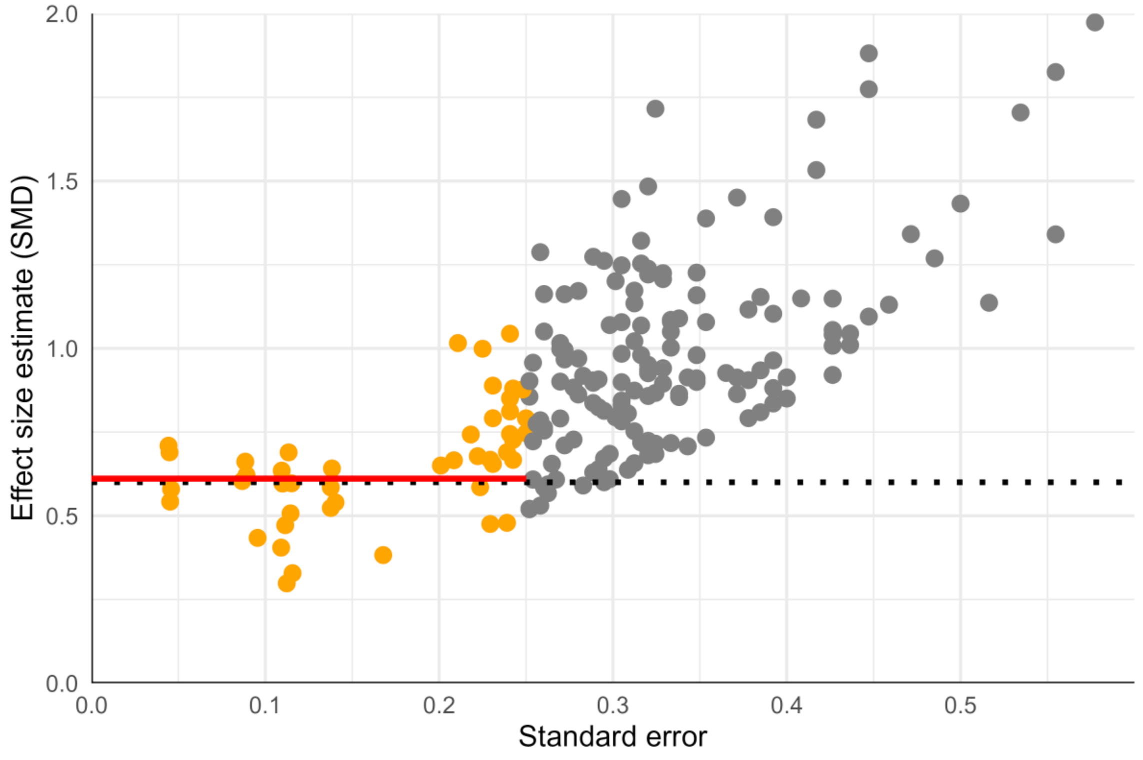
- Regression-based adjustment methods for small-study effects
  - PET-PEESE (Stanley & Doucouliagos, 2014)
  - Weighted average of the adequately powered (WAAP, Stanley et al., 2017)
  - Weighted and iterated least squares (WILS, Stanley & Doucouliagos, 2022)
  - Endogenous kink model (EK, Bom & Rachinger, 2019)
- Selection models
  - *p*-value selection models (e.g., Hedges, 1992; Vevea & Hedges, 1995)
  - *p*-curve (Simonsohn et al., 2014) , *p*-uniform, *p*-uniform* (van Aert et al., 2023)

Univariate Regression-Based Methods

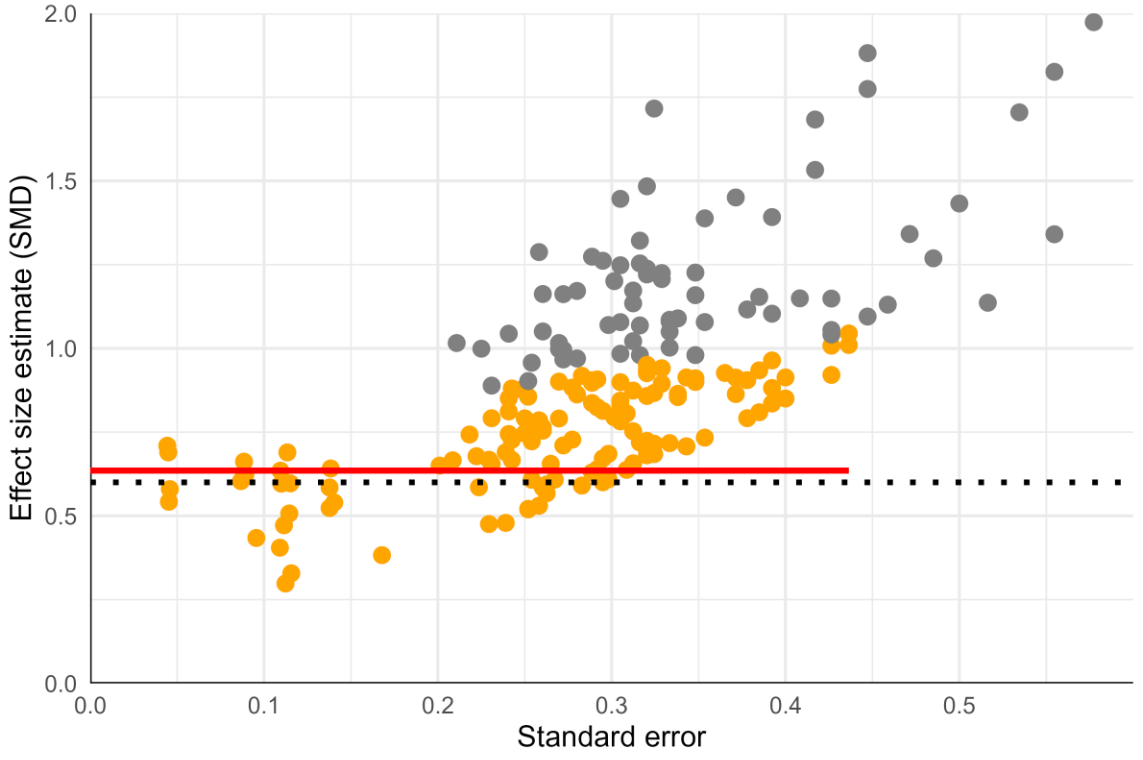


Precision-effect test
PET estimate: 0.493
Precision-effect estimator with SE
PEESE estimate: 0.569
Endogenous kink meta-regression
EK estimate: 0.565

Univariate Regression-Based Methods

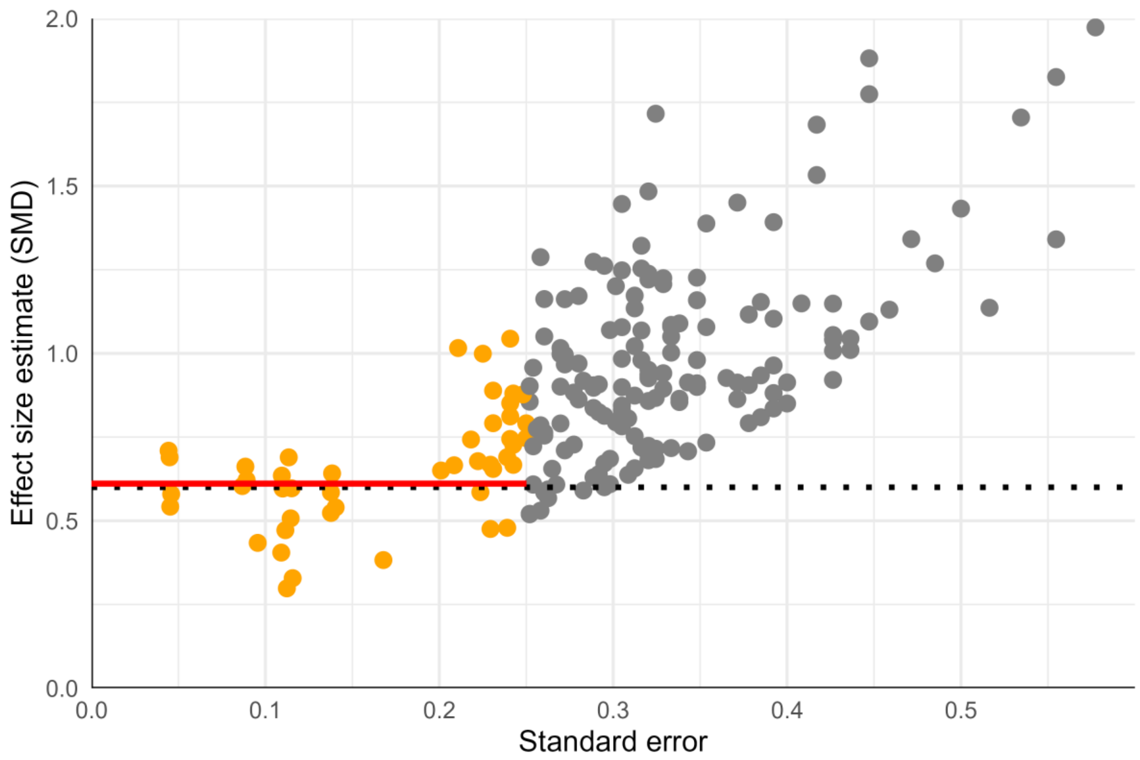


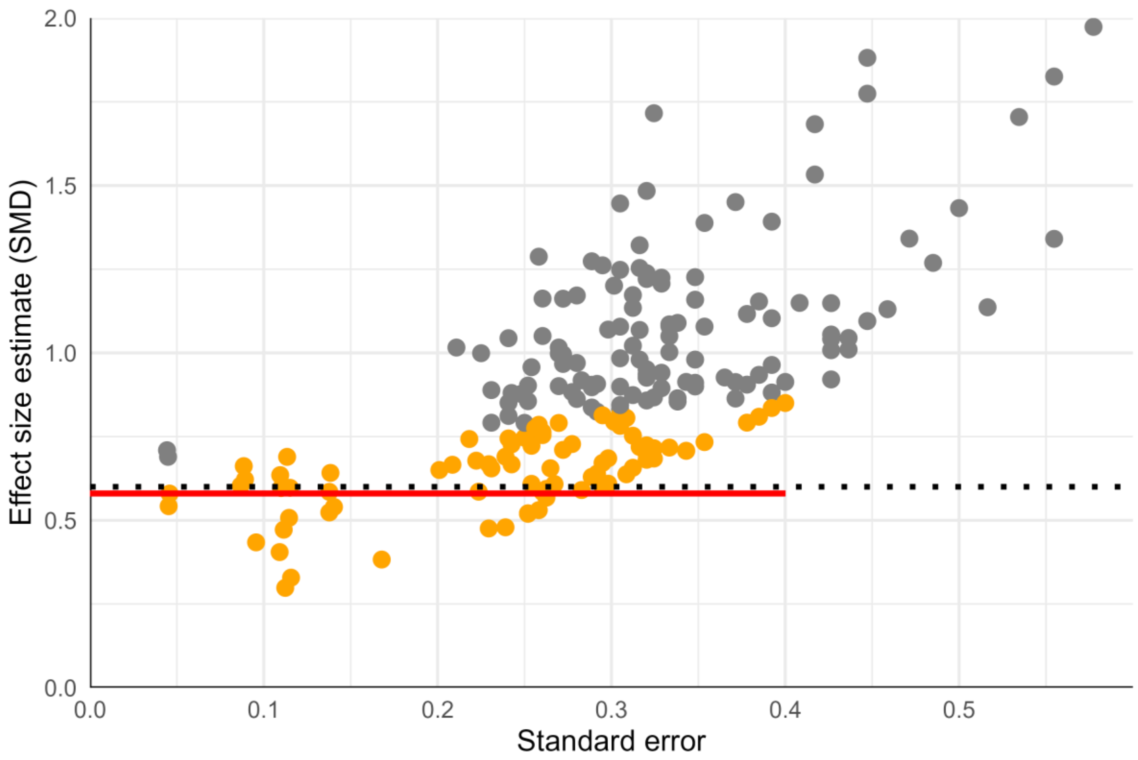
Weighted average of the adequately powered
WAAP estimate: 0.611



Weighted and iterated least squares
WILS estimate (1st iteration): 0.635

Univariate Regression-Based Methods

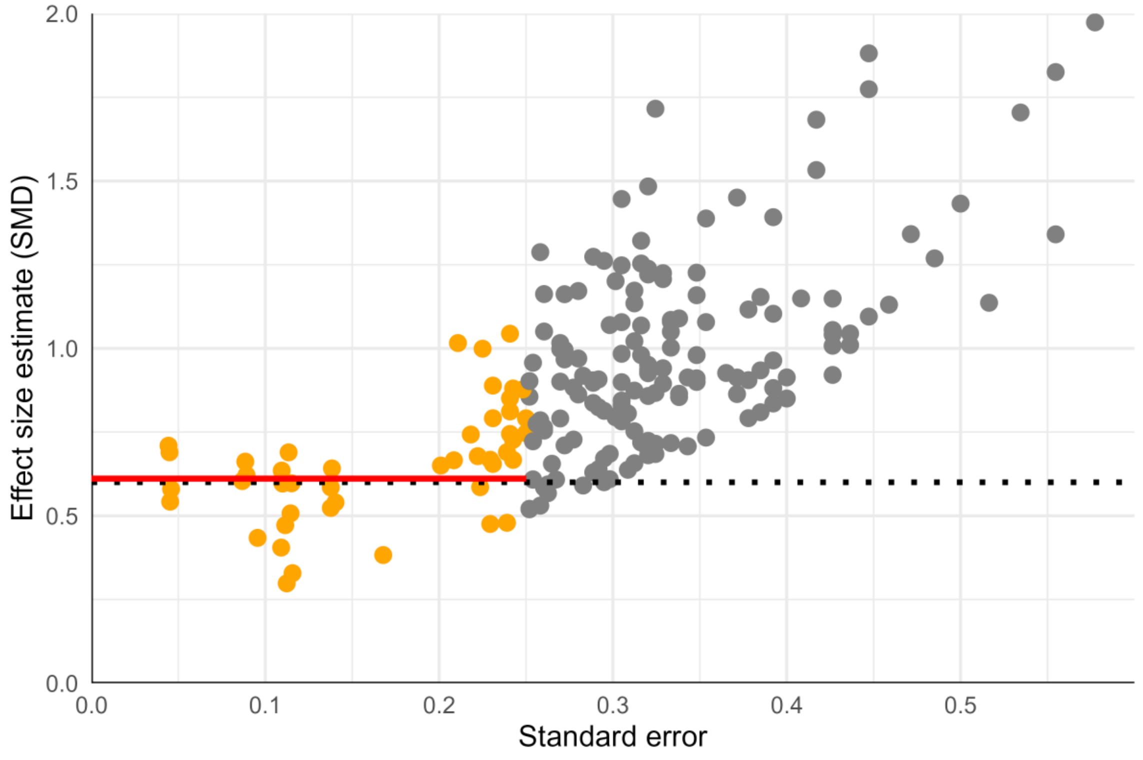




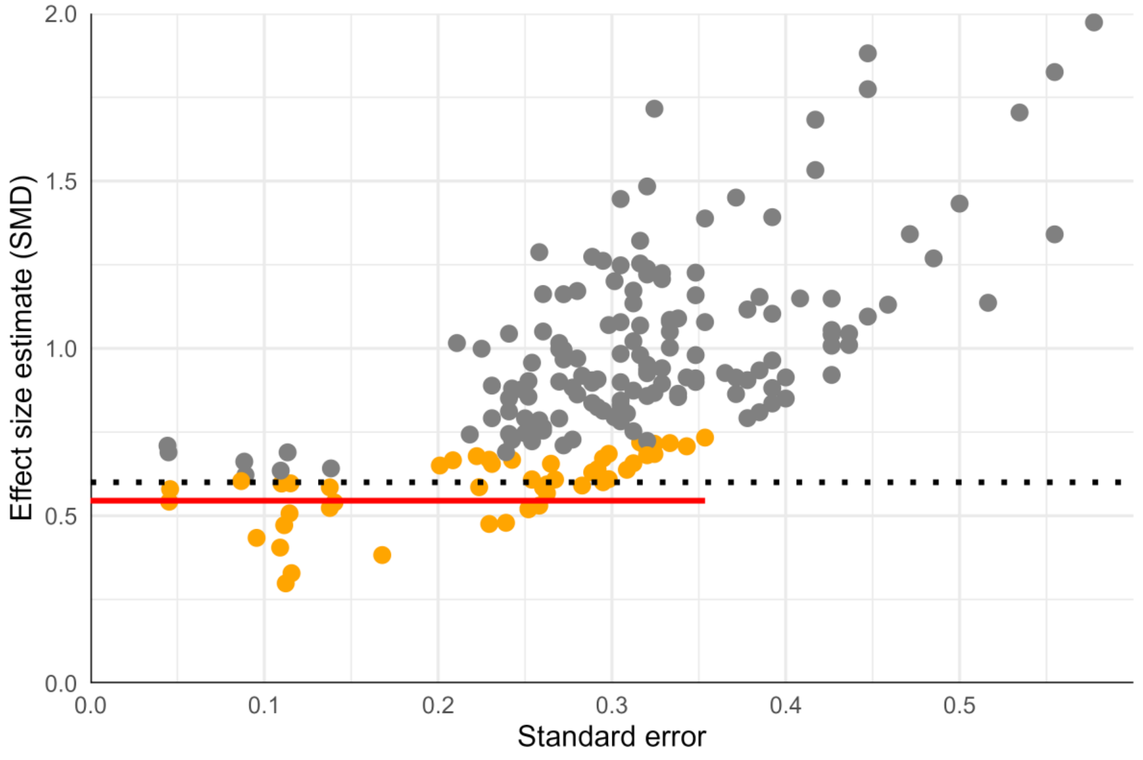
Weighted average of the adequately powered
WAAP estimate: 0.611

Weighted and iterated least squares
WILS estimate (2rd iteration): 0.580

Univariate Regression-Based Methods

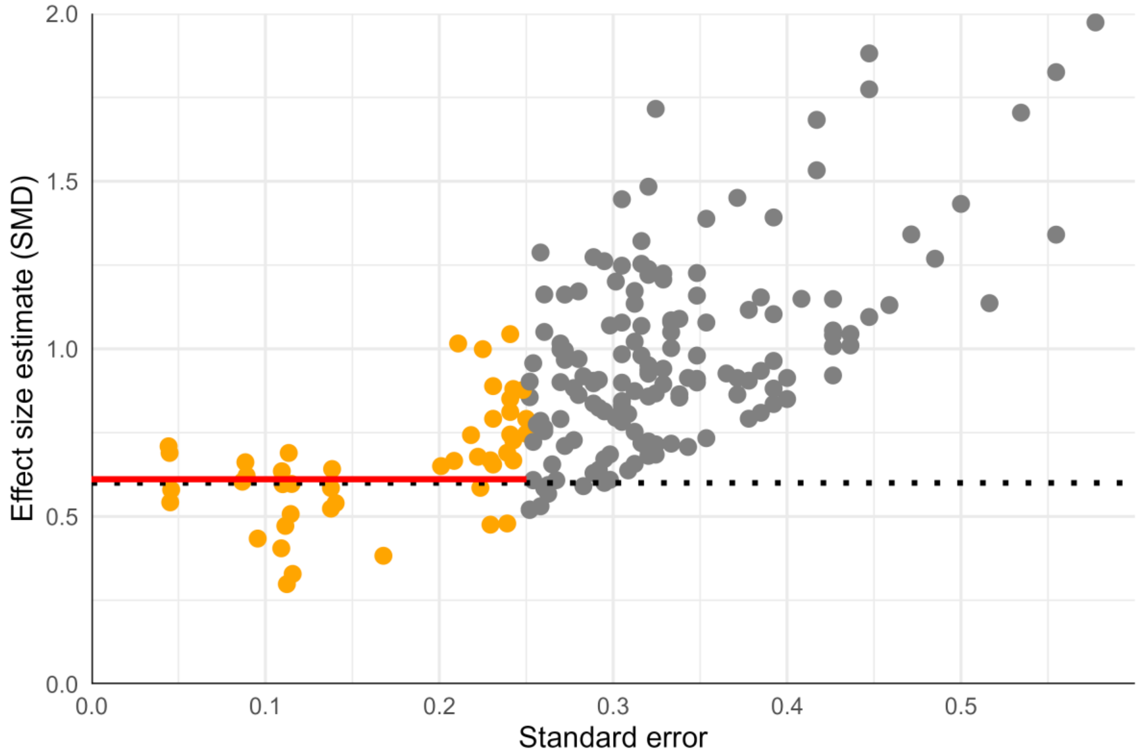


Weighted average of the adequately powered
WAAP estimate: 0.611

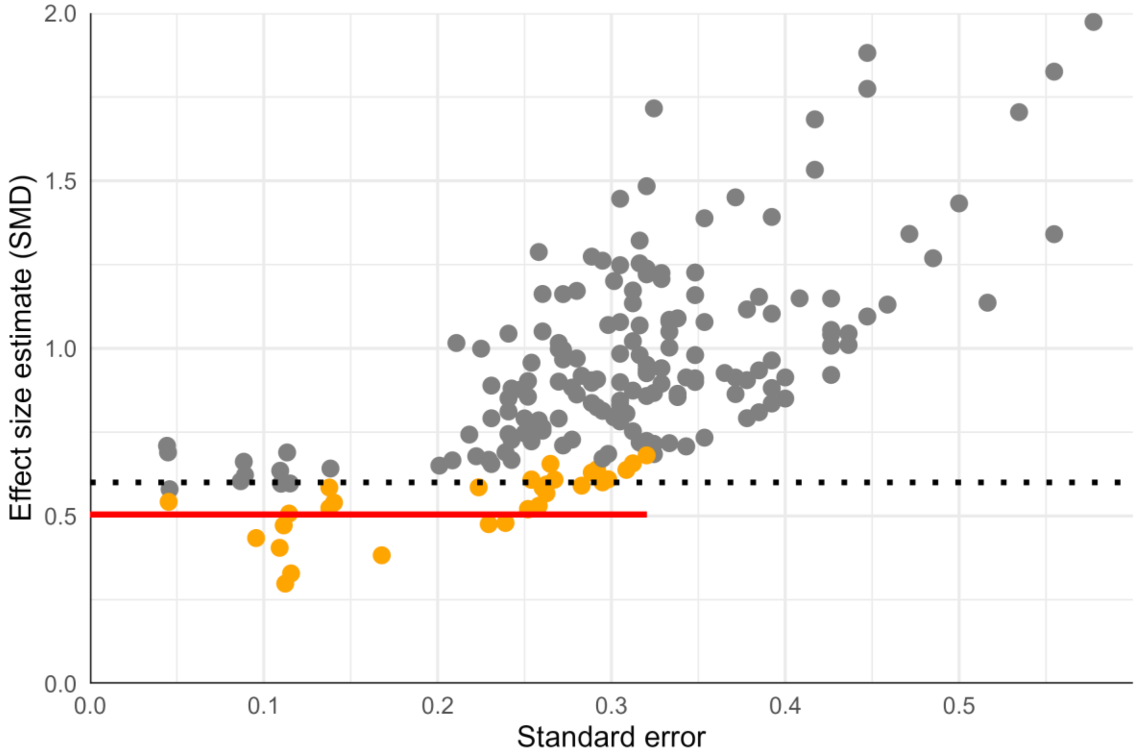


Weighted and iterated least squares
WILS estimate (3rd iteration): 0.545

Univariate Regression-Based Methods

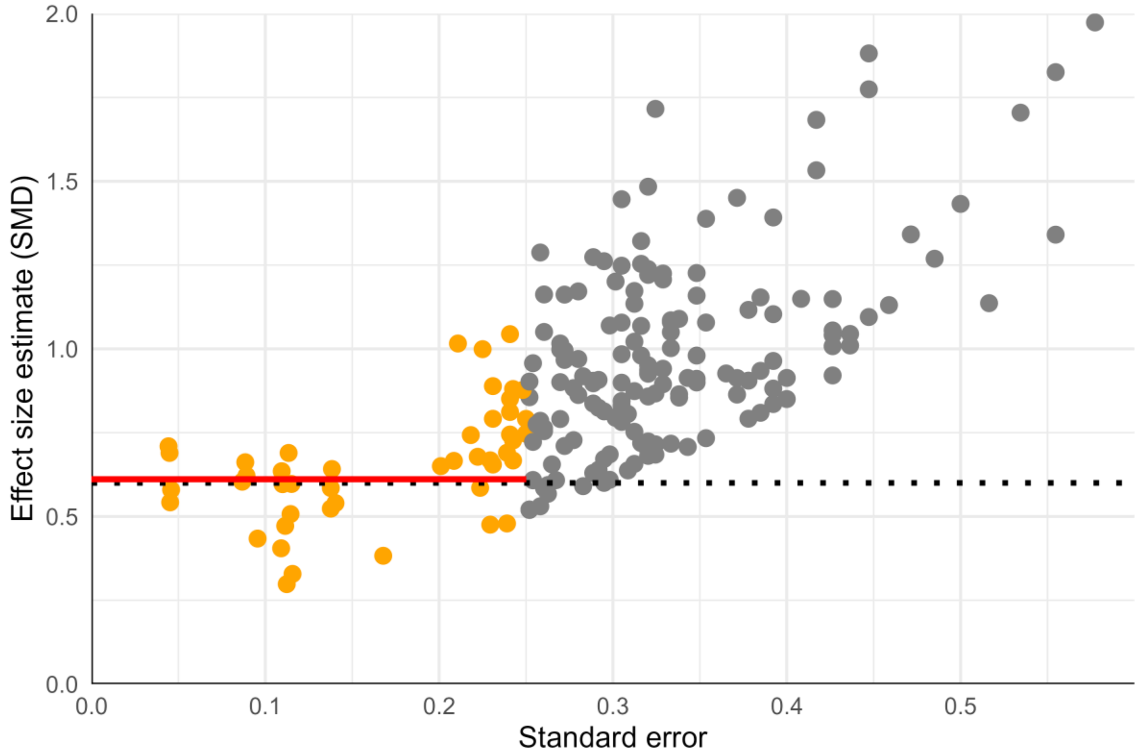


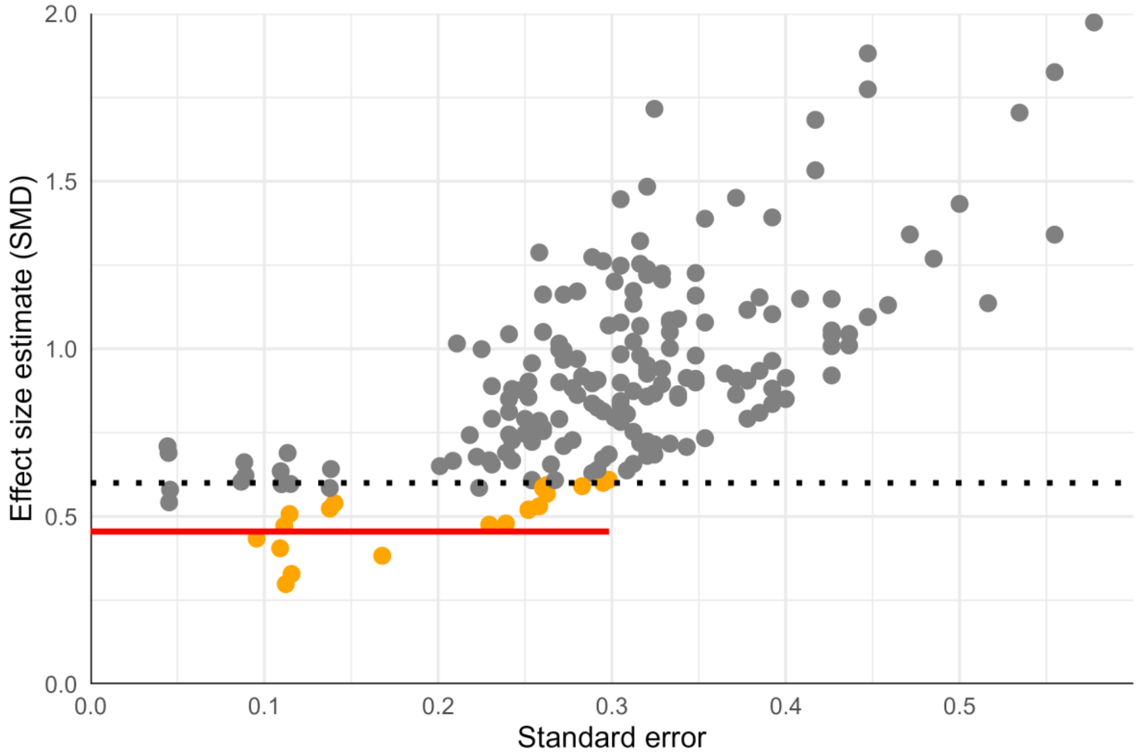
Weighted average of the adequately powered
WAAP estimate: 0.611



Weighted and iterated least squares
WILS estimate (4^{th} iteration): 0.504

Univariate Regression-Based Methods

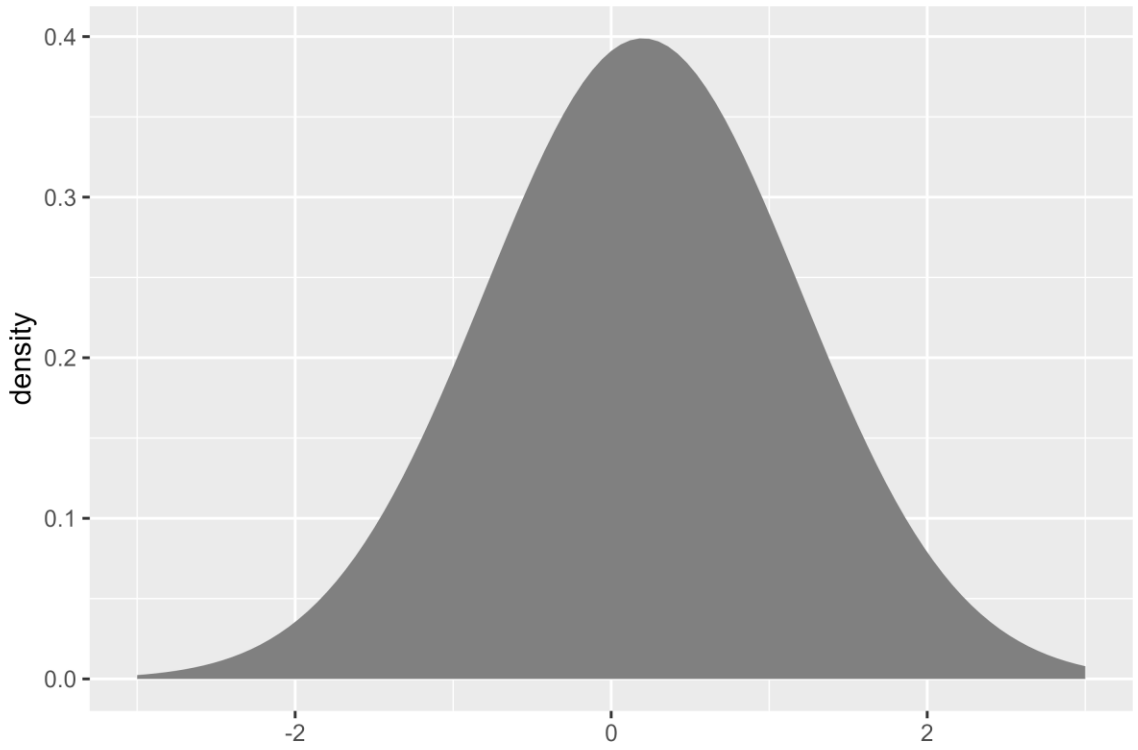




Weighted average of the adequately powered
WAAP estimate: 0.611

Weighted and iterated least squares
WILS estimate (5^{th} iteration): 0.455

Univariate Selection Model (3PSM)



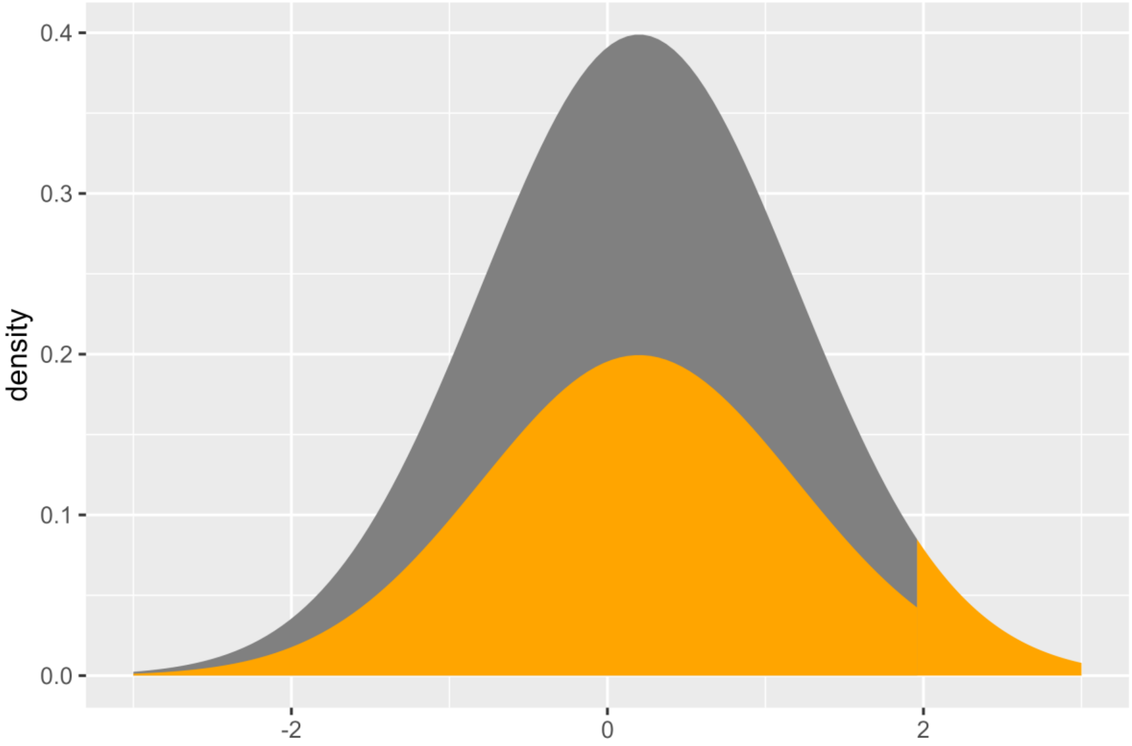
- The data model

$$Y_i^* \sim N(\mu, \tau^2 + V_i^*)$$

- The selection model (weight function)

$$\mathrm{W}(y|\lambda) = \begin{cases} 1 & if\ 0 < p_i \leq .025 \\ \lambda & if\ .025 < p_i \leq 1 \end{cases}$$

Univariate Selection Model (3PSM)



- The data model

$$Y_i^* \sim N(\mu, \tau^2 + V_i^*)$$

- The selection model (weight function)

$$W(y|\lambda) = \begin{cases} 1 & if\ 0 < p_i \leq .025 \\ \lambda & if\ .025 < p_i \leq 1 \end{cases}$$

- The observed density

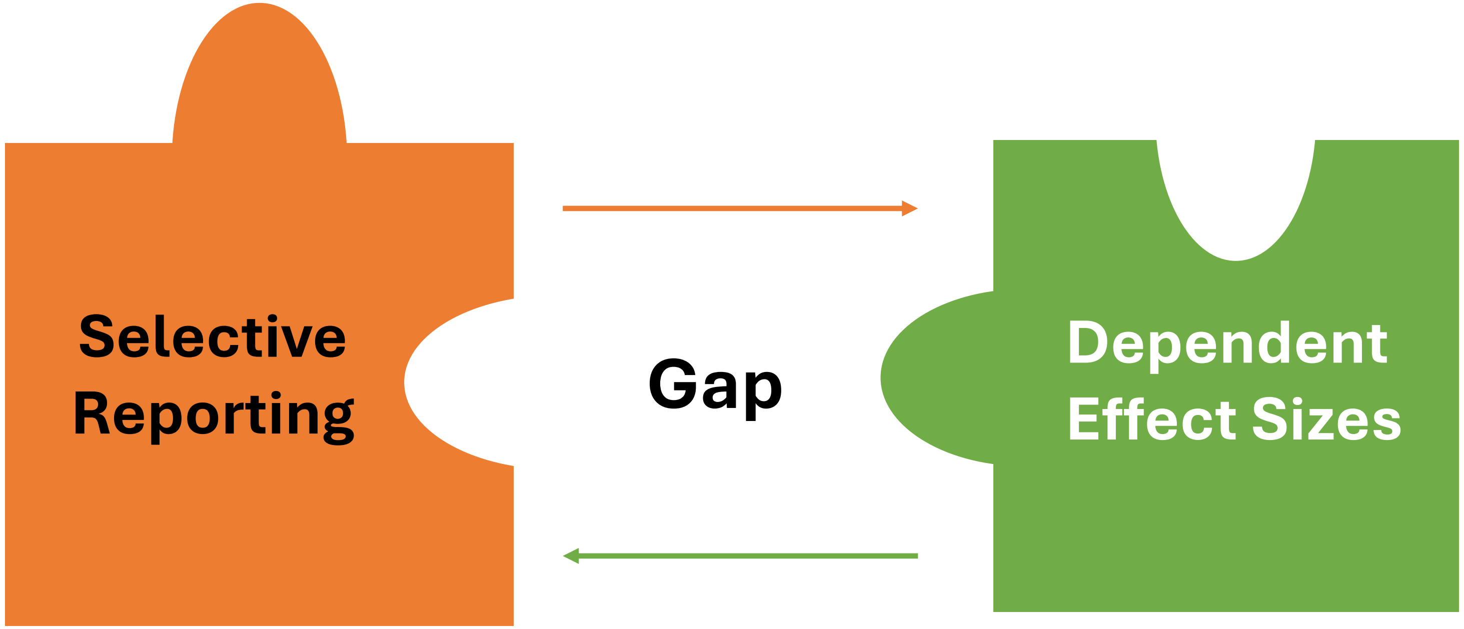
$$g(y|\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{f(y|\boldsymbol{\theta})W(y|\boldsymbol{\lambda})}{\int f(y|\boldsymbol{\theta})W(y|\boldsymbol{\lambda})dy}$$

Problem Statement

Selective Reporting

Regression-based methods
Selection models
P-uniform, p-uniform*...

Problem Statement



Regression-based methods
Selection models
P-uniform, p-uniform*...

Multilevel methods
Multivariate methods
Robust variance estimation

Purpose of the Study

- To ***propose a new weighting scheme*** for the correlated and hierarchical effects model in RVE framework to account for effect size dependencies

- To ***adapt univariate regression-based adjustment methods*** using the proposed working model and weighting scheme.

- To ***evaluate the performance*** of these adjustment methods, including novel adaptations, in an extensive simulation study that emulates the features of real-world meta-analyses assuming ***p-value selection forms***.

CHE-ISCW and Novel Adaptations

CHE-ISCW

- In univariate meta-analysis, **fixed effects model weights** were proposed to be used in random effects meta-regression models to allocate relatively more weights to large studies that are less susceptible to selective reporting bias (Henmi & Copas, 2010).

CHE-ISCW

- **The CHE model** (correlated and hierarchical effects)
  - $T_{ij} = \mu + u_j + v_{ij} + e_{ij}$ $\quad u_j \sim N(0, \tau^2) \quad v_{ij} \sim N(0, \omega^2) \quad Var(e_{ij}) = s_{ij}^2 \quad Cov(e_{hj}, e_{ij}) = \rho s_j^2$
  - The weight: $\boldsymbol{w_j} = \dfrac{\boldsymbol{n_j}}{(\hat{\boldsymbol{\tau}}^2 + \boldsymbol{\rho s_j^2})\boldsymbol{n_j} + \hat{\boldsymbol{\omega}}^2 + (\mathbf{1} - \boldsymbol{\rho})\boldsymbol{s_j^2}}$
- **The ISCW weights** (inverse sampling covariance weights)
  - $\boldsymbol{S_j} = \rho s_j^2 \boldsymbol{J_j} + (1 - \rho) s_j^2 \boldsymbol{I_j}$ $\quad \boldsymbol{W_j} = \boldsymbol{S_j}^{-1}$
  - The weight: $\tilde{\boldsymbol{w}}_{\boldsymbol{j}} = \dfrac{\boldsymbol{n_j}}{\boldsymbol{\rho s_j^2 n_j} + (\mathbf{1} - \boldsymbol{\rho})\boldsymbol{s_j^2}}$
- **Cluster-robust standard error**

Novel Adaptations

CHE-ISCW

Univariate	Multivariate
01 EK	01 EK
02 PET-PEESE	02 PET-PEESE
03 WAAP	03 WAAP
04 WILS	04 WILS

Simulation Study

Research Questions

- How do **univariate adjustment methods** perform in the context of dependent effect sizes in the presence of one-step or two-step selection?

- In the dependent effect size context and under one- or two-step selection, how do the **adapted estimators based on CHE-ISCW** perform compared to their univariate counterparts?

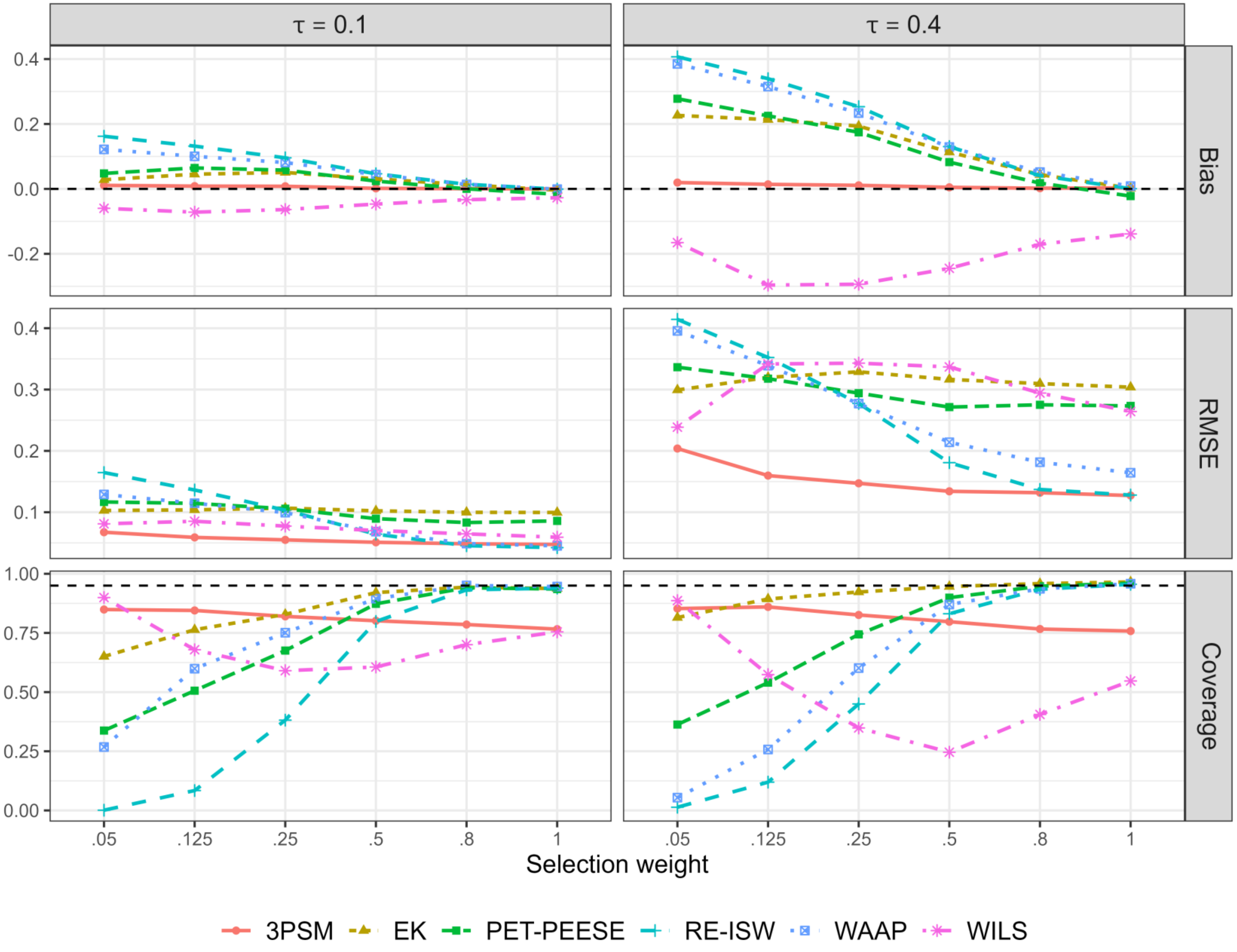
- How do promising multivariate adapted adjustment methods perform compared to the most effective univariate estimators?

Simulation Methods

- **Data Generation**
  - Generated meta-analytic dataset with dependent effect sizes
  - Censored under one-step and two-step p-value selection
- **Estimators**
  - Univariate regression methods: RE-ISW, PET-PEESE, EK, WAAP, WILS
  - Other univariate methods: trim and fill, p-uniform, p-uniform*, 3PSM, 4PSM
  - Multivariate: CHE-ISCW, adapted PET-PEESE, adapted EK, adapted WAAP, adapted WILS
- **Performance criteria**
  - Bias
  - Accuracy
  - Confidence interval coverage and width

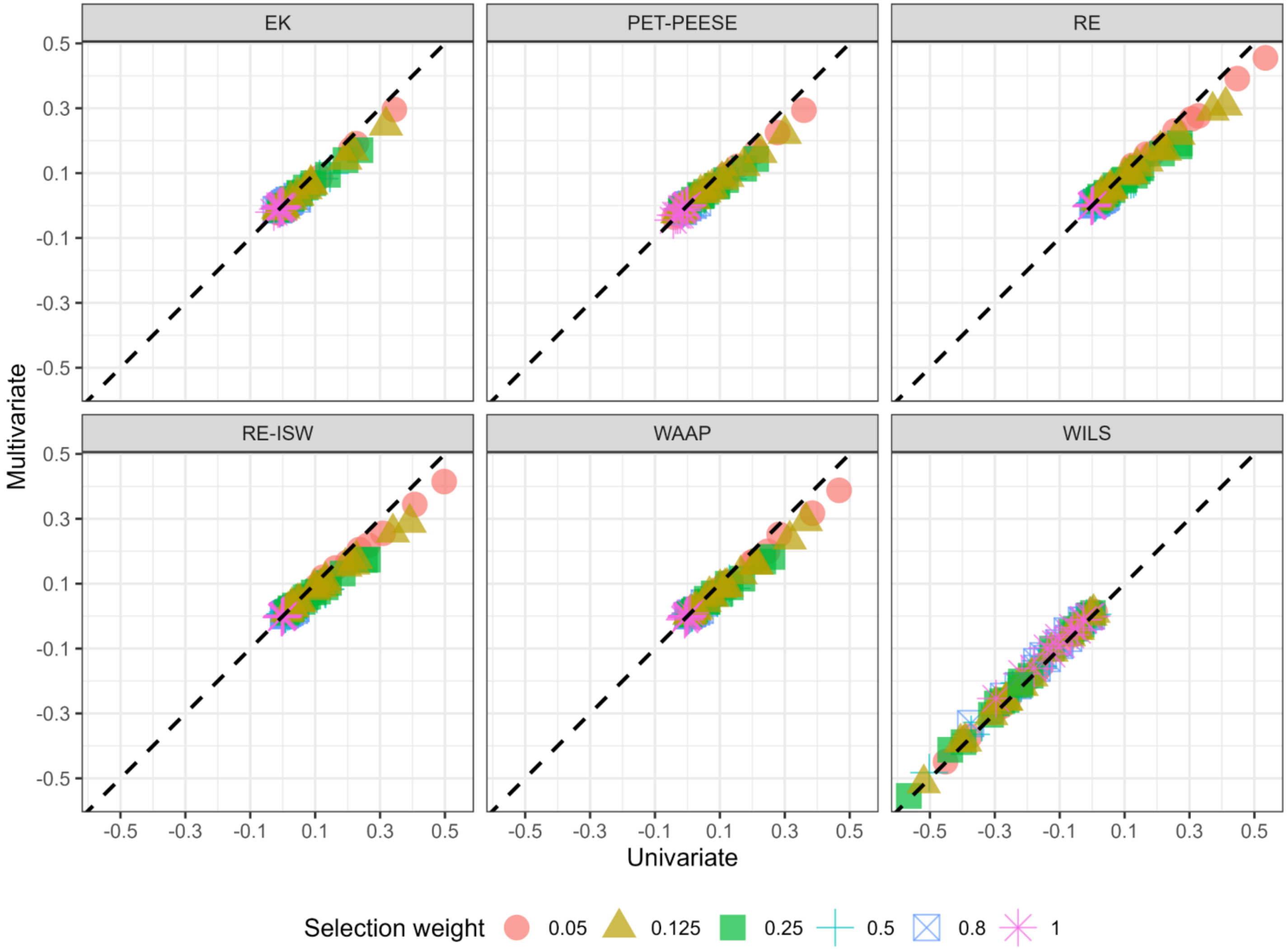
Results

Highlights 1: Should not ignore dependence



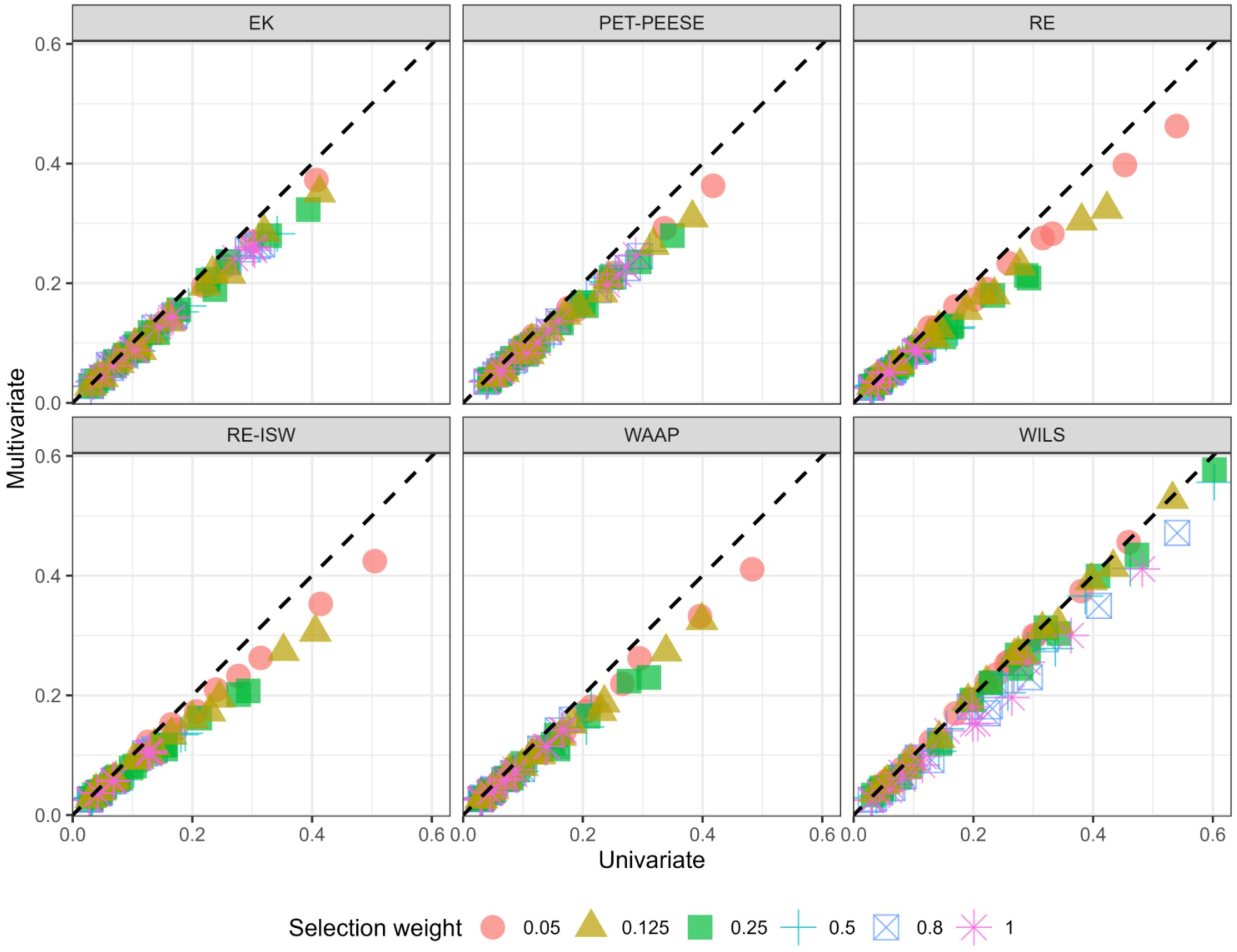
Average effect size: 0.2
Number of studies: 30
Average outcome corr: 0.4
One-step selection

Highlights 2: CHE-ISCW improves bias, accuracy, and coverage



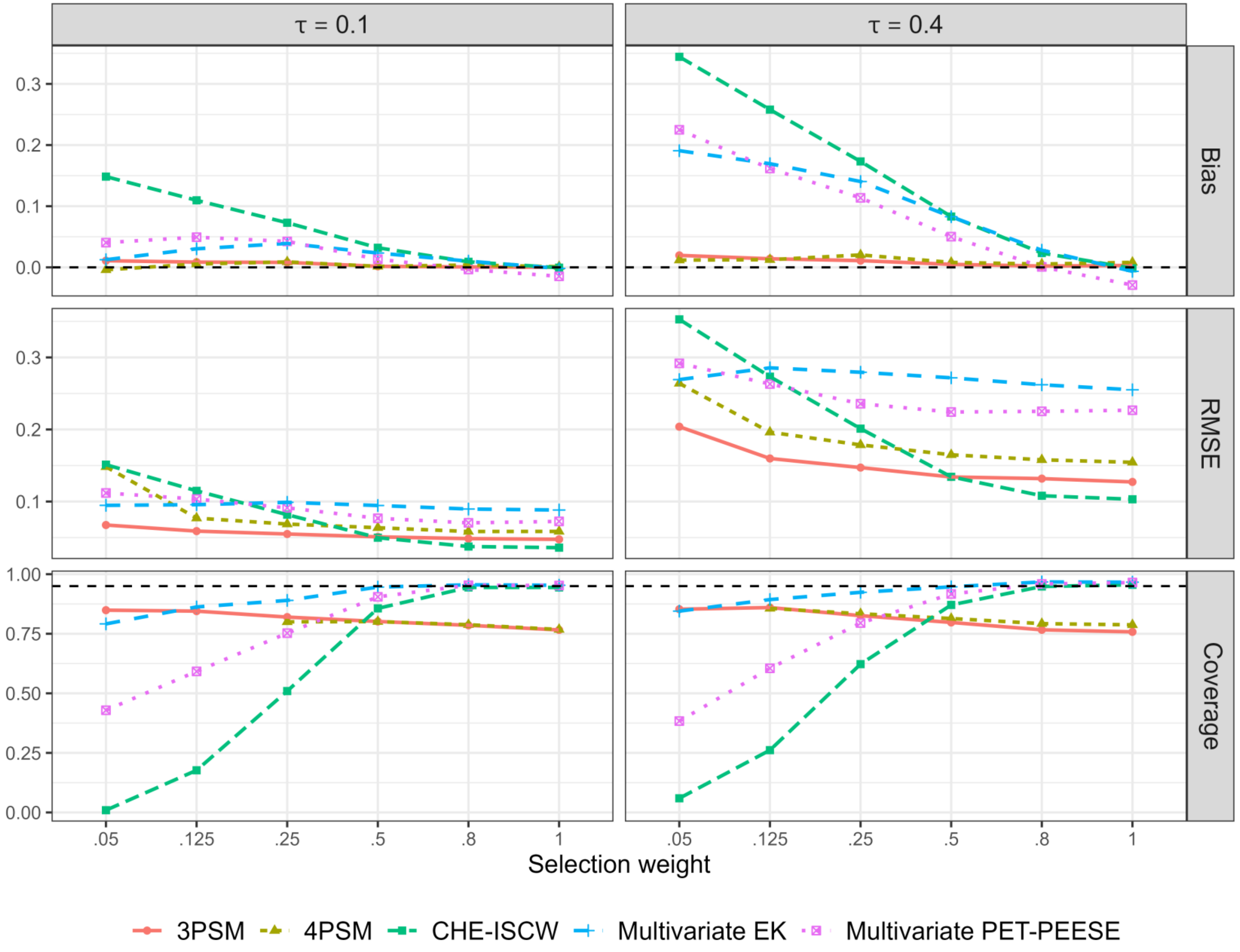
Number of studies: 30
One-step selection

Highlights 2: CHE-ISCW improves bias, accuracy, and coverage



Number of studies: 30
One-step selection

Highlights 3: No Clear Winner!



Average effect size: 0.2
Number of studies: 30
Average outcome corr: 0.4
One-step selection

Discussion

Implications

- Meta-analysts should not ignore effect size dependencies when correcting for selective reporting bias.
- Sensitivity analyses are recommended in practice because none of the methods performs adequately across all simulation conditions.
- While methodological work is yet needed for further developing more robust adjustment methods for selection bias, the most efficient strategy for addressing selective reporting is to prevent its occurrence.

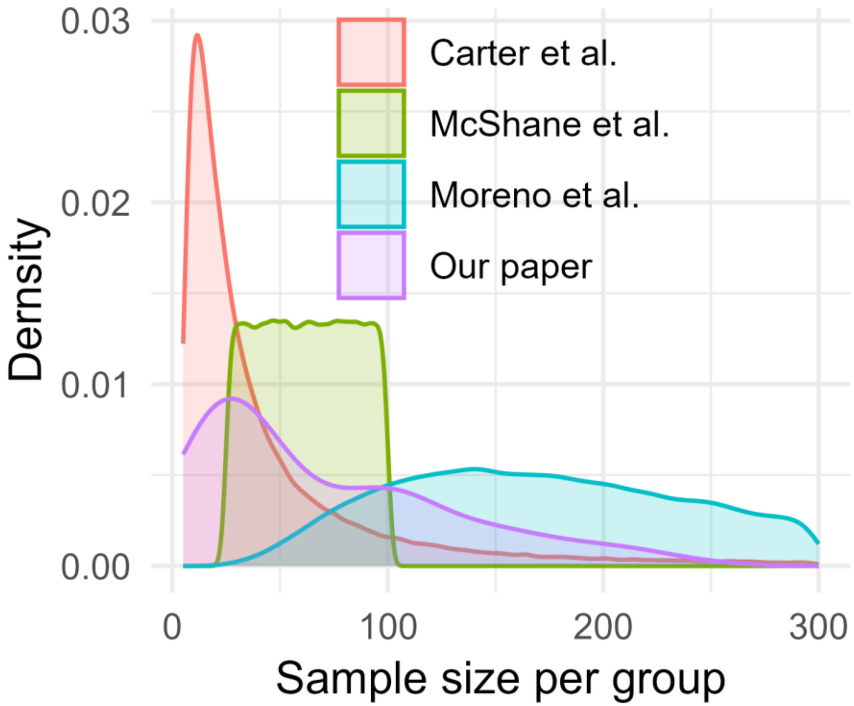
Limitations and Future Directions

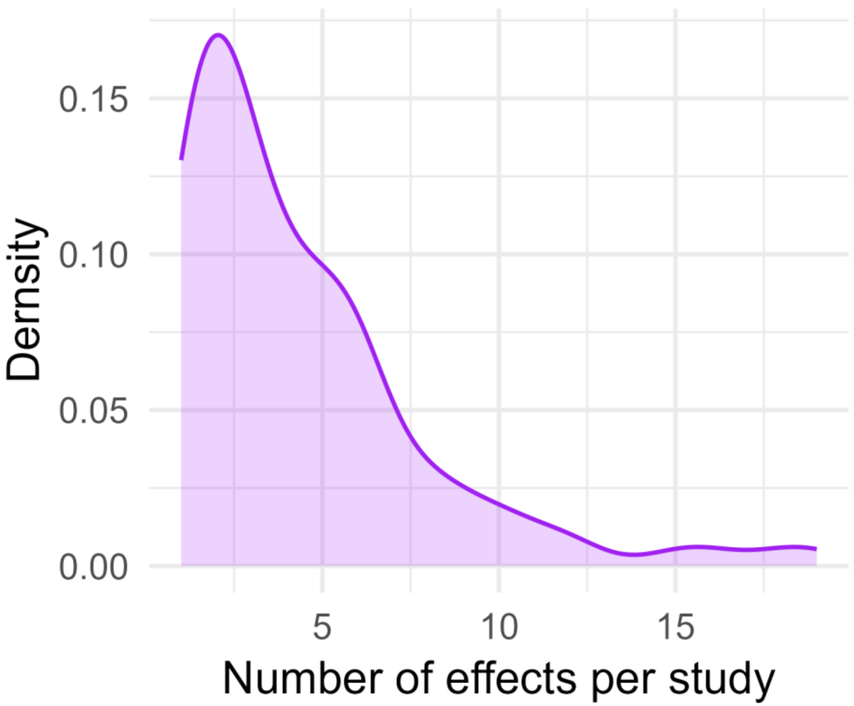
- The simulation is limited to one- and two-step p-value selection of outcomes within study. Further research should consider other types of **selection mechanisms**.
- The CHE-ISCW working model only includes the **unexplained heterogeneity**. Future research could consider incorporating moderators to explain the heterogeneity.
- This study only examined the recovery of average effect size parameter. Research is needed to **evaluate heterogeneity estimators** in the presence of selective reporting and effect size dependencies.

Thank you

Chen, M., & Pustejovsky, J. E. (2024, October 25). Adapting Methods for Correcting Selective Reporting Bias in Meta-Analysis of Dependent Effect Sizes. https://doi.org/10.31222/osf.io/jq52s

Supplement

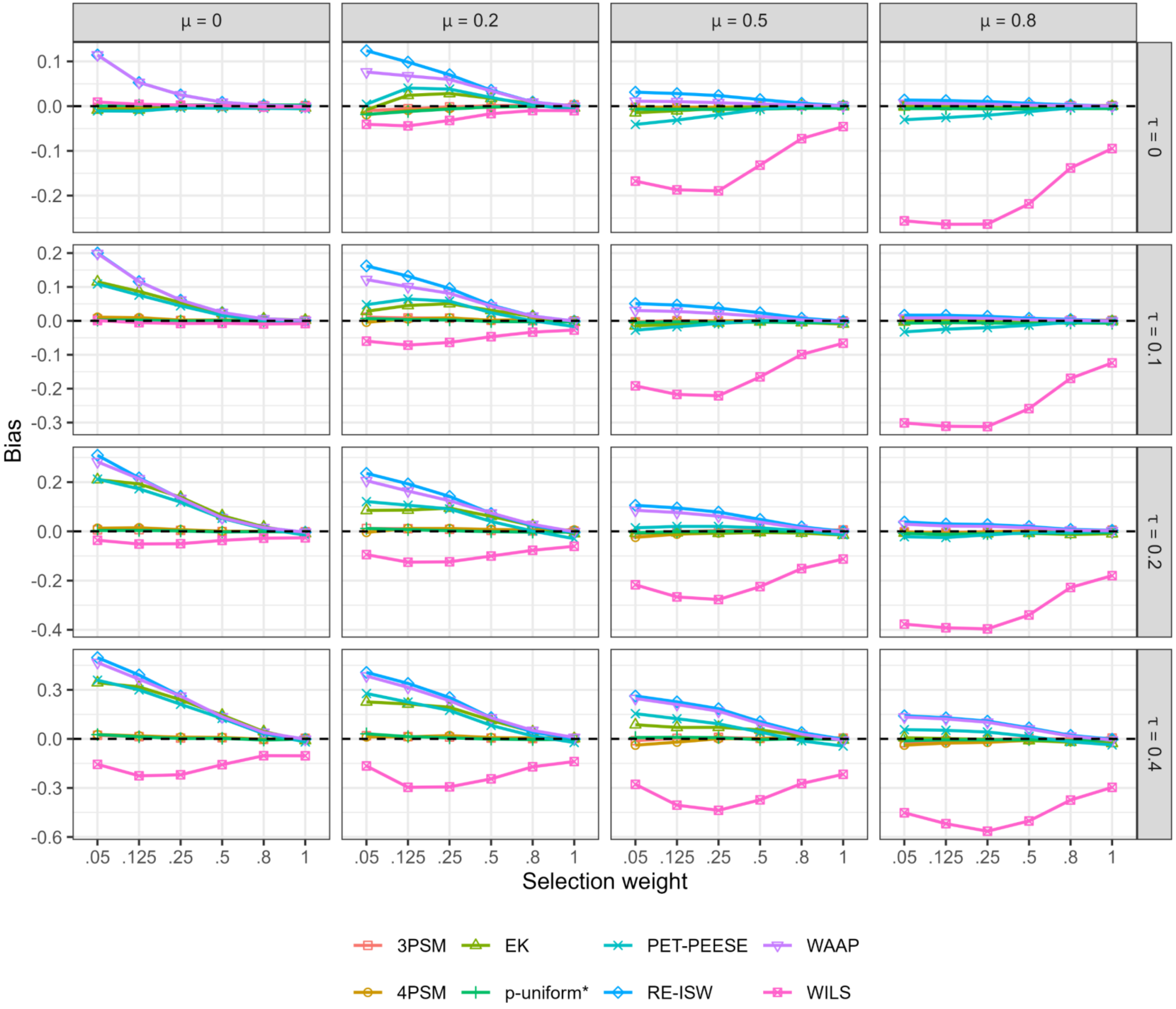

0.03
0.02
0.01
0.00
Dernsity
0
100
200
300
Sample size per group
Carter et al.
McShane et al.
Moreno et al.
Our paper


0.15
0.10
0.05
0.00
Dernsity
5
10
15
Number of effects per study

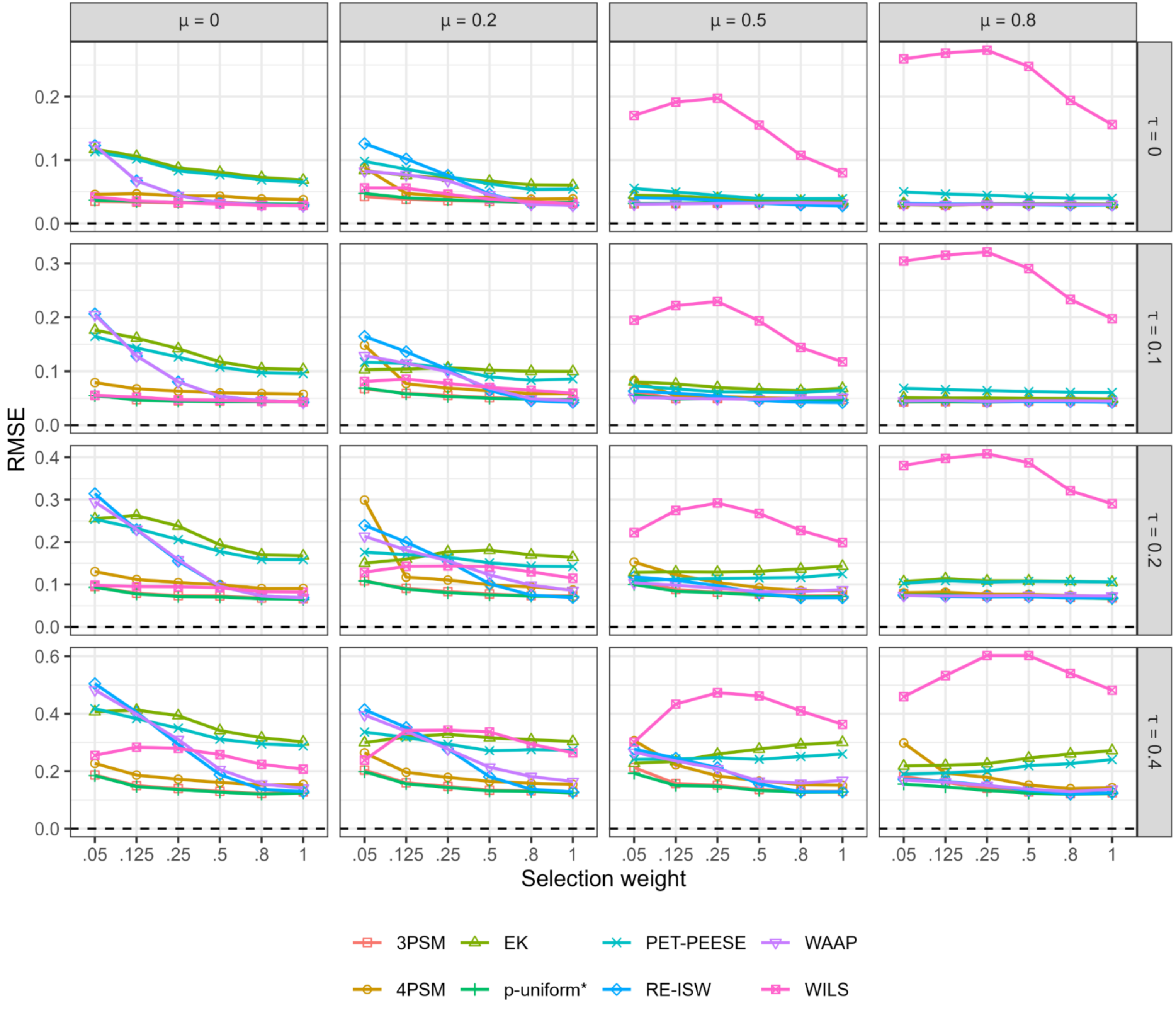
Experimental Design

Simulation parameters	Values
Overall average effect	0, 0.2, 0.5, 0.8
Between-study heterogeneity	0, 0.1, 0.2, 0.4
Number of studies	10, 30, 60, 100
Average correlation between outcomes	0.2, 0.4, 0.8
Selection weight for $.025 < p \leq .5$	1, 0.8, 0.5, 0.25, 0.125, 0.05
Ratio of selection weights	1, 0.5

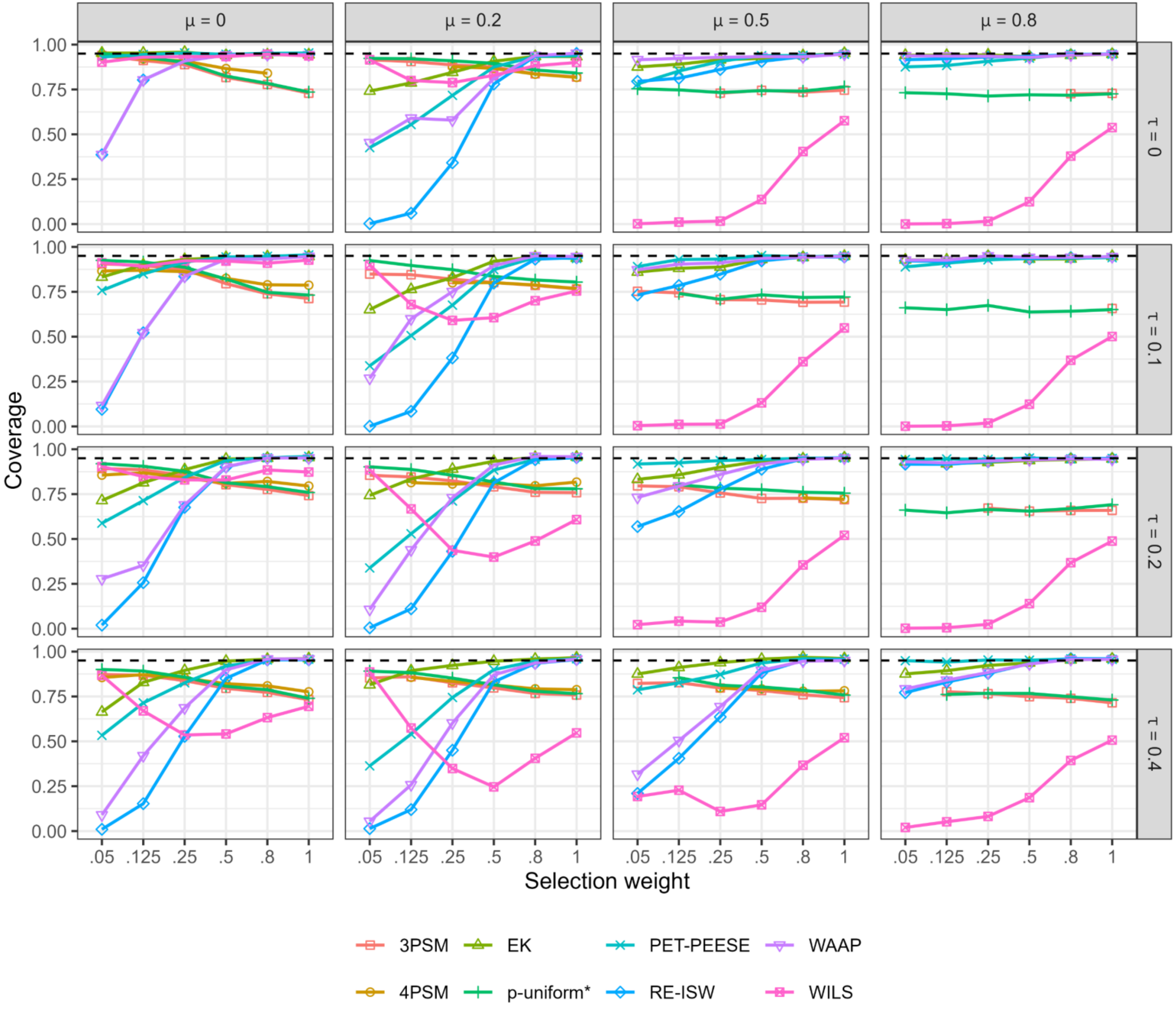
Full factorial with 2,304 conditions, each condition with 2000 replications



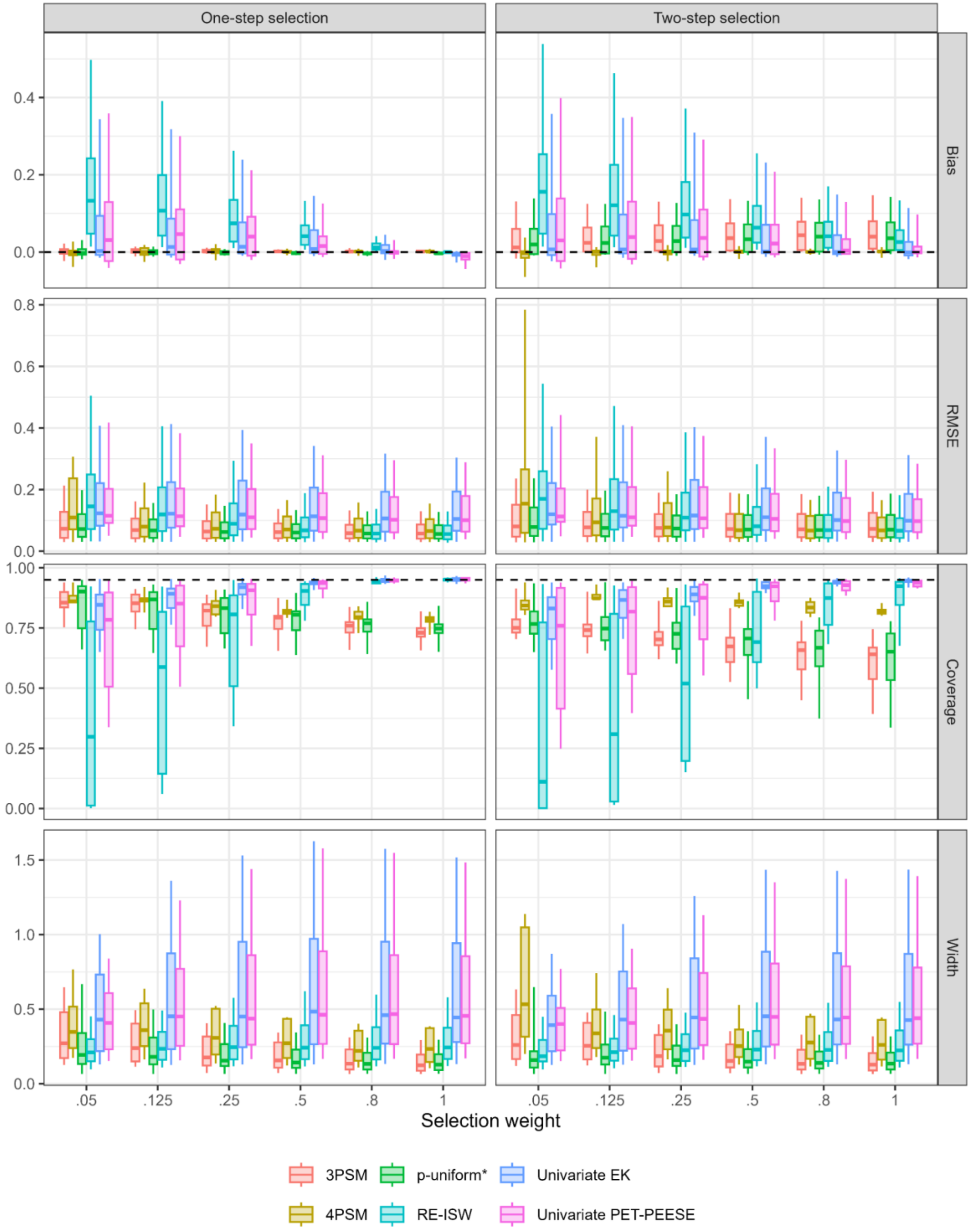
μ = 0
μ = 0.2
μ = 0.5
μ = 0.8
τ = 0
τ = 0.1
τ = 0.2
τ = 0.4
Bias
Selection weight
.05
.125
.25
.5
.8
1
3PSM
EK
PET-PEESE
WAAP
4PSM
p-uniform*
RE-ISW
WILS



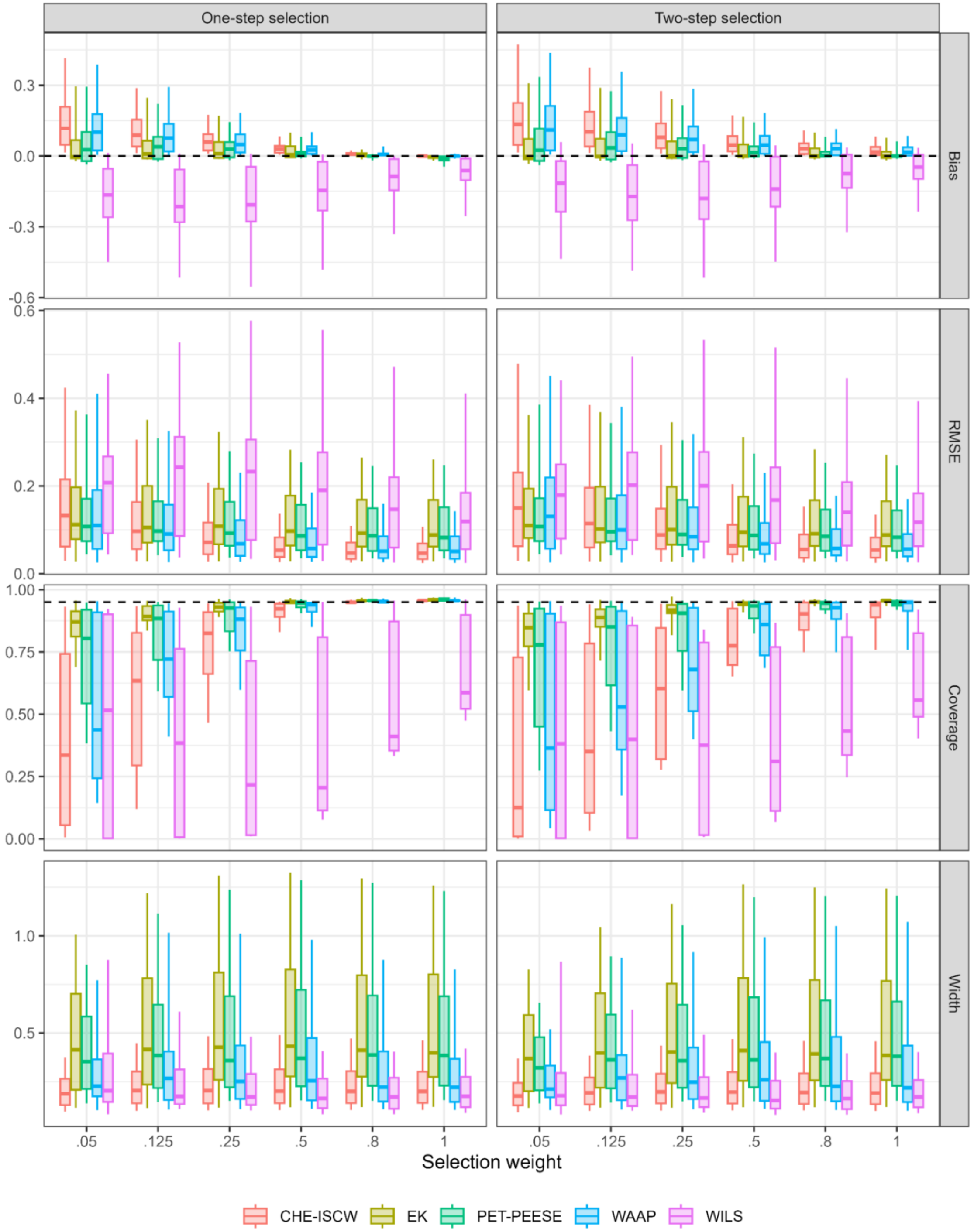
μ = 0
μ = 0.2
μ = 0.5
μ = 0.8
τ = 0
τ = 0.1
τ = 0.2
τ = 0.4
RMSE
Selection weight
.05 .125 .25 .5 .8 1
3PSM
EK
PET-PEESE
WAAP
4PSM
p-uniform*
RE-ISW
WILS



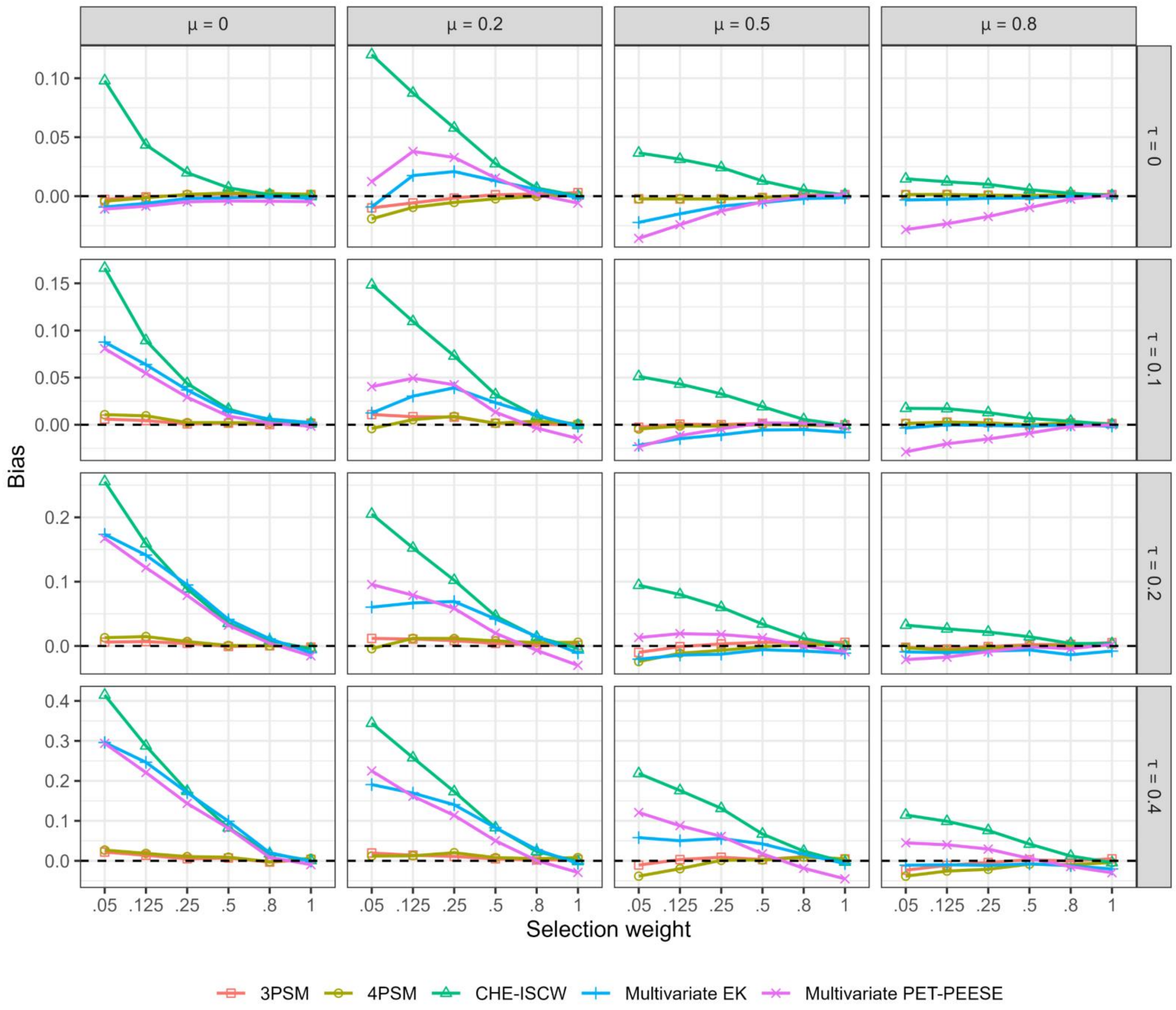
μ = 0
μ = 0.2
μ = 0.5
μ = 0.8
τ = 0
τ = 0.1
τ = 0.2
τ = 0.4
Coverage
1.00
0.75
0.50
0.25
0.00
.05 .125 .25 .5 .8 1
Selection weight
3PSM
EK
PET-PEESE
WAAP
4PSM
p-uniform*
RE-ISW
WILS

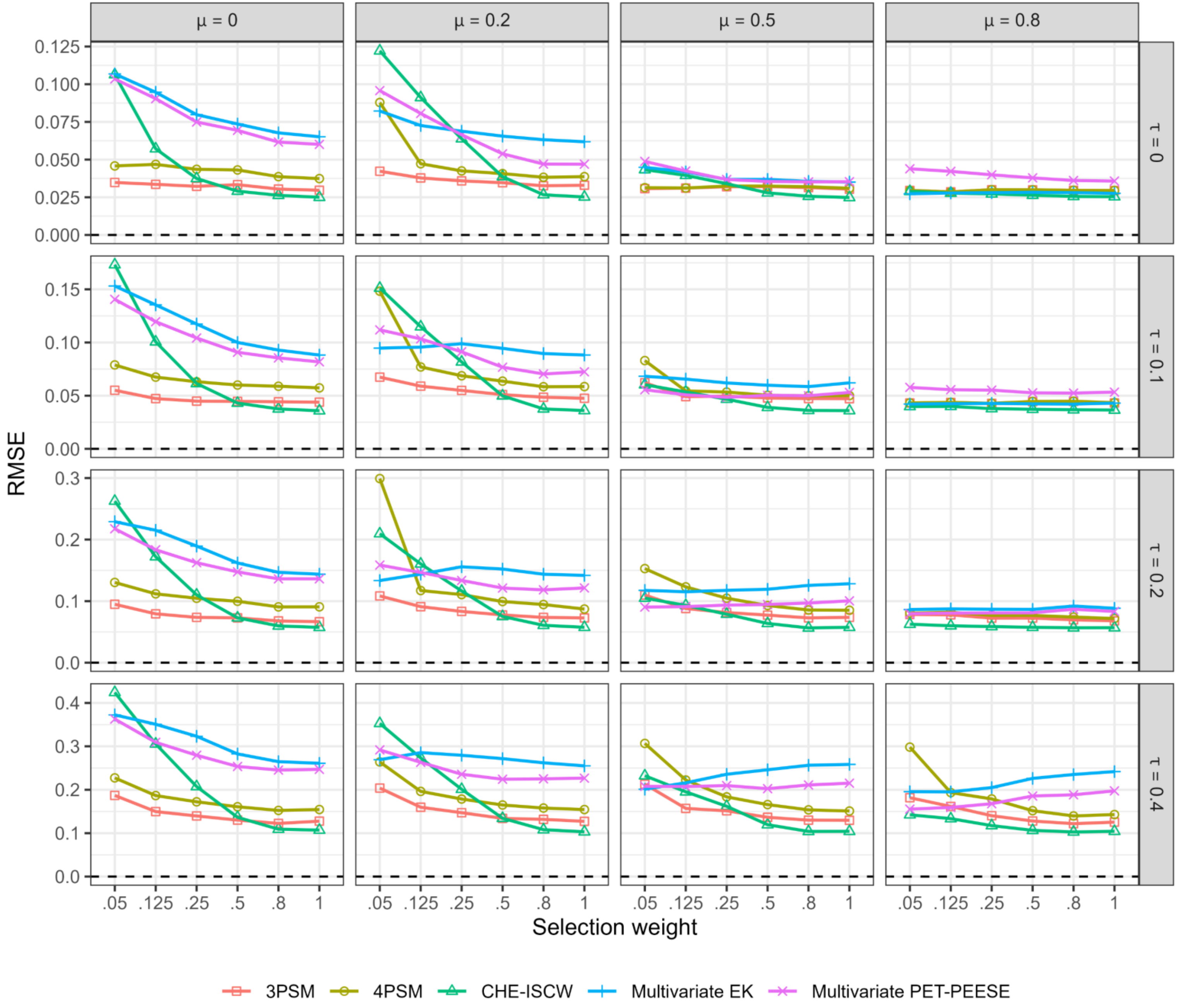


One-step selection
Two-step selection
Bias
RMSE
Coverage
Width
0.4
0.2
0.0
0.8
0.6
1.00
0.75
0.50
0.25
0.00
1.5
1.0
0.5
.05
.125
.25
.5
.8
1
Selection weight
3PSM
4PSM
p-uniform*
RE-ISW
Univariate EK
Univariate PET-PEESE

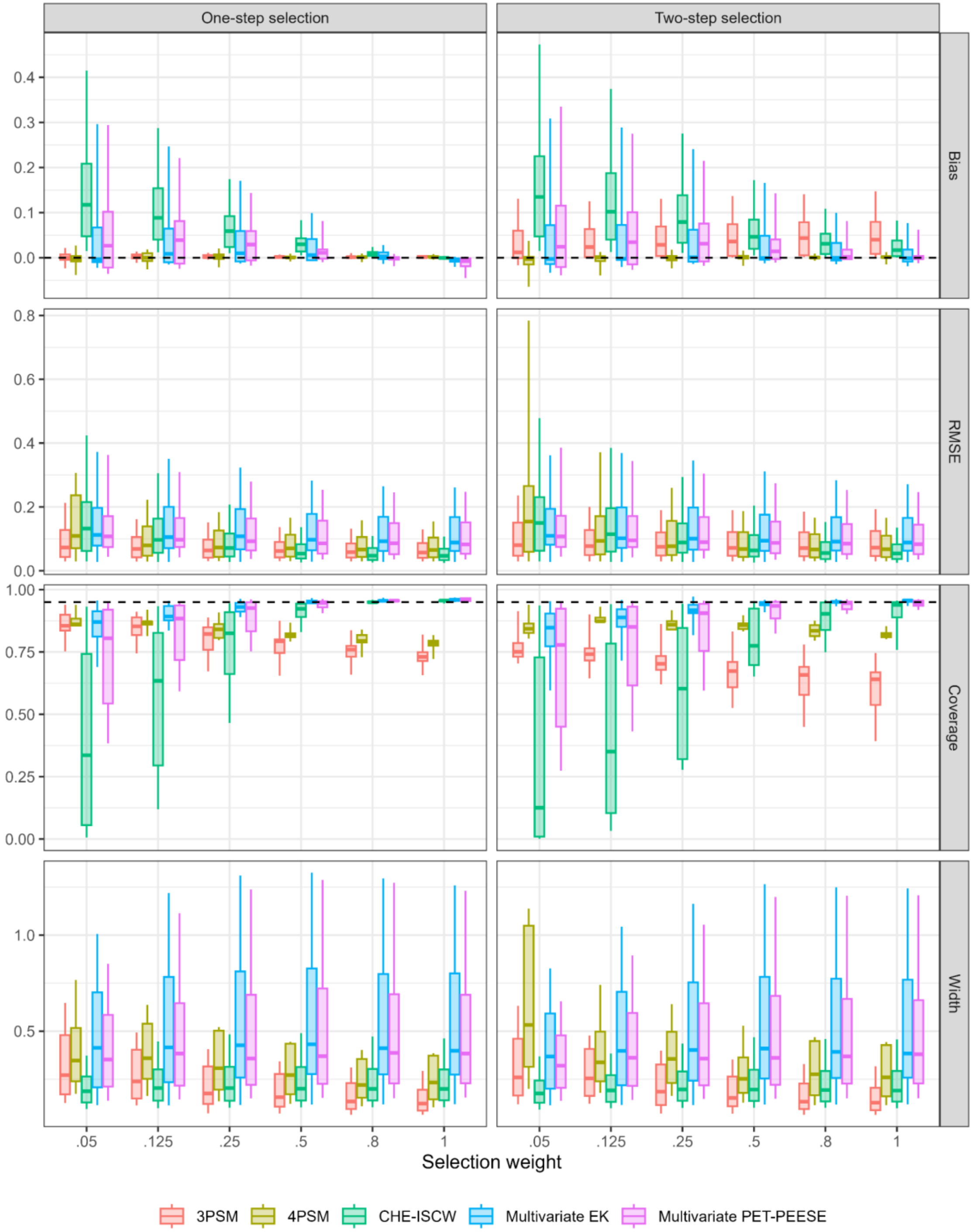


One-step selection
Two-step selection
Bias
RMSE
Coverage
Width
Selection weight
CHE-ISCW
EK
PET-PEESE
WAAP
WILS


μ = 0
μ = 0.2
μ = 0.5
μ = 0.8
τ = 0
τ = 0.1
τ = 0.2
τ = 0.4
Bias
0.10
0.05
0.00
0.15
0.2
0.1
0.0
0.4
0.3
.05 .125 .25 .5 .8 1
Selection weight
3PSM
4PSM
CHE-ISCW
Multivariate EK
Multivariate PET-PEESE



μ = 0
μ = 0.2
μ = 0.5
μ = 0.8
τ = 0
τ = 0.1
τ = 0.2
τ = 0.4
RMSE
Selection weight
.05 .125 .25 .5 .8 1
3PSM
4PSM
CHE-ISCW
Multivariate EK
Multivariate PET-PEESE



One-step selection
Two-step selection
Bias
RMSE
Coverage
Width
0.4
0.3
0.2
0.1
0.0
0.8
0.6
0.4
0.2
0.0
1.00
0.75
0.50
0.25
0.00
1.0
0.5
.05
.125
.25
.5
.8
1
Selection weight
3PSM
4PSM
CHE-ISCW
Multivariate EK
Multivariate PET-PEESE