DEPARTMENT OF ECONOMICS AND FINANCE

SCHOOL OF BUSINESS AND ECONOMICS

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

Performance of Bias-Detection Methods in Psychological Meta-Analysis: A Large-Scale Simulation Study

Sanghyun Hong
W. Robert Reed
R.C.M van Aert
M.A.L.M van Assen

WORKING PAPER

No. 3/2026

Department of Economics and Finance
UC Business School
University of Canterbury
Private Bag 4800, Christchurch
New Zealand

WORKING PAPER No. 3/2026

Performance of Bias-Detection Methods in Psychological Meta-Analysis: A Large-Scale Simulation Study

Sanghyun Hong[1]
W. Robert Reed[1]
R.C.M van Aert[2]
M.A.L.M van Assen[2†]

June 2026

Abstract: Meta-analyses are routinely assessed for publication bias, yet the relative performance of bias-detection methods under realistic conditions remains unclear, particularly when bias may arise from both selective publication and questionable research practices (QRPs). We present a large-scale simulation study evaluating 21 commonly used bias-detection methods across 864 conditions varying in true effect size, effect size heterogeneity, number of studies, publication bias, and QRPs. Data were generated using the simulation framework of Carter et al. (2019), which jointly models publication bias and QRPs. Results show that heterogeneity is the primary determinant of method performance because it strongly influences false positive rates of many methods. When heterogeneity was low to moderate, TESS performed best overall. When heterogeneity was high, selection models performed best in small to medium meta-analyses, and caliper tests performed best in large meta-analyses. Precision–based methods, including variants of Egger's test, performed poorly and often exhibited inflated false-positive rates under heterogeneity. These findings suggest that bias-detection methods should be selected based on heterogeneity and the number of studies rather than applied routinely, with TESS providing a strong general-purpose test across many settings.

Accompanying Shiny app: A Shiny app that allows the reader to explore additional results can be found here: https://pbiasmethods.shinyapps.io/main/
Accompanying R package: pbiasr is a lightweight source package that allows one to apply the 21 publication-bias methods described in this paper to the user's own data. Details are provided here: https://github.com/sho-125/BiasDetectionMethods

Data availability statement: Data and code to reproduce the results in this paper are posted here: https://osf.io/tvryp

JEL Categories: C12, C15, C18, C52, C87
Acknowledgements: This work was made possible by the use of the RCH facilities (DOI:10.18124/CANTERBURYNZ-UCRCH, RRID:SCR_027870) at the University of Canterbury.

[1] Department of Economics and Finance & UCMeta, University of Canterbury, NEW ZEALAND
[2] Department of Methodology and Statistics, Tilburg University, NETHERLANDS

† Corresponding author: Marcel van Assen. Email: M.A.L.M.vanAssen@tilburguniversity.edu

Meta-analyses play a central role in cumulative science. By synthesizing results across primary studies, they are widely used to summarize evidence, guide theory development, and inform policy and practice (Borenstein et al., 2009). However, the validity of a meta-analysis depends critically on whether the available studies constitute a representative sample of the underlying research evidence. When this assumption is violated, meta-analytic conclusions may be systematically biased. Consequently, the ability to detect bias is fundamental to the credibility of meta-analytic evidence.

One prominent source of distortion is publication bias, whereby studies yielding statistically significant results are more likely to be published than studies with nonsignificant or unfavorable findings (Rothstein et al., 2005). Such selective publication has long been known to inflate estimated effect sizes (Lane & Dunlap, 1978), distort estimates of effect size heterogeneity (Augusteijn et al., 2019; Jackson, 2007), and compromise statistical inference (Ioannidis, 2008). Evidence from multiple sources suggests that publication bias and selective reporting occur in psychology and related fields (Bartoš et al., 2024; Dickersin, 2005; Franco et al., 2014). Yet large-scale reviews of meta-analyses report relatively low rates of detected publication bias (e.g., Wu et al., 2026; Wu and Reed, 2026), highlighting the possibility that commonly used bias-detection methods may lack power or perform poorly under realistic conditions.

A second, closely related source of distortion arises from questionable research practices (QRPs), such as optional stopping, selective outcome reporting, and flexible analytic decisions (Simmons et al., 2011). These practices alter the distribution of effect sizes and *p*-values, particularly around conventional significance thresholds, and can bias meta-analytic estimates even when all conducted studies are ultimately published (Hartgerink et al., 2016; Maassen et al., 2025). Survey research has shown that self-admitted usage of QRPs is substantial among psychologists in the United States, Italy, and Brazil (Agnoli et al., 2017; John et al., 2012; Rabelo et al., 2020).

Because publication bias and QRPs can operate simultaneously and often generate similar statistical patterns in meta-analytic data, distinguishing between these sources of bias is difficult in practice. We therefore adopt the perspective of an applied researcher: methods are evaluated based on their ability to detect the presence of bias, regardless of its source.

Understanding the performance of bias-detection methods is important for several reasons. First, concern about publication bias (e.g., Bartoš et al., 2024; Fanelli, 2012) and QRPs s(e.g., Allum et al., 2023; Gopalakrishna et al., 2022; John et al., 2012) has been raised across many disciplines. Second, reporting guidelines such as the PRISMA statement (Page et al.,

2021) and the Meta-Analysis Reporting Standards (MARS; Appelbaum et al., 2018) recommend that meta-analyses routinely assess publication bias. Consistent with these recommendations, between 58% and 90% of published meta-analyses report some form of bias assessment (Wu et al., 2026).

Third, researchers now have a large number of bias-detection methods available to them – some widely used (e.g., funnel plots and Egger's regression test), others less so (e.g., caliper tests and tests of excess statistical significance) – complicating the decision of which one(s) to use.

Previous simulation studies have provided important insights into the performance of individual bias-detection methods (e.g., Begg & Mazumdar, 1994; Deeks et al., 2005; Francis, 2012; Kromrey & Rendina-Gobioff, 2006; Macaskill et al., 2001; Pustejovsky & Rodgers, 2019; Renkewitz & Keiner, 2019; Rücker et al., 2011; Schwarzer et al., 2002; Sterne et al., 2000; van Aert et al., 2016; van Assen et al., 2015). These studies show that meta-analyses with fewer than 10 studies often lack power to detect bias and that effect size heterogeneity can inflate Type I error rates and reduce statistical power for bias-detection methods.

However, this literature has several important limitations. Most studies examine only a small subset of available bias-detection methods and focus exclusively on publication bias while ignoring the role of QRPs. A notable exception is Renkewitz and Keiner (2019), who evaluated the performance of six tests using simulations that incorporated both publication bias and the QRP of optional stopping (i.e., repeatedly adding data until a statistically significant result is obtained). In addition, existing studies typically restrict attention to relatively narrow ranges of data-generating conditions. As a result, applied researchers have limited practical guidance when choosing among competing bias-detection methods.

The present study addresses these limitations by conducting a large-scale simulation-based comparison of bias-detection methods from an applied perspective. Our main contribution is to provide condition-specific recommendations for selecting bias-detection methods based on observable characteristics of a meta-analysis, particularly I^2 and the number of studies. To this end, we adopt the simulation framework developed by Carter et al. (2019). The Carter et al. framework is well suited to this study because it jointly models publication bias and QRPs across a wide range of realistic conditions while intentionally avoiding alignment with the assumptions of any specific bias-detection method. Its use is further facilitated by the availability of publicly available data and materials provided by Carter et al. (2019), including a project website (https://osf.io/rf3ys) and interactive tools (http://www.shinyapps.org/apps/metaExplorer/). A final reason for adopting this framework is

that we were not involved in its development, thereby providing an independent testing environment that reduces the risk of tailoring the simulation design to favor specific methods.

Our analysis evaluates the performance of 21 commonly used bias-detection tests while systematically varying true effect size, effect size heterogeneity, number of studies, publication bias, and the use of QRPs. Nineteen of these tests were selected via a literature search conducted in July 2021 using the keyword "publication bias test". Two additional tests related to the MAIVE estimators were added following the recent publication of Irsova et al. (2025). In contrast to Renkewitz and Keiner (2019), we examine the impact of four QRPs rather than only optional stopping.

We find that effect size heterogeneity, more than the number of studies, is the primary determinant of comparative performance because it strongly influences false positive rates of many methods. When heterogeneity is low to moderate, the test of excess statistical significance (*TESS*; Stanley et al., 2021) performs best overall. When heterogeneity is high, selection models (notably the three-parameter selection model) perform best in small- to medium-sized meta-analyses, whereas caliper-based methods perform best when the number of studies is large. In contrast, precision–based tests, including Egger-type regressions, perform poorly and exhibit inflated false positive rates under heterogeneity. These findings suggest that bias-detection methods should be selected based on heterogeneity and the number of studies, helping applied meta-analysts balance the risks of false positives and false negatives.

The remainder of this paper is organized as follows. We first describe the simulation environment, including the data-generating process, the modeling of QRPs and publication bias, and the construction of the meta-analytic datasets. We then introduce the 21 bias-detection methods evaluated in the study and define the performance metrics used to compare them. Next, we present results on applicability and convergence before reporting comparative performance across small, medium, and large meta-analyses stratified by heterogeneity. The discussion section interprets the main findings, offers condition-specific recommendations for applied researchers, and addresses limitations. The paper concludes by summarizing the study's contributions and discussing their implications for research practice.

Methods

Simulation environment

In this section, we describe the simulation environment used to generate the meta-analytic datasets and evaluate the performance of bias-detection methods. A more detailed

description of the simulation framework is provided by Carter et al. (2019). The simulation process can be viewed as a four-stage pipeline: (a) generation of a primary study, (b) potential application of QRPs, (c) selective publication to introduce publication bias, and (d) aggregation into a meta-analytic dataset.

Simulation framework. Using publicly available code provided by Carter et al. (2019), we created 864 (4 × 3 × 8 × 3 × 3) simulation conditions, defined by all combinations of true effect size (δ), heterogeneity (τ), number of primary studies (K), degree of QRPs, and extent of publication bias. For each simulation condition, we generated 1,000 meta-analytic datasets.

` Table 1 summarizes the simulation factors and their levels. The values for true effect size and heterogeneity were adopted from Carter et al., who calibrated these parameters to match empirical characteristics commonly observed in psychological research. We deviated from Carter et al. by varying the number of studies per meta-analysis across eight levels (K = 5 to 800), compared with their four levels (10, 30, 60, 100). This extension allows us to evaluate method performance across a wider range of meta-analytic sizes, including very small meta-analyses, which are common in some medical contexts (Hong & Reed, 2026), and very large meta-analyses, which are more typical in fields such as economics (Wu et al., 2026).

Table 1: Factors in the simulation design with their levels

Factor	Levels	Number of Levels
Mean true effect (δ)	{0.0, 0.2, 0.5, 0.8}	4
Heterogeneity (τ)	{0.0, 0.2, 0.4}	3
Number of studies (K)	{5, 10, 30, 60, 100, 200, 400, 800}	8
QRPs	{none, medium, high}	3
Publication bias	{none, medium, high}	3

Creation of raw data for individual studies. Before applying QRPs and publication selection, each primary study was generated following Carter et al. First, the sample size per group was drawn from a truncated inverse gamma distribution (truncated at n = 5 and n = 1,905; shape = 1.15326986, scale = 0.04622745). Control and treatment observations were then generated as $Y_{Ci} \sim N(0,1)$ and $Y_{Ti} \sim N(\mu, 1)$, where $\mu \sim N(\delta, \tau^2)$. Group means and variances were computed, and an independent two-sample *t*-test was performed to obtain the *p*-value and the study-level effect size (Cohen's *d*; see Box 1 of Carter et al., 2019). This data-generating process produces study-level effect sizes that vary across studies due to

heterogeneity (τ) while maintaining the interpretation of effects as standardized mean differences.

Introduction of QRPs. Studies were evaluated by a simulated researcher who could apply QRPs in an attempt to obtain statistical significance. The researcher first tested the primary hypothesis using the originally specified analysis. If the result was statistically significant in the expected direction (two-tailed α = .05), the process stopped and the result was reported as originally specified. If the initial result was not statistically significant, the researcher iteratively applied QRPs in an attempt to obtain a statistically significant result, following either a medium-QRP strategy or a strong-QRP strategy. This procedure mimics researcher degrees of freedom in which analytic decisions are conditioned on statistical significance.

The medium-QRP strategy consisted of two QRPs: the use of two weakly correlated dependent variables (r = .2) and up to three additional data-collection efforts, with three observations per group added in each effort. The strong-QRP strategy included these same two QRPs but allowed up to five (instead of three) additional data-collection efforts and added two further QRPs: optional removal of outliers (observations with absolute z-scores greater than 2) and conditioning on a binary moderator that was uncorrelated with the dependent variable following a significant interaction effect. See Carter et al. for a detailed description of these QRPs and the order in which they were applied. The medium- and strong-QRP strategies resulted in average Type I error rates of .09 and .27 at the primary study level. Not all studies in a meta-analysis in case of a QRP condition were impacted by QRPs. In the medium-QRP condition, studies were generated using no QRPs in 30% of cases, the medium-QRP strategy in 50% of cases, and the strong-QRP strategy in 20% of cases. In the strong-QRP condition, the corresponding proportions were 10%, 40%, and 50%. These mixtures reflect heterogeneity in researcher behavior, rather than assuming that all researchers engage in QRPs to the same extent.

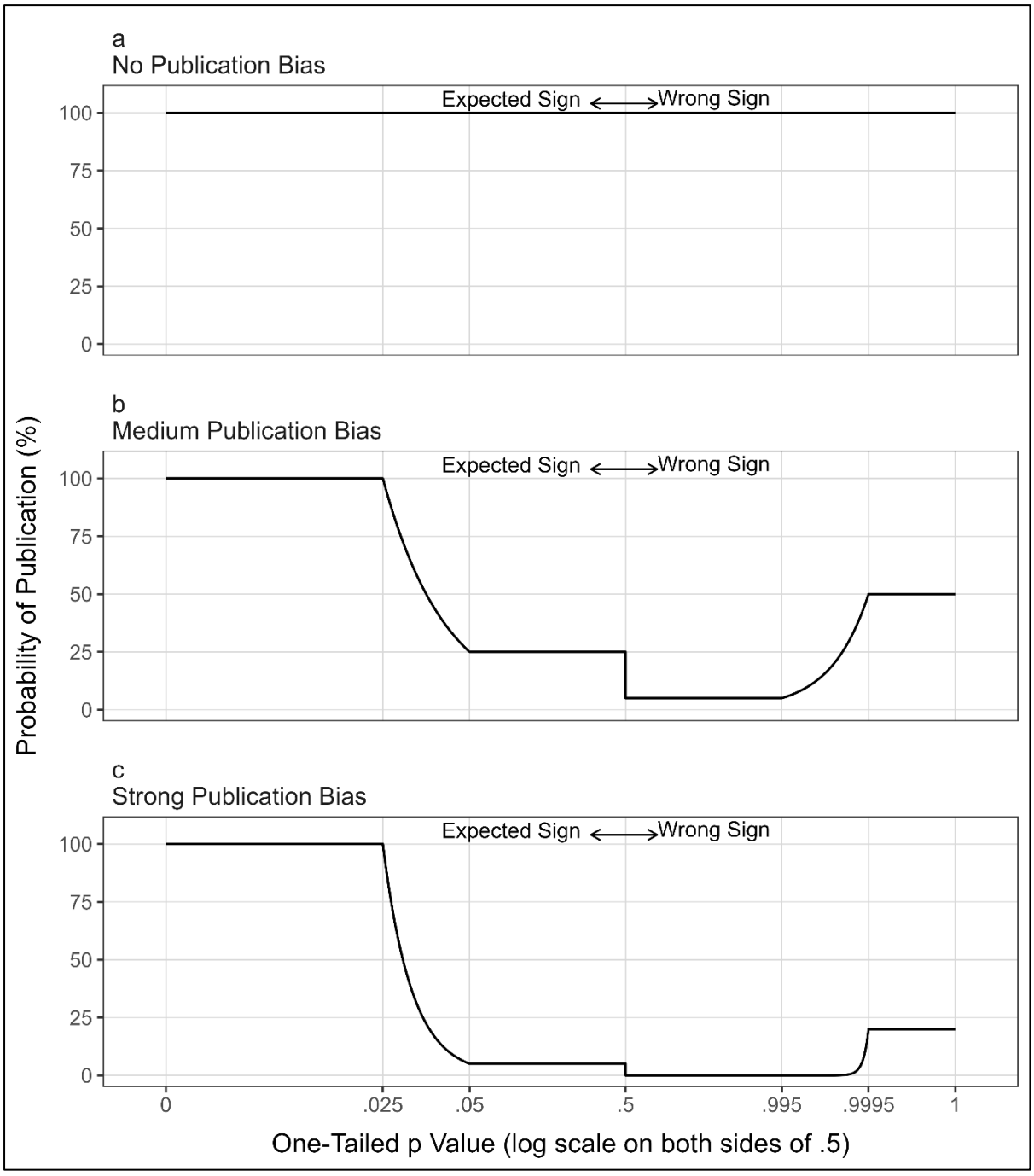
Introduction of publication bias (selection). In the third stage, studies were subjected to a censoring function that acted as a gatekeeper, simulating selective publication due to peer review and the file-drawer process. The probability of publication depended on both the sign of the estimated effect and its one-tailed p-value, with the steepness of the censoring function determined by the specified level of publication bias (medium or high). Two weight functions, w_{high} and w_{medium}, were used to model publication probability as a function of the one-tailed p-value:

w_1	for $0 \leq p < .025$
$w_2 = f_1(p)$	for $.025 \leq p < .05$
w_3	for $.05 \leq p < .5$
w_4	for $.5 \leq p < .995$
$w_5 = f_2(p)$	for $.995 \leq p < .9995$
w_6	for $.9995 \leq p \leq 1$

For the high publication bias condition, statistically significant studies were always published, whereas non-significant studies were published rarely or not at all, with $w_1 = 1$, $w_3 = .05$, $w_4 = 0$, and $w_6 = .2$. For medium publication bias, statistically non-significant studies were published at a higher rate, with $w_1 = 1$, $w_3 = .25$, $w_4 = .05$, and $w_6 = .5$. Parameters w_2 and w_5 were defined as two different exponential functions of the *p*-value that connected the adjoining intervals (see Figure 1). Importantly, the publication selection mechanism did not correspond to the assumptions of any of the selection models evaluated in this study (Andrews & Kasy, 2019; Vevea & Hedges, 1995; van Assen et al., 2015).

Bundling the individual studies into meta-analytic datasets. For each simulation condition, this procedure was repeated until the target number of studies (*K*) was reached, at which point the retained study-level effect sizes constituted a single meta-analytic dataset. These simulated meta-analyses served as the evaluation datasets for the bias-detection methods examined in this study, allowing direct comparison of method performance across a wide range of meta-analytic conditions. We next describe the bias-detection methods that we assessed. The methods are discussed in detail in the Supplementary Materials, including the R-code to apply them.

Figure 1: Publication bias schedule relating probability of publication as a function of one-tailed p-values. This is a copy of Figure 2 in Carter et al., (2019).



Bias detection methods

A wide range of statistical tests have been proposed to detect publication bias and related distortions in meta-analytic data, but their relative performance under realistic conditions remains an open question. Although these methods differ in their assumptions, data

requirements, and data characteristics for detecting bias, they can be broadly grouped into four classes.

Precision–based tests examine whether primary study effect sizes are systematically related to their precision (e.g., standard error or functions of sample size). Such a relationship is expected when statistically significant results are more likely to be published (nine tests). *Selection models* assume that the probability of publication depends on study characteristics such as a study's *p*-value and jointly model the effect-size distribution and the selection process (five models). *Excess statistical significance tests* compare the observed number of statistically significant results to the number expected based on estimated study-level statistical power (three tests). Finally, *caliper tests* assess discontinuities in the distribution of *p*-values by comparing the frequency of just-significant and just-nonsignificant results (four tests). In total, 21 bias-detection tests were analyzed in this study. These tests are listed in Table 2 and briefly described below, with additional technical details provided in the Supplementary Materials. Unless explicitly noted otherwise, all tests were implemented using a two-tailed significance level of $\alpha = .05$.

Precision–based tests. These tests are used to detect small-study effects (Sterne et al., 2011) and are often interpreted as tests for publication bias. Under publication bias, primary study results are more likely to be published when they are statistically significant, that is, when $\hat{d}_i / se(\hat{d}_i)$ exceeds a critical value (typically around 2, corresponding to a one-tailed *p*-value of .025). As a result, in the presence of publication bias or QRPs, studies with lower precision are expected to report larger observed effect sizes; thus, a negative relation between effect size and its precision is expected, or positive relation between effect size and standard error.

Begg and Mazumdar's (1994) rank-correlation test (BEGG) belongs to this category because it tests the association between primary study standardized effect sizes and their standard errors using Kendall's tau. We also included five variants of Egger's (1997) regression test (EGGER1–EGGER5), which regress observed effect sizes on precision. The last two variants use an instrumental variable–type approach with sample size serving as a predictor (Irsova et al., 2025). The endogenous-kink test (EK; Bom & Rachinger, 2019) is another regression-based variant of Egger's approach that uses a spline specification, allowing the relationship between effect size and standard error to operate only above a prespecified threshold value of the standard error.

Whereas the Egger and EK methods detect bias using tests of the regression coefficient linking effect size to standard error, the remaining two methods—SKEW1 and SKEW2 (Lin

& Chu, 2018) evaluate asymmetry in the residuals from the corresponding regression model. SKEW1 tests for skewness in the residual distribution, whereas SKEW2 combines a test for residual skewness with a test of the slope in an Egger-type regression. Under publication bias, residuals are expected to be right-skewed and the regression slope is expected to be positive.

Selection models. Selection models detect bias by explicitly modeling the probability that a study is observed (i.e., published) as a function of its statistical significance. These models treat the observed meta-analytic sample as a selectively filtered subset of a larger population of studies and jointly estimate parameters governing both the effect-size distribution and the selection mechanism. Across models, selection is represented using weight functions that assign different publication probabilities to regions of the *p*-value distribution. Evidence of bias is inferred when the estimated selection probabilities vary across regions of statistical significance, indicating that some results are more likely to be published than others.

The 3- and 4-parameter selection models (PSM3 and PSM4; Iyengar & Greenhouse, 1988; Vevea & Hedges, 1995; Vevea & Woods, 2005) as well as the 3- and 5-parameter Andrews and Kasy (2019) models (AK1 and AK2) assume a normal distribution of true effect sizes, whereas *p*-uniform (PUNIF; van Assen et al., 2015) assumes a fixed-effect model. The models differ in the weight functions used to represent publication selection, which depend on the *p*-value of an effect size. The PSM models assume directional publication bias with weights $\{1, w_1\}$ for intervals $\{p \leq .025, p > .025\}$ for PSM3; and $\{1, w_1, w_2\}$ for $\{p \leq .025, .025 < p \leq .5, p > .5\}$ for PSM4. Note that these *p*-values are one-tailed and the weight of effect sizes with $p \leq .025$ is set to 1 to identify the model. The AK models assume symmetrical (bidirectional) publication bias. The AK1 model uses weights $\{1, w_1\}$ for intervals $\{p \leq .05, p > .05\}$ based on two-tailed p-values, whereas the AK2 model uses weights $\{1, w_1, w_2, w_3, w_4\}$ for intervals $\{p \leq .025, .025 < p \leq .5, .5 < p \leq .975, p > .975\}$ based on one-tailed *p*-values. *p*-uniform assigns equal weight to all statistically significant effect sizes but ignores all non-significant results. For the PSM3, PSM4, AK1, and AK2 models, publication bias is inferred when the estimated weights differ from 1, as determined by a likelihood ratio test. In contrast, the publication bias test in *p*-uniform applies Fisher's test to the distribution of conditional *p*-values.

As noted above, the selection weight functions in the simulation design do not correspond to the selection mechanisms assumed by PSM3, PSM4, AK1, or AK2. Accordingly, the simulations do not favor these selection models through alignment between the data-generating process and model assumptions. We also note that van Assen et al. (2015) caution against applying *p*-uniform in the presence of effect-size heterogeneity, which characterizes

most scenarios in our simulation study and under which *p*-uniform has been shown to perform poorly.

Tests of excess statistical significance. This category assesses excess statistical significance (ESS) by comparing the number of statistically significant results to the number expected based on study-level statistical power. The logic of these tests is that, regardless of whether bias arises from publication bias or QRPs, both processes increase the proportion of statistically significant results relative to what would be expected given the underlying effect sizes and sample sizes (Stanley et al., 2021, p. 778). The first approach, TES (also known as exploratory test, Ioannidis & Trikalinos, 2007), computes expected power using the meta-analytic average effect and sampling errors but does not account for effect size heterogeneity. Stanley et al. (2021) introduced PSST and TESS to improve upon TES by explicitly incorporating heterogeneity into the calculation of expected significance. Like TES, PSST identifies bias by comparing the observed proportion of statistically significant findings to the expected proportion. If π and P denote the expected and observed proportion of statistically significant studies, with ESS = P - π, then the test statistic of the proportion of statistical significance test (PSST) is calculated as

$$Z_{PSST} = \frac{ESS}{\pi(1-\pi)/K} \quad ,$$

and the test statistic of the test of excess statistical significance (TESS) is calculated as

$$Z_{TESS} = \frac{ESS - .05}{.05(1-.05)/K}$$

where TESS tests whether the excess significance rate exceeds .05. Primary-study results are classified as statistically significant and positive using $z_{0.975} = 1.96$. The resulting Z_{PSST} and Z_{TESS} statistics are then compared to $z_{0.95} = 1.645$, the standard-normal critical value for a one-sided test with $\alpha = .05$.

Caliper-based tests. Caliper-based tests detect discontinuities in the distribution of reported *p*-values around conventional significance thresholds (Gerber & Malhotra, 2008; Schneck, 2017). These tests compare the frequency of just-significant primary study results to the frequency of just-nonsignificant results using a proportion one-tailed test with $\pi = .5$. Under publication bias or QRPs, the number of just-significant results is expected to exceed the number of just-nonsignificant results. The four caliper-based tests differ in the width of the interval examined. CALI05, CALI10, CALI15, and CALI20 compare frequencies within intervals equal to ±5%, ±10%, ±15%, and ±20% of the critical z-value, respectively. When the critical value is $z = 1.96$ and results are expressed as two-sided *p*-values, these intervals

correspond to *p*-value ranges of (.040–.050, .050–.063), (.031–.050, .050–.078), (.024–.050, .050–.096), and (.019–.050, .050–.117), respectively.

Table 2. Description of Bias Detection Tests

CATEGORY	TEST	BRIEF DESCRIPTION
Precision–Based Tests	BEGG	Rank-correlation test assessing the association between standardized effect sizes and their standard errors (Begg & Mazumdar, 1994).
	EGGER1	Fixed-effects Egger regression testing whether the slope differs from zero in a regression of effect sizes on standard errors (Egger et al., 1997).
	EGGER2	Random-effects Egger regression allowing for between-study heterogeneity in the residual variance (Egger et al., 1997).
	EGGER3	Egger-type regression estimated by weighted least squares (WLS), using weights, $w_i = 1/SE_i^2$ (Egger et al., 1997; Stanley & Doucouliagos, 2014).
	EGGER4	Instrumental variable-style Egger regression using fitted standard errors derived from sample size (Irsova et al., 2025).
	EGGER5	Instrumental variable-style Egger regression using fitted standard errors derived from sample size and inverse-variance weights (Irsova et al., 2025)
	EK	Similar to an Egger regression except that the standard error is entered as a spline function (Bom & Rachinger, 2019).
	SKEW1	Skewness test evaluating asymmetry in residuals from Egger's regression (Lin & Chu, 2018).
	SKEW2	Combined test using both the slope of Egger's regression and residual skewness for improved detection power (Lin & Chu, 2018).
Selection Models	PSM3	Three-parameter selection model estimating a step-function publication process conditional on statistical significance (Iyengar & Greenhouse, 1988; Vevea & Hedges, 1995).
	PSM4	Four-parameter selection model allowing asymmetric publication probabilities across significance regions (Vevea & Woods, 2005).
	AK1	Andrews–Kasy selection model with symmetric selection around zero (Andrews & Kasy, 2019).
	AK2	Andrews–Kasy selection model allowing asymmetric selection for positive and negative results (Andrews & Kasy, 2019).
	PUNIF	Tests whether the conditional distribution of statistically significant *p*-values deviates from uniformity (van Assen et al., 2015).

CATEGORY	TEST	BRIEF DESCRIPTION
Excess Statistical Significance Tests	TES	Test for Excess Significance comparing observed and expected numbers of statistically significant results (Ioannidis & Trikalinos, 2007).
	PSST	Proportion of statistically significant studies test comparing observed and expected significance rates (Stanley et al., 2021).
	TESS	Proportion of statistically significant studies test comparing the difference between the proportions of observed and expected significances with 5% (Stanley et al., 2021).
Caliper Tests	CALI05	Caliper test using a ±5% window around the significance threshold (Gerber & Malhotra, 2008; Schneck, 2017).
	CALI10	Caliper test using a ±10% window around the significance threshold (Gerber & Malhotra, 2008; Schneck, 2017).
	CALI15	Caliper test using a ±15% window around the significance threshold (Gerber & Malhotra, 2008; Schneck, 2017).
	CALI20	Caliper test using a ±20% window around the significance threshold (Gerber & Malhotra, 2008; Schneck, 2017).

Performance measures

We evaluate the performance of bias-detection methods using three complementary performance measures: *Balanced Accuracy* (*BA*), *Distance*, and *Relative Performance*. We illustrate these measures using a simple artificial example. For each of the 864 simulation conditions, we compute each method's success rate. In bias conditions, the success rate equals the proportion of true positives (TP), or statistical power. In no-bias conditions, the success rate equals the proportion of true negatives (TN), or 1 − Type I error rate. The top panel of Table 3 reports success rates for three hypothetical methods (A, B, and C) across three conditions—two with bias and one without bias.

To interpret these performance measures, it is useful to consider the receiver operating characteristic (ROC) (Fawcett, 2006; Hanley & McNeil, 1982). In ROC space, the horizontal axis represents the Type I error rate (1 − TN), and the vertical axis represents the true positive rate (TP), or statistical power. Methods closer to the upper-left corner of ROC space perform better, as they achieve higher power while maintaining lower Type I error rates.

For each bias-detection method and subgroup of conditions, we compute two quantities: (i) the average true positive rate (TP) across all conditions with bias and (ii) the average Type I error rate (1 − TN) across all conditions without bias. These averages define a coordinate pair in the ROC space. For Methods A, B, and C in Table 3, the average TP values are (.8 + .7)/2 =

0.75, (.7 + .9)/2 = 0.80, and (.5 + .6)/2 = 0.55, respectively. Combined with Type I error rates of 0.10, 0.10, and 0.01, this yields the following ROC coordinates: A: (.10, .75), B: (.10, .80), C: (.01, .55). Graphically, Methods A and B have similar Type I error rates, but Method B achieves higher power. Method C, by contrast, controls Type I error more strictly (0.01), albeit at the cost of substantially lower power.

Table 3: Toy example to illustrate the performance measures

Performance measure	Method A	Method B	Method C
Success rates (probability space)			
Success (Bias condition 1) = TP	0.8	0.7	0.5
Success (Bias condition 2) = TP	0.7	0.9	0.6
Success (No-bias condition) = TN	0.9	0.9	0.99
Balanced Accuracy (BA) = (Avg TP + Avg TN)/2	**0.825**	**0.85**	**0.77**
Location (log-odds space)			
Location (Bias condition 1) = log(TP/(1−TP))	1.386	0.847	0
Location (Bias condition 2) = log(TP/(1−TP))	0.847	2.197	0.405
Location (No-bias condition) = log((1−TN)/TN)	−2.197	−2.197	−4.595
Distance = Avg bias location – Avg no-bias location	**3.314**	**3.719**	**4.798**
Condition-level shortfall			
Bias condition 1: shortfall from best	0	0.539	1.368
Bias condition 2: shortfall from best	1.350	0	1.792
No-bias condition: shortfall from best	2.398	2.398	0
Relative Performance: $P_i(c)$			
Relative Performance: $P_i(c = 0)$	**1/3**	**1/3**	**1/3**
Relative Performance: $P_i(c = 1)$	**1/3**	**2/3**	**1/3**
Relative Performance: $P_i(c = 2)$	**2/3**	**2/3**	**1**

Balanced accuracy (BA). Based on these success rates, we compute Balanced Accuracy (BA) = (TP + TN) / 2. Balanced Accuracy averages power and the true negative rate, thereby giving equal weight to bias and no-bias conditions. We adopt equal weighting because the frequency of bias in practice is unknown. For example, for Method A: TP = (.8 + .7)/2 = .75, TN = .9, and BA = (.75 + .9)/2 = .825. As shown in the last row of the top panel of Table 3, Method B achieves the highest Balanced Accuracy (BA = 0.85).

Balanced Accuracy evaluates performance directly in probability space. Under this metric, a one-percentage-point improvement is treated as equally valuable regardless of where

it occurs. Thus, moving from .50 to .51 is counted the same as moving from .98 to .99. This property makes Balanced Accuracy easy to interpret, but it also implies that improvements in the middle of the probability scale are treated the same as improvements near the extremes.

Distance. Distance also relies on TP and TN but evaluates performance in log-odds (logit) space using the transformation $\log(P/(1 - P))$, where P is the probability of correctly classifying a condition. Because differences between probabilities are magnified in the tails of the distribution, the Distance metric places greater weight on methods that achieve very low Type I error and very high power compared to Balanced Accuracy. Thus, whereas Balanced Accuracy measures average classification accuracy, Distance measures how far apart the bias and no-bias conditions are on the logit scale.

To compute Distance, we first convert success rates TN and TP to locations in logit space using the logit transformation. Following recommendations (e.g., Agresti, 2013), success rates of 0 and 1 are replaced by .0005 and .9995 to avoid infinite log-odds values. Distance is then defined as the difference between the average logit location for conditions with bias and the average logit location for conditions without bias, with larger values indicating greater separation—and therefore better discrimination—between the bias and no-bias conditions. For example, for Method A, Distance = (1.386 + 0.847)/2 − (−2.197) = 3.314. Note that the largest possible distance is 15.201 when replacing 0 and 1 by .0005 and .9995, respectively.

According to Table 3, Method C achieves the largest Distance (4.798). Although Method C has lower power under bias than Method B, its much stronger performance under no-bias conditions (.99 vs. .90) produces a larger separation on the logit scale and therefore a higher Distance value.

Relative Performance: $P_i(c)$. The rows under "Condition-level shortfall" quantify how far each method falls short of the best-performing method in that specific condition, measured in logit space. The method with the highest location in that condition is treated as the benchmark and assigned a shortfall of 0. For the remaining methods, the shortfall c equals the difference between the benchmark location and the method's own location, with larger values indicating weaker discrimination relative to the best method. For example, in Bias condition 1, Method A is the benchmark because it has the highest location, with Method B's shortfall c = 1.386 − 0.847 = 0.539 and Method C's c = 1.386 − 0 = 1.386. Small shortfalls indicate that a method performs nearly as well as the best method in that condition, whereas large shortfalls indicate a meaningful loss of discrimination ability.

Relative Performance, $P_i(c)$ evaluates how frequent method *i* performed well, relative to other methods, across conditions as a function of a given threshold value, *c*. $P_i(c = 0)$ is the proportion of conditions in which method *i* performs best (i.e., has the largest Distance, so that its shortfall equals 0). Table 3 shows that each method performs best in one condition, so $P_i(c = 0)$ = 1/3 for all three methods when c = 0. Note that $\sum_i P_i(c = 0)$ may exceed 1 if there are ties. $P_i(c = 1)$ is the proportion of conditions in which method *i* performed well, that is, with that method's Distance within 1 logit unit of the best-performing method. For example, $P_B(c = 1)$ = 2/3 because Method B has a shortfall of 0.539 < 1 when compared to Method A in Bias condition 1. By varying *c* along the horizontal axis, we obtain performance curves that show not only which method performs best (c = 0), but also which methods perform close to best for a given tolerance level *c.*

To simplify interpretation of the Relative Performance curves, we focus on two values of the tolerance parameter c: $c = 0$ and $c = 1$. When $c = 0$, Relative Performance represents the percentage of simulation conditions in which a method performed best. When $c = 1$, a method is classified as relatively "good" within a given simulation condition if its performance is within 1 logit unit of the best-performing method in that condition. Expressed in terms of correct classification rates, this means that if the best-performing method correctly classifies 95%, 75%, and 60% of cases in a given condition, then another method would be considered "good" *in that condition* if it correctly classifies at least 87.5%, 52.5%, and 35.6% of cases, respectively. Because a difference of one logit unit can correspond to relatively large differences in correct classification probabilities—especially when overall classification rates are moderate—this criterion represents a relatively lenient definition of "good." The purpose of this threshold is therefore not to identify methods that perform nearly as well as the best method in absolute terms, but rather to identify methods whose performance is broadly comparable to the best method in terms of discrimination ability. In the Results section, we report methods that were classified as "good" in at least 50% of simulation conditions within each subgroup.

Summary. In summary, the three performance measures capture complementary aspects of method performance. Balanced Accuracy summarizes overall performance in probability space by weighting power and Type I error equally. Distance evaluates performance in logit space and measures how well a method discriminates between bias and no-bias conditions by quantifying the separation between these two conditions, thereby placing greater emphasis on methods that achieve very high power and very low Type I error. Relative

Performance evaluates how often a method performs best or near-best across simulation conditions and therefore reflects the robustness of a method's performance. Taken together, these measures provide a more complete picture of performance than any single metric alone.

Presentation of results: Matching performance to observable characteristics.

A central objective of our study is to provide practical guidance on which bias-detection tests perform best under specific observable conditions. Based on previous research, we do not expect a single test to dominate across all settings (Carter et al., 2019; Hong & Reed, 2021). Accordingly, following Hong & Reed (2021), we categorize meta-analyses along two observable dimensions—the number of studies (K) and effect heterogeneity (measured by I^2)—to identify the best-performing methods within each category. We focus on these two dimensions because they are directly observable to researchers, unlike features such as the true presence of publication bias, and because prior research has shown that these characteristics have been shown to influence the performance of meta-analytic methods (e.g., Carter et al., 2019; Hong & Reed, 2021).

Table 4 categorizes meta-analyses into nine groups based on the number of studies (Small, Medium, and Large) and effect heterogeneity (Small, Medium, and Large, based on I^2). The nine groups were created such that each contained approximately equal numbers of meta-analyses. Each group is identified by a two-letter code (e.g., SL, MS, LM), where the first letter indicates the number of studies and the second letter indicates the level of effect heterogeneity. Thus, "SL" refers to meta-analyses with a small number of primary studies ($K \leq 10$) but a large degree of effect heterogeneity ($I^2 \geq 0.60$).

We report two numbers in each (K, I^2) cell. The first number represents the total number of simulation environments (meta-analysis types) in that cell, and the second number represents the number of those environments without QRPs. Thus, for "SS", the entry "115/35" indicates that there are 115 simulation environments corresponding to meta-analyses with $K \leq 10$ and $I^2 < 0.29$, of which 35 were simulated without QRPs.

Table 4. Nine categories of simulation environments according to the values of K and I^2.

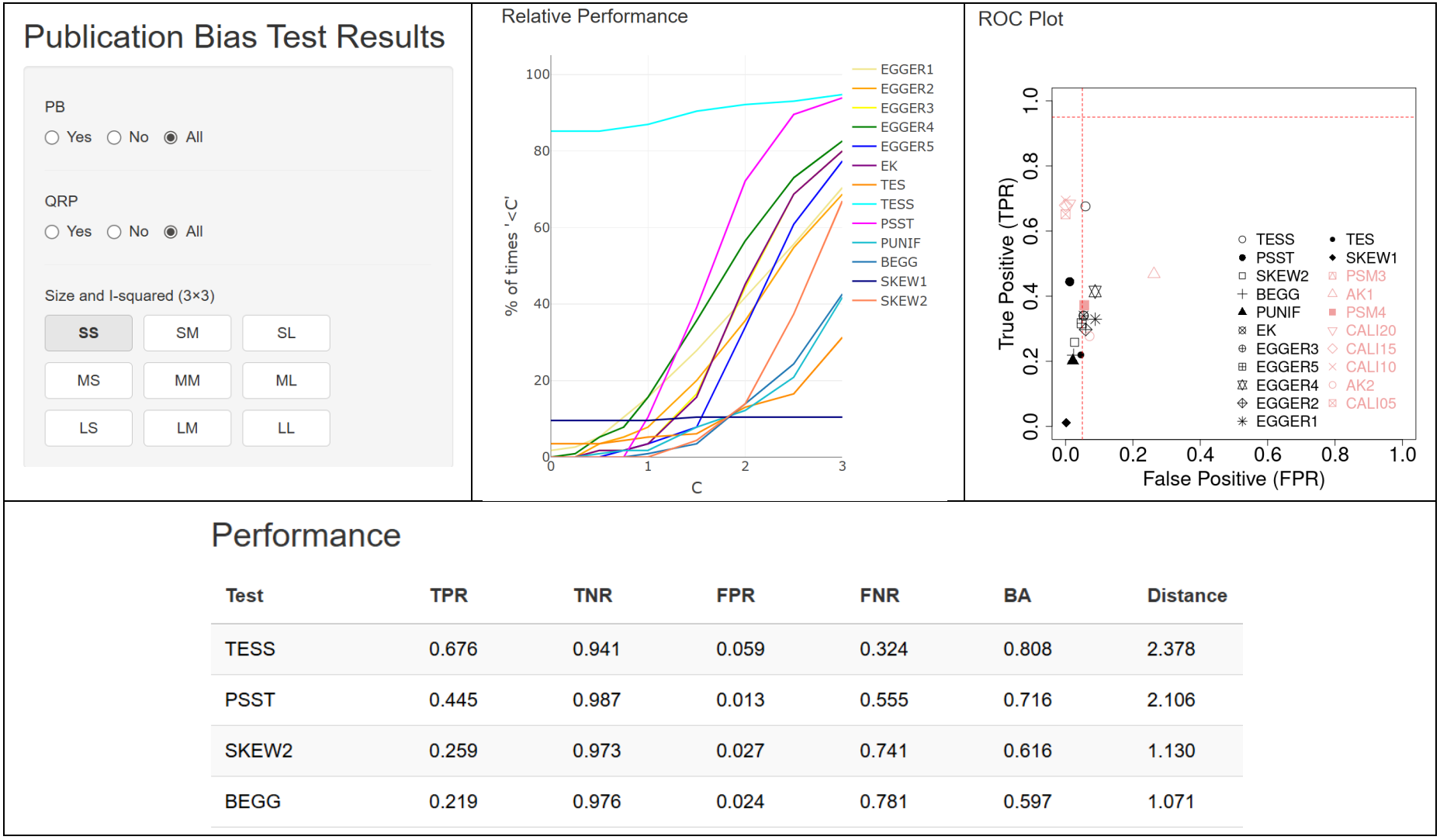
		Heterogeneity (I^2)		
		Small ($I^2 < 0.29$)	Medium ($0.29 \leq I^2 < 0.60$)	High ($I^2 \geq 0.60$)
Number of Studies (K)	Small ($K \leq 10$)	115/35	81/29	105/36
	Medium ($10 < K \leq 100$)	89/33	105/33	91/28
	High ($K > 100$)	86/33	100/33	92/28

We discuss the results for all 864 simulation conditions, grouped according to (K, I^2) categories. We do this for two reasons. First, our goal is to evaluate relative performance of bias-detection methods under realistic research conditions, where multiple sources of bias can occur. Second, presenting full subsets for every possible combination of conditions would substantially increase the length, detail, and complexity of our discussion, obscuring the main patterns we wish to highlight. For readers interested in exploring more detailed subsets of conditions, we provide an interactive Shiny application (see Figure 2 for example interface and output). The app allows users to filter simulation scenarios by the presence or absence of publication bias and QRPs, allowing readers to explore results beyond the aggregated analyses presented here.

Results

We begin by reporting applicability (i.e., whether a method's minimum data requirements are satisfied) and convergence rates for all methods across the nine scenarios. If a method exhibits convergence or applicability below 90% in a given scenario, it is excluded from the performance comparisons for that scenario for two reasons. First, methods that frequently fail to produce results are of limited practical value and cannot be recommended for applied use. Second, because nonconvergent or inapplicable iterations must be discarded, performance metrics computed only on the remaining iterations would be biased, potentially making such methods appear more effective than methods that successfully return results even in more difficult cases. We then report performance results—Balanced Accuracy, Distance, and Relative Performance—separately for small (S), medium (M), and large (L) meta-analyses, with results further stratified by small, medium, and large levels of heterogeneity.

Figure 2: Example Shiny application interface and results display (https://pbiasmethods.shinyapps.io/main/**)**



Performance

Test	TPR	TNR	FPR	FNR	BA	Distance
TESS	0.676	0.941	0.059	0.324	0.808	2.378
PSST	0.445	0.987	0.013	0.555	0.716	2.106
SKEW2	0.259	0.973	0.027	0.741	0.616	1.130
BEGG	0.219	0.976	0.024	0.781	0.597	1.071

Convergence and applicability

Convergence issues arise when a method's estimation does not yield a stable solution, as can occur for likelihood-based approaches such as selection models. Applicability refers to whether a method's minimum data requirements are satisfied, which may fail when there are insufficient numbers of studies in the relevant *p*-value intervals. For example, selection models require at least one effect size in each interval of the selection function. For caliper tests, Type I error and power are both effectively zero when four or fewer effect sizes fall within the combined caliper intervals (i.e., $(1/2)^4 > \alpha = .05$), so we require at least five effect sizes for these tests to be applicable. These requirements limit the applicability of selection models and caliper tests, particularly in small meta-analyses. For example, the CALI20 test considers only effect sizes with *p*-values in the interval (.019, .117). The expected proportion of effect sizes in this interval is approximately .09 when the true effect is zero and at most .342 when the true effect size is between 1.6 and 1.65. Thus, even under favorable conditions, a meta-analysis typically requires at least 15 studies for five or more effect sizes to fall within the CALI20 interval.

Table 5 presents applicability and convergence rates for all methods across the nine categories of meta-analyses (SS, SM, SL, …, LM, LL). Each percentage represents the proportion of simulation iterations in which the method produced a valid test result, that is, the method was applicable and the estimation procedure converged. Grey-shaded cells indicate cases where applicability/convergence fell below the 90% threshold, and these methods were excluded from subsequent performance comparisons for that scenario.

The nine precision–based methods and the three excess statistical significance methods showed no applicability or convergence problems, with rates exceeding 90% in all scenarios and often reaching 100%. The same was true for *p*-uniform. In contrast, selection models and caliper tests exhibited substantial applicability and convergence limitations in several scenarios. The applicability/convergence rate for PSM3 fell below 90% in the SS, SM, MS, and ML categories, and it was therefore excluded from performance comparisons in those scenarios. Similarly, PSM4 met the applicability/convergence criterion only in the LL category. AK1 met the criterion only in ML and the large meta-analysis categories (LS, LM, LL), whereas AK2 was excluded from all performance comparisons due to consistently poor applicability/convergence. As expected, caliper methods also showed limited applicability in some scenarios, with applicability/convergence rates increasing with the width of the caliper interval (e.g., CALI20 > CALI05), as well as with the number of studies (*K*) and the level of

heterogeneity (I^2). These results indicate that, in practice, many selection-model and caliper-based methods are not usable in small meta-analyses, which has important implications for applied researchers choosing among bias-detection procedures.

Table 5: Applicability/Convergence Rates

	SS	SM	SL	MS	MM	ML	LS	LM	LL
BEGG	100	100	100	100	100	100	100	100	100
EGGER1	100	100	100	100	100	100	100	100	100
EGGER2	99.1	99.1	99.4	95.2	93.7	99.5	96.3	95.7	99.8
EGGER3	100	100	100	100	100	100	100	100	100
EGGER4	100	100	100	100	100	100	100	100	100
EGGER5	100	100	100	100	100	100	100	100	100
EK	100	100	100	100	100	100	100	100	100
SKEW1	99.9	99.9	100	99.1	99.4	100	99.5	100	100
SKEW2	99.9	99.9	100	99.1	99.4	100	99.5	100	100
PSM3	39.5	58.7	91.4	82.0	87.1	98.0	97.3	98.7	99.9
PSM4	7.7	21.6	67.6	39.2	38.3	82.9	57.6	57.7	91.6
AK1	45.8	59.3	89.7	84.5	87.5	97.5	98.0	98.7	99.8
AK2	0.5	3.0	32.8	11.3	10.6	47.2	24.7	19.3	55.1
PUNIF	95.3	97.1	99.5	99.1	99.9	100	100	100	100
TES	100	100	100	100	100	100	100	100	100
PSST	100	100	100	100	100	100	100	100	100
TESS	100	100	100	100	100	100	100	100	100
CALI05	0.3	0.2	24.8	43.5	54.2	89.3	98.7	99.5	100
CALI10	2.4	1.9	54.1	73.7	79.9	99.2	100	100	100
CALI15	6.2	5.4	68.4	86.5	90.4	99.9	100	100	100
CALI20	10.6	9.8	76.0	93.0	95.4	100	100	100	100

NOTE: The values in the table represent the percent of iterations where the respective bias detection method is both appliable and convergent. Grey-shaded areas indicate that applicability/convergence rates are less than 90%.

Performance comparison: Small meta-analyses ($K \leq 10$)

We begin with small meta-analyses ($K \leq 10$), where differences between bias-detection methods are most pronounced and where method choice is therefore most consequential. Figures 3 and 4 present the performance metrics for the 21 bias-detection tests across the three small meta-analysis subgroups. In both figures, results are displayed in vertically ordered panels corresponding to increasing levels of effect heterogeneity: small heterogeneity ($I^2 < 0.29$) at the top (SS), medium heterogeneity ($0.29 \leq I^2 < 0.60$) in the middle (SM), and large heterogeneity ($I^2 \geq 0.60$) at the bottom (SL).

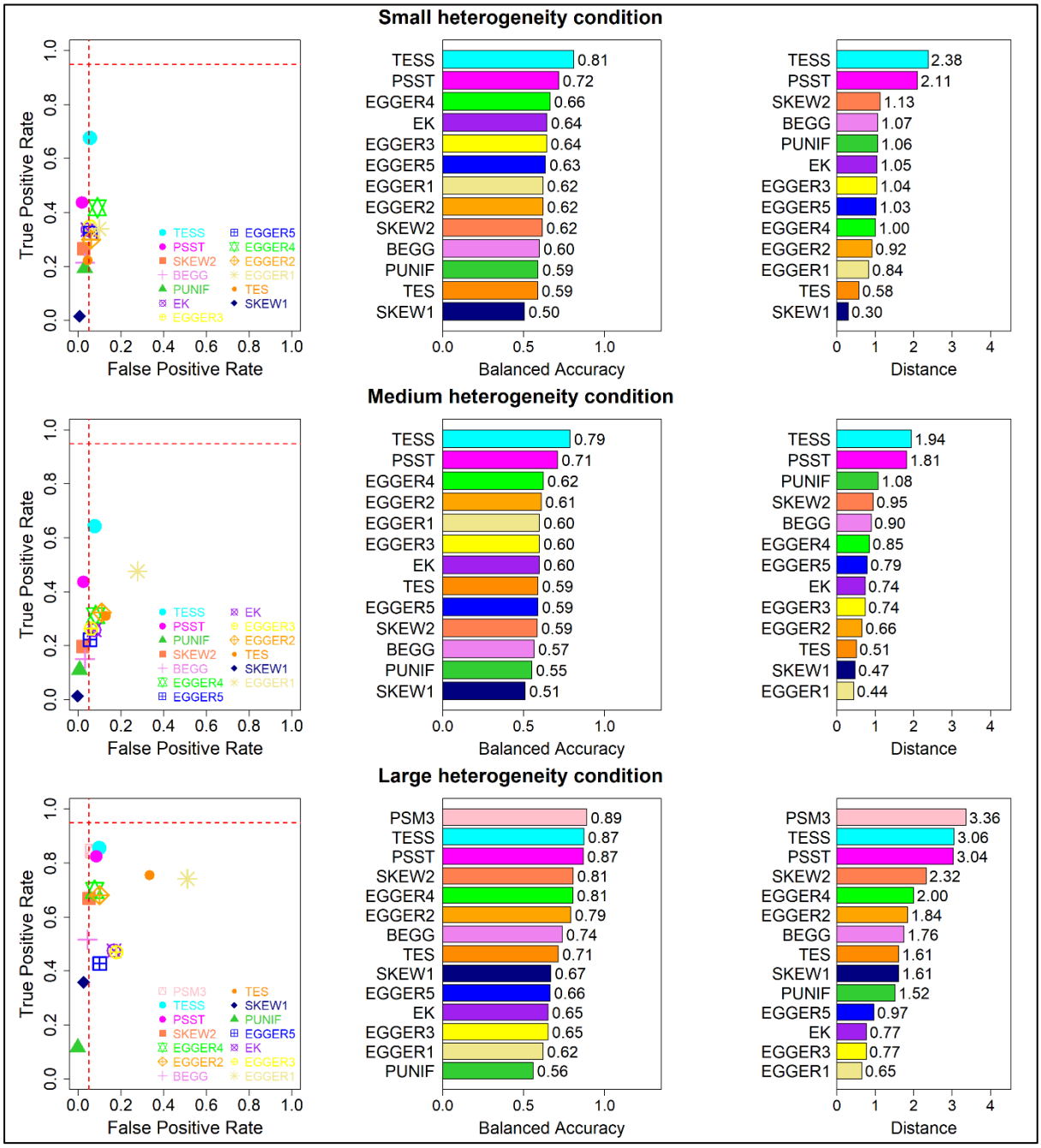
Figure 3 shows that for SS and SM, the excess statistical significance methods TESS (ranked 1) and PSST (ranked 2) clearly outperformed the other methods in terms of both Balanced Accuracy and Distance. The difference between the two methods is primarily due to their true positive rates. TESS had TP of 0.676 and 0.650 for SS and SM, respectively. The corresponding values for PSST were 0.445 and 0.443. The results also show that TES performed substantially worse than TESS and PSST, indicating that accounting for heterogeneity when estimating expected significance is important for excess-significance tests.

Although PSM3 could not be evaluated in the SS and SM scenarios due to applicability and convergence limitations, it performed best in the SL scenario, closely matching TESS and PSST on true positive rate, while having the lowest false positive rate of the three (.059). Both TESS and PSST saw their false positive rates increase under high heterogeneity (.105 and .084, respectively). Relatedly, increasing heterogeneity inflated false positive rates for most precision–based methods, in some cases to unacceptably high levels (e.g., .511 for EGGER1). This result highlights an important interaction between heterogeneity and method performance: when heterogeneity is large, the selection model PSM3 (and as we shall see later, caliper tests with wider intervals) gains performance relative to excess-significance and precision–based approaches.

Figure 4 shows the Relative Performance curves for the three subgroups. For both SS (Figure 4a) and SM (Figure 4b), TESS dominated the other methods, performing best in 85% (SS) and 78% (SM) of all conditions. Using the more lenient criterion of $c = 1$, indicated by the vertical line in the figure, TESS remained among the best-performing methods in 87% (SS) and 85% in (SM) of conditions. No other test met the criterion for being classified as "good" in at least 50% of simulation conditions. SKEW1 performed poorly overall; however, it occasionally ranked best in no-bias conditions because its false positive rate was close to zero, resulting in high true negative rates but very low power.

Figure 4c shows that PSM3, TESS, and PSST outperformed all other methods under high heterogeneity, performing best in 49%, 43%, and 26% of conditions, respectively (note that the sum exceeds 100% because ties were possible). Using the more lenient criterion of c = 1, these same three methods remained dominant, ranking within one logit unit of the best-performing methods in 71%, 67%, and 60% of conditions, respectively. No other method was classified as "good" in at least half of the simulation conditions.

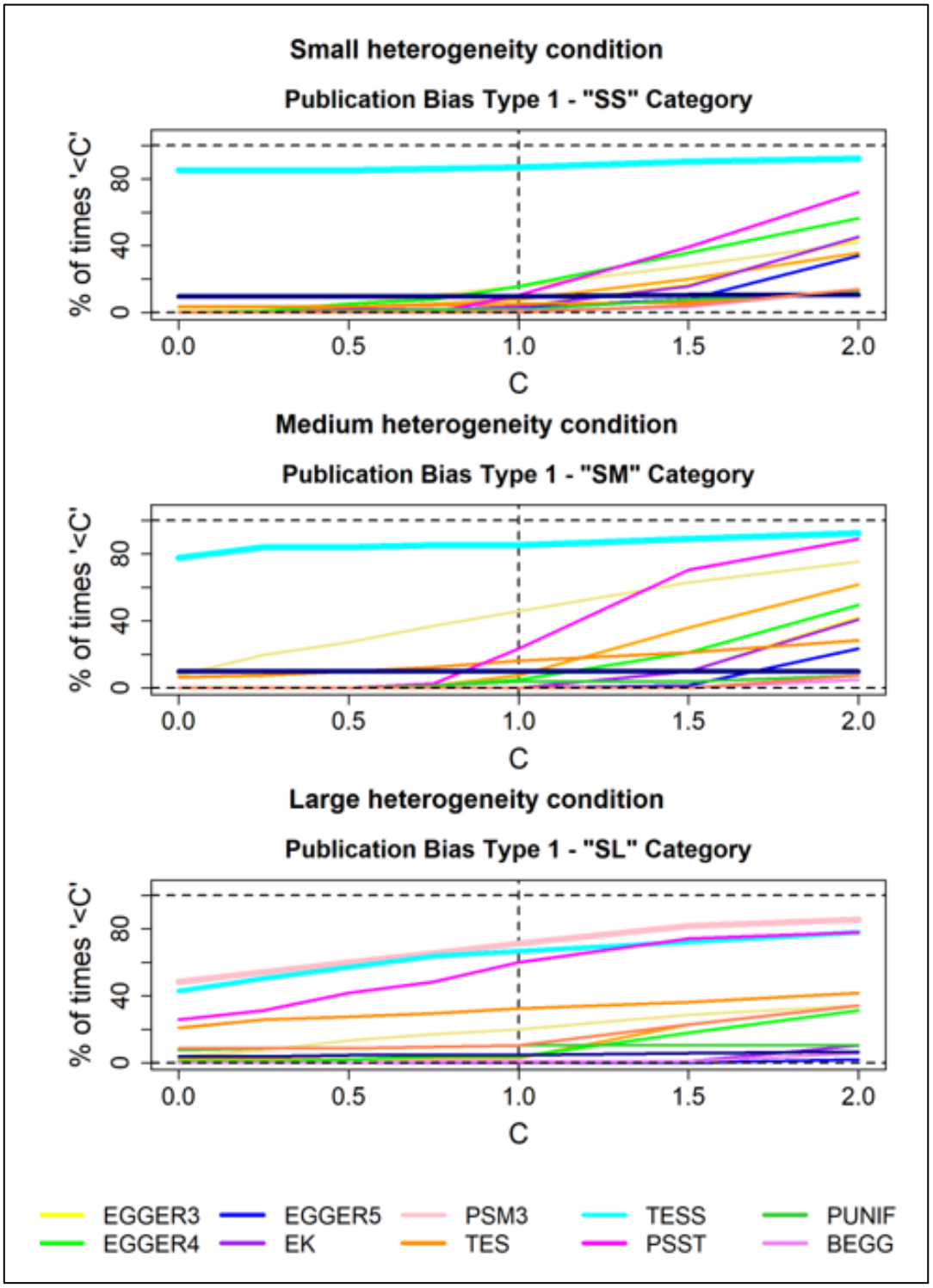
Figure 3: ROC curve and bar graphs of Balanced Accuracy and Distance of bias detection methods for Small meta-analyses, $K \leq 10$.



NOTE: The three panels report methods' performance for (from top to bottom) Small heterogeneity, Medium heterogeneity, Large heterogeneity, and with respect to (from left to right) ROC space, Balanced Accuracy, and Distance.

.

Figure 4: Distance curves of bias detection methods for small meta-analyses: small heterogeneity (SS), medium heterogeneity (SM), large heterogeneity (SL)



NOTE: This figure plots Relative Performance curves as a function of the distance parameter, c The Relative Performance statistic, $P_i(c)$, measures the proportion of instances where the respective test is within c logit units of the best-performing test. $P_i(c = 0)$, equals the proportion of times in which method i performed best, $P_i(c = 1)$ equals the proportion of times in which that method was within distance 1 of the best method.

Performance comparison: Medium meta-analyses ($10 < K \leq 100$)

We next examine medium-sized meta-analyses ($10 < K \leq 100$), where increased sample size improves the ability of most methods to detect bias. Results are reported in Figures 5 and 6. As in the previous section, the panels are ordered vertically according to increasing heterogeneity: MS, MM, and ML. Compared with small meta-analyses, performance improves markedly across most methods as the number of studies increases: true positive rates rise and separation in ROC space (Distance) becomes more pronounced, indicating improved discrimination between bias and no-bias conditions.

In subgroup MS (Figure 5a), PSST and TESS again outperformed the other methods with respect to Balanced Accuracy (.958 and .937, respectively) and Distance (4.857 and 5.171, respectively). Compared with small meta-analyses, performance improved substantially for these and most other methods as the number of studies increased, reflecting higher power and improved separation between bias and no-bias conditions. False positive rates remained below .05 for both (.020 for PSST and .018 for TESS). SKEW1 and *p*-uniform performed substantially worse than the other methods in this subgroup.

With medium heterogeneity (MM, Figure 5b), CALI20 joined TESS and PSST as one of the best-performing methods, though TESS ranked first on both Balanced Accuracy and Distance. Performance differences across methods were substantial in this subgroup, indicating that method choice remains important even in medium-sized meta-analyses when heterogeneity is present. Among the precision-based methods, the Egger variants performed relatively poorly, whereas SKEW2 and BEGG, together with CALI15, ranked approximately fourth through sixth.

At large heterogeneity ($I^2 > 0.60$, ML, Figure 5c), we observe a shift similar to that seen in the small-*K*, large-heterogeneity scenarios. PSM3 moves to the top of the ROC plots and achieves the highest Balanced Accuracy and Distance values, followed by CALI20, while TESS ranks third and PSST somewhat lower. This confirms the pattern seen with small meta-analyses: as heterogeneity increases, selection models, and now caliper methods with wide intervals, become increasingly competitive. In this setting, performance differences are driven less by gains in power—which are now generally high across methods—and more by differences in false positive rate control. Methods that better accommodate heterogeneity retain separation in ROC space, whereas methods that are more sensitive to between-study variability shift rightward due to inflated Type I error.

Precision–based methods again perform relatively poorly in ML. Although their power improves relative to small meta-analyses, this gain does not translate into better overall discrimination because inflated false positive rates offset the gains in power. False positive rates were somewhat too low for CALI20 (.019) and CALI15 (.021), somewhat too high for TESS (.085), and closest to the nominal .05 level for PSM3 (.068). AK1 performed particularly poorly in this setting, and the relative performance of BEGG declined compared to the MS and MM scenarios.

Turning to the Relative Performance curves in Figure 6, when heterogeneity was small (MS), TESS and PSST performed best in 63% and 62% of conditions, respectively. Using the $c = 1$ criterion, TESS (83%) and PSST (88%) were among the best-performing methods in most conditions. Among the remaining methods, only TES achieved the criterion of being good in at least 50% of simulation conditions. With medium heterogeneity (MM), TESS and PSST performed even better, ranking best in 71% and 70% of conditions, respectively, and both were classified as good performers in 84% and 85% of conditions. Again, only TES (54%) was within one logit unit of the best performers in half of simulation conditions. With large heterogeneity (ML), the best-performing methods were PSM3 (81%), PSST (67%), and TESS (59%). Using the $c = 1$ criterion, these same methods were again classified as good performers—PSM3 (85%), PSST (69%), and TESS (64%)—with CALI20 (51%), TES (66%) and SKEW2 (58%) also meeting the 50% threshold.

The results for CALI15 and CALI20 reveal an apparent paradox: these methods perform well according to Balanced Accuracy and Distance but not according to the Relative Performance measure. This apparent contradiction arises because Balanced Accuracy and Distance are "pure" performance measures, in the sense that they do not depend on the prior probability of bias or on the decision criterion used to declare bias, whereas Relative Performance depends on both the prior probability of bias and the decision criterion. In our simulations, the prior probability of no bias is relatively small (8/91 = 9%) in subgroup ML. Because the caliper methods use a conservative decision rule, as reflected in their low false positive rates, they perform particularly well in no-bias conditions. In contrast, PSST uses a more liberal decision rule (i.e., a higher false positive rate), which results in higher true positive rates and therefore better performance in the majority of conditions where bias is present.

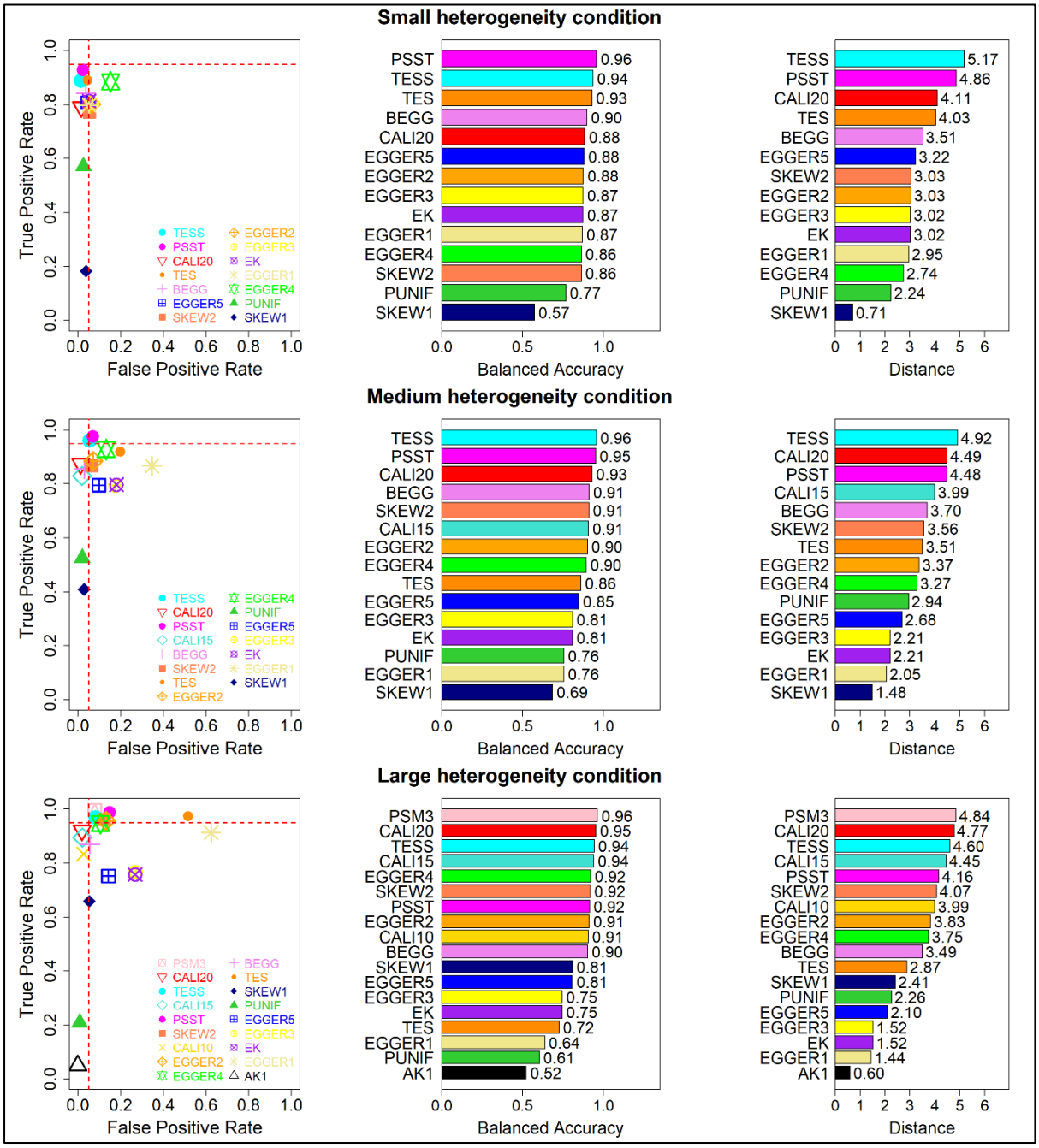
The paradox is illustrative, as shown in the following extreme example. Suppose a hypothetical bias test (called "silly") always declares bias, and suppose bias is present in 90% of conditions. It is feasible that "silly" could outperform all other methods at $c = 0$ (with a score of 90%) and could also be among the best methods at $c = 1$, even though its pure performance

measures would be very poor (Balanced Accuracy = .45 and Distance = 0). Note that Distance equals zero because the test does not distinguish between bias and no-bias conditions (it always declares bias), so its true positive rate and false positive rate are both equal to 1, resulting in no separation between bias and no-bias conditions.

This example highlights that Relative Performance curves should be interpreted with caution and in conjunction with absolute performance measures such as Balanced Accuracy and Distance. More generally, performance rankings can depend on the evaluation metric used, so Relative Performance should be interpreted alongside metric-based measures rather than in isolation.

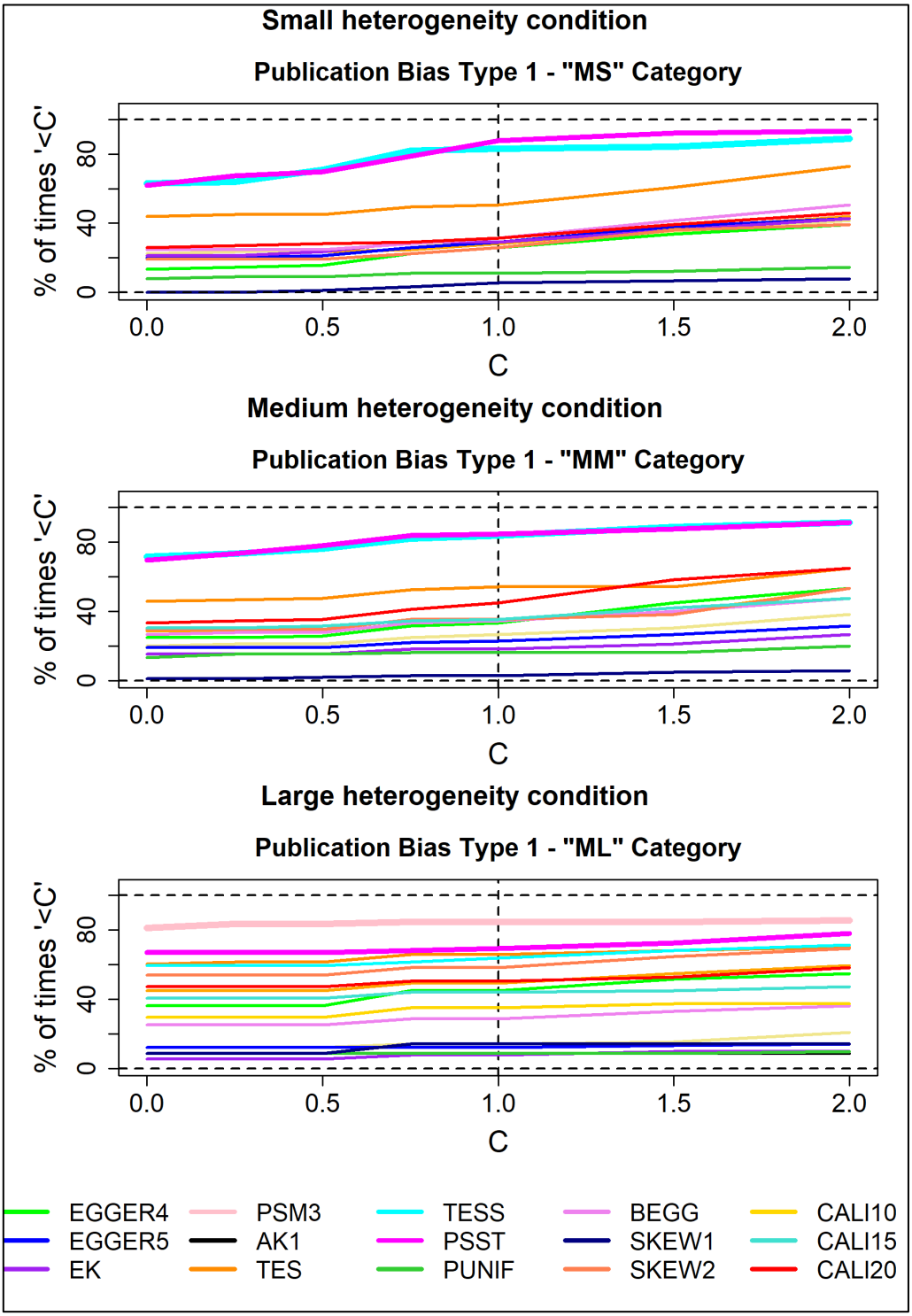
In summary, the results for medium-sized meta-analyses largely parallel those observed for small meta-analyses, with clearer separation in performance as feasibility and statistical power increase. When heterogeneity is small to moderate ($I^2 < 0.60$), TESS and PSST provide the best overall performance, both in terms of average discrimination (Balanced Accuracy and Distance) and the frequency with which they outperform competing methods. As heterogeneity exceeds 0.60, PSM3 moves into the top position, with CALI20 also emerging as a top-performing method. However, even under high heterogeneity, TESS remains competitively ranked and therefore provides a practical alternative when selection models are infeasible or fail to converge.

Figure 5: ROC curve and bar graphs of Balanced Accuracy and Distance of bias detection methods for Medium meta-analyses, $10 < K \leq 100$.



NOTE:
The three panels report methods' performance for (from top to bottom) Small heterogeneity, Medium heterogeneity, Large heterogeneity, and with respect to (from left to right) ROC space, Balanced Accuracy, and Distance.

Figure 6: Distance curves of bias detection method for Medium-sized meta-analyses: small heterogeneity (MS), medium heterogeneity (MM), large heterogeneity (ML)



NOTE: This figure plots Relative Performance curves as a function of the distance parameter, c The Relative Performance statistic, $P_i(c)$, measures the proportion of instances where the respective test is within c logit units of the best-performing test. $P_i(c = 0)$, equals the proportion of times in which method i performed best, $P_i(c = 1)$ equals the proportion of times in which that method was within distance 1 of the best method.

Performance comparison: Large meta-analyses ($K > 100$)

Figures 7 and 8 report the performance of the different bias-detection methods for large meta-analyses ($K > 100$). In contrast to small- and medium-sized meta-analyses, the defining feature of the large-K setting is that statistical power to detect bias is extremely high for most methods, and true positive rates frequently approach or equal 1 across LS, LM, and LL. As a result, differences in Balanced Accuracy and Distance are driven primarily by variation in false positive rates rather than by differences in power.

This pattern is visible across all three heterogeneity panels of Figure 7. In the LS case (low heterogeneity), most methods cluster near the upper-left corner of the ROC plots, and Balanced Accuracy values exceed .95 for many procedures. As heterogeneity increases from LS to LM and LL, methods diverge more distinctly in ROC space—not because detection power changes substantially, but because false positive rates increase for some methods while remaining low for others. Thus, in large meta-analyses, control of false positives—particularly in the presence of heterogeneity—becomes the primary determinant of method performance.

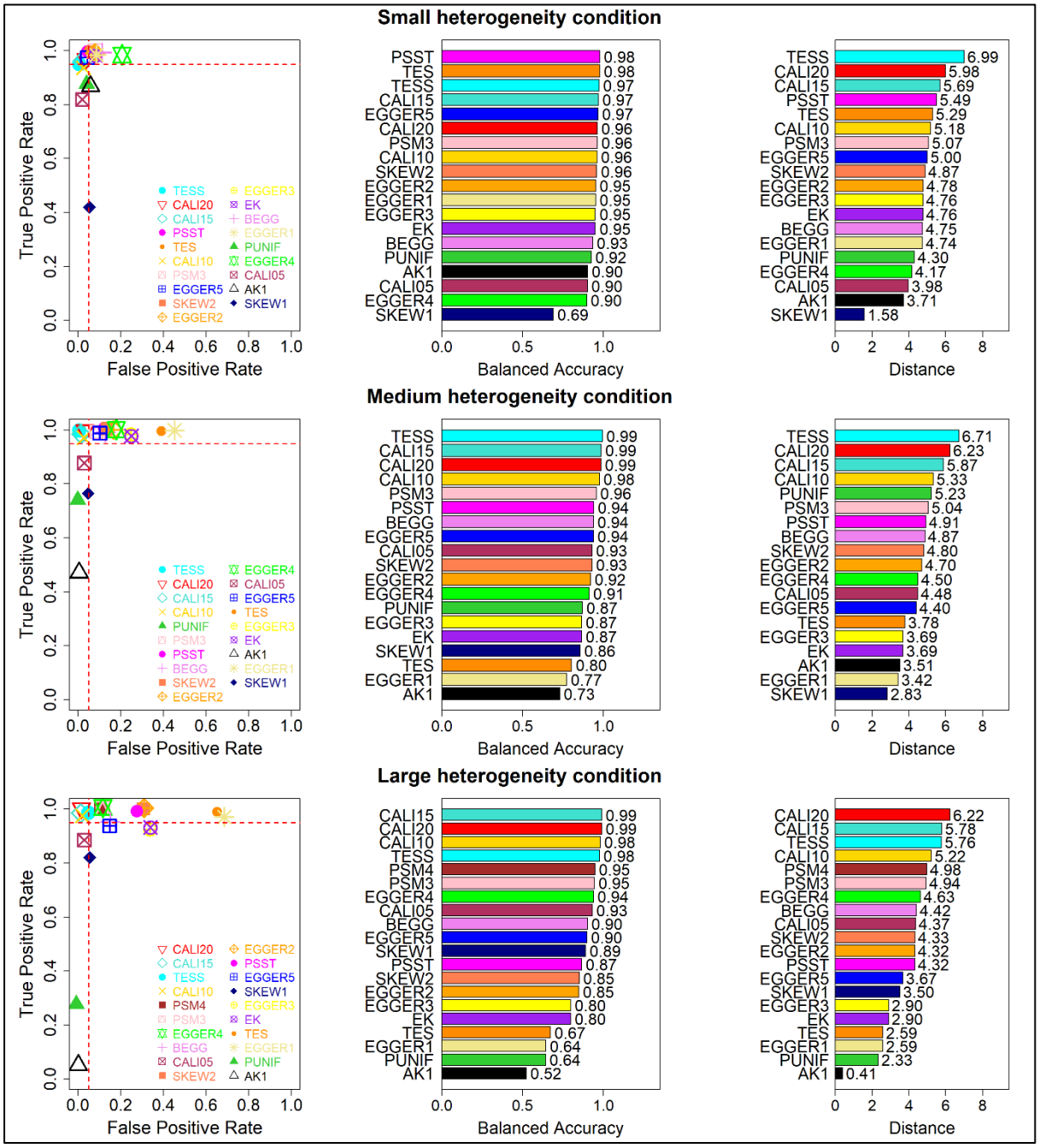
For low heterogeneity (Figure 7a), TESS performed third on Balanced Accuracy, but best on Distance, followed by CALI20 and CALI15. False positive rates were low for all three—.024 for CALI15 and .034 for CALI20—and extremely low for TESS (.001). For medium heterogeneity (LM, Figure 7b), false positive rates increased for most methods relative to the low-heterogeneity case, but not for TESS, CALI20 and CALI15. TESS again performed best, now for both Balanced Accuracy and Distance, with CALI15 and CALI20 also ranking among the top three methods on both metrics.

With high heterogeneity (LL, Figure 7c), false positive rates increased further, particularly for the precision–based methods, but also for TES (.655), PSST (.267), and even PSM3 (.109). In contrast, the four best-performing methods—CALI20, CALI15, TESS, and CALI10 (ranked first through fourth based on Distance)—all maintained false positive rates below .05 (.020, .018, .046, and .024, respectively). Because statistical power was close to 1 in this setting, methods with lower false positive rates also achieved the highest Balanced Accuracy. PSM3, which performed best under high heterogeneity in small and medium-sized meta-analyses, dropped out of the top group in the large-K setting because its false positive rate (.109) was higher than that of the best-performing methods. *p*-uniform and AK1 performed very poorly; they rarely rejected a true null and also had low true positive rates, resulting in poor overall discrimination.

Turning to the Relative Performance curves in Figure 8, many methods performed best in more than half of the conditions because true positive rates were close to 1 for most methods in many scenarios. Consequently, Relative Performance is less informative in the large-K setting, where performance differences are driven almost entirely by false positive rates rather than power. In this setting, Balanced Accuracy and Distance provide more informative measures of performance.

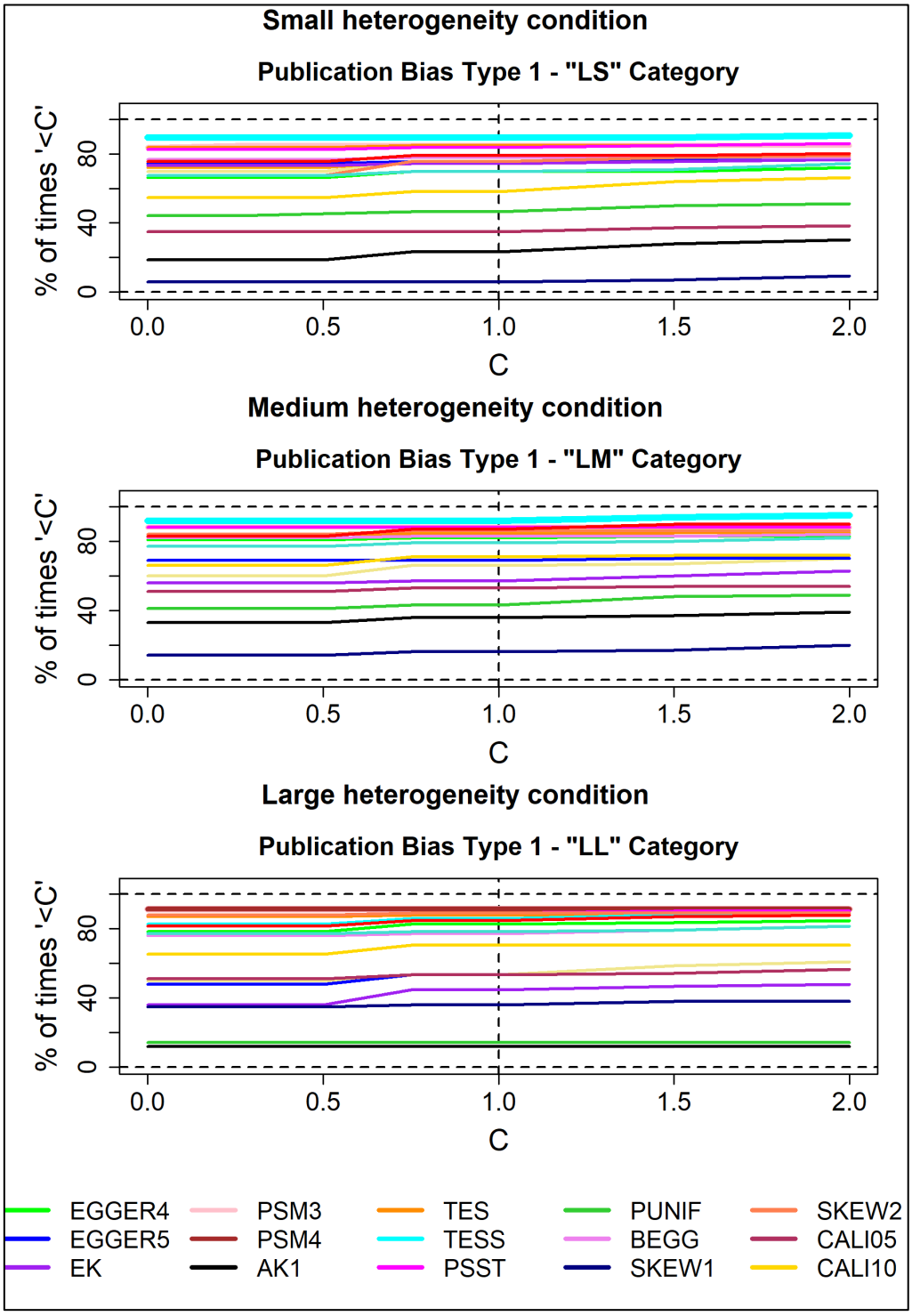
In summary, the results for large meta-analyses reinforce and strengthen a pattern observed in smaller meta-analyses: as *K* increases, performance differences are increasingly determined by how effectively bias-detection methods control false positives. In large samples, caliper procedures—particularly CALI20 and CALI15—emerge as the best-performing methods, while TESS remains broadly robust across heterogeneity levels. Precision–based methods continue to perform comparatively poorly due to persistent inflation of Type I error as heterogeneity increases.

Figure 7: ROC curve and bar graphs of Balanced Accuracy and Distance of bias detection methods for Large meta-analyses, $K > 100$.



NOTE: The three panels report methods' performance for (from top to bottom) Small heterogeneity, Medium heterogeneity, Large heterogeneity, and with respect to (from left to right) ROC space, Balanced Accuracy, and Distance.

Figure 8: Distance curves of bias detection method for Large-sized meta-analyses: small heterogeneity (LS), medium heterogeneity (LM), large heterogeneity (LL)



NOTE: This figure plots Relative Performance curves as a function of the distance parameter, c The Relative Performance statistic, $P_i(c)$, measures the proportion of instances where the respective test is within c logit units of the best-performing test. $P_i(c = 0)$, equals the proportion of times in which method i performed best, $P_i(c = 1)$ equals the proportion of times in which that method was within distance 1 of the best method.

Discussion

Overview of main findings

This study evaluated 21 bias-detection methods across 864 simulation conditions incorporating publication bias and QRPs. By organizing the results into nine subgroups defined by meta-analytic size (*K*) and observed heterogeneity (I^2), we evaluate method performance from the perspective of applied researchers who must select a bias-detection test based on observable study characteristics. Three main conclusions emerge from the analysis.

Overall, TESS performed best for small and medium-sized meta-analyses, most consistently achieving the highest or near-highest discrimination across heterogeneity levels. In large meta-analyses, PSM3 and CALI20 often emerged as top performers—particularly when heterogeneity was high ($I^2 \geq 0.60$)—but even in these settings TESS remained competitive and rarely fell far behind the leading method. Accordingly, while no single method dominated in every condition, TESS provided the most consistently strong performance across the widest range of conditions and therefore represents the most reliable general-purpose test.

Second, heterogeneity was the primary determinant of relative performance. While increasing the number of studies uniformly improved statistical power across methods and reduced convergence problems, these gains affected most bias-detection tests similarly. As a result, increases in the number of studies generally improved performance for all methods without substantially changing their relative ranking. In contrast, shifts in ranking were largely driven by changes in I^2. Methods that did not adequately account for heterogeneity exhibited sharply inflated false positive rates as I^2 increased, which substantially reduced their overall discrimination ability.

Third, precision–based methods—including all variants of Egger's regression—consistently underperformed. Their vulnerability to heterogeneity led to substantially inflated false positive rates in many realistic settings. This finding is particularly important because Egger-type tests remain the most frequently used quantitative bias-detection tools in practice (Wu et al., 2026). Continued reliance on these procedures may therefore lead to systematically biased inferences about the presence of publication bias, with false positives likely when heterogeneity is large— a condition that is common in psychological meta-analyses (van Erp et al., 2017) and in many other fields (Wu et al., 2026).

Practical Guidance for Applied Meta-Analysts

Applied meta-analysts are recommended to select bias-detection methods based on their expected performance under the characteristics of the meta-analysis at hand, rather than relying on convention or software defaults. Simulation studies such as ours can help inform these choices. Our results show that the performance of bias-detection methods varies systematically with key meta-analytic characteristics, particularly heterogeneity and the number of studies. These characteristics can therefore be used to guide the selection of methods that are expected to perform best in a given application. The recommendations below summarize which methods performed best under different meta-analytic conditions and are intended to provide practical guidance for applied researchers whose meta-analyses resemble the conditions examined in our simulation study.

Recommendations. Our results support the following condition-specific recommendations based on observable characteristics of a meta-analysis. When heterogeneity is low to moderate ($I^2 < 0.60$), TESS is recommended as the primary bias-detection method regardless of the number of studies. When heterogeneity is high ($I^2 \geq 0.60$) and the number of studies is 100 or fewer, PSM3 is recommended when feasible, with TESS serving as a practical alternative when selection models are infeasible or fail to converge. When heterogeneity is high ($I^2 \geq 0.60$) and the meta-analysis includes more than 100 studies, caliper tests—particularly CALI20 and CALI15—are recommended, with TESS again serving as a strong alternative

Table 6 summarizes these recommendations and presents a practical decision rule for selecting a bias-detection method based on two observable characteristics of a meta-analysis. This decision rule is a central contribution of the study and reflects a broader finding: no single method performs best across all conditions; instead, the optimal method depends primarily on heterogeneity and, secondarily, on the number of studies. In practice, researchers should therefore examine I^2 and the number of studies before selecting a bias-detection method, rather than relying on default procedures or routinely used tests.

Table 6. Decision rule for selecting a bias-detection method

Observed heterogeneity (I^2)	No. of studies (K)	Recommended method	Alternative
$I^2 < 0.60$	Any K	TESS	—
$I^2 \geq 0.60$	$K \leq 100$	PSM3	TESS
$I^2 \geq 0.60$	$K > 100$	CALI20 or CALI15	TESS

NOTE. Recommendations are based on overall performance across the simulation conditions examined in this study. TESS showed consistently strong performance across most conditions

and can be used when selection models or caliper tests are infeasible, fail to converge, or are not available in standard software.

Across all settings, precision–based methods should not be relied upon for bias detection due to their inflated false positive rates under heterogeneity. Consequently, caution is also warranted when interpreting funnel plot asymmetry, the most commonly used approach (60%; Wu et al., 2026), which qualitatively depicts the association between primary study standard errors and observed effect sizes. Because funnel plots reflect the same small-study effects captured by precision–based tests and because individual study estimates are imprecise, visual inspection of funnel plot asymmetry is an unreliable method for detecting bias, particularly when heterogeneity is present (Tang & Liu, 2000; Terrin et al., 2005; Van Assen et al., 2023)

Using Multiple Bias Detection Methods. Best practice guidelines often recommend applying multiple bias-detection procedures (e.g., Wu et al., 2026). This requires researchers to define a decision rule on how the results of the applied procedures are interpreted prior to conducting these analyses to avoid selective reporting. Preregistering the bias-detection procedures that will be applied as well as the decision rule is advised to counteract selective reporting. Note that when requiring agreement across procedures, this has the important drawback that it reduces statistical power, particularly in small meta-analyses where power is already limited.

If researchers elect to apply multiple tests, we recommend drawing them from the top-performing methods for the relevant (K, I^2) regime and from distinct methodological classes (e.g., excess statistical significance tests, selection models, and caliper tests), rather than employing multiple variants of the same underlying approach. Using multiple variants of the same method (e.g., several versions of Egger's regression) does not provide independent evidence of bias and may instead reinforce method-specific biases.

We do not exclude the possibility that using a combination of several bias-detection procedures outperforms the use of one procedure in particular conditions. For instance, caliper tests seem particularly sensitive to p-values just below .05, whereas selection models more generally model publication bias. Combining them into one decision rule may improve overall performance. However, we did not test the performance of combinations of procedures. Future research is needed into understanding under which conditions the procedures and combinations of them detect bias, and our large dataset may be used for that purpose.

Explaining the best performance of TESS, PSM3, and CALI20

We attribute the strong performance of TESS to three main features. First, TESS focuses directly on the proportion of statistically significant results, which reflects the combined effects of publication bias and QRPs. Second, TESS does not rely on distributional assumptions about the selection process, unlike selection models. Third, TESS has relatively high statistical power because the standard error associated with testing an expected excess significance (ESS) of .05 is small. However, in large meta-analyses with high heterogeneity, TESS exhibited very low Type I error rates, suggesting that its performance in these settings could potentially be improved by modestly increasing its nominal significance level.

When it converged, PSM3 outperformed TESS (and all other methods) for meta-analyses with up to 100 studies and high heterogeneity. This strong performance was somewhat unexpected because the publication bias mechanism in our simulation environment does not match the selection process assumed by the selection model, and the selection model does not explicitly model QRPs. We attribute the good performance of PSM3 to its focus on the relative probability of statistically significant versus nonsignificant results, a signal that is affected by both publication bias and QRPs. In large meta-analyses, however, PSM3 performed worse than TESS due to its higher Type I error rate. As statistical power approached 1 in these conditions, overall performance became increasingly determined by false positive rates, suggesting that performance of PSM3 in large meta-analyses could be improved by using a more stringent significance threshold.

Among the caliper tests, the test with the largest interval (CALI20) showed the best overall performance. In meta-analyses with 100 studies or fewer, caliper tests suffer from two limitations. First, the tests may not be applicable because the intervals around the significance threshold may contain too few effect sizes. Second, statistical power is relatively low because the standard error of the caliper test is large compared to TESS (caliper tests test a null proportion of .50, whereas TESS tests a null proportion of .05). In large meta-analyses with more than 100 studies and high heterogeneity, CALI20 performed best overall, and second best after TESS when heterogeneity was low to moderate. We attribute the good performance of CALI20 in these settings to three features: it does not rely on model assumptions, it tests for an excess of statistically significant results relative to nonsignificant results near the significance threshold, and it maintains an appropriate Type I error rate in high-heterogeneity settings. Our results therefore support the use of caliper tests, particularly CALI20, for detecting bias in large and heterogeneous collections of results, including meta-analyses that

combine results across diverse or loosely related research topics, as in Gerber and Malhotra (2008).

The results show that the performance of bias-detection methods depends primarily on heterogeneity and, secondarily, on the number of studies. Methods that perform well under low heterogeneity do not necessarily perform well when heterogeneity is high, and methods that perform well in small meta-analyses are not always optimal in large meta-analyses. These patterns explain why no single method performs best across all conditions and why a condition-specific approach is necessary

Accompanying Shiny app and R package

For readers interested in exploring alternative data structures — for example subsetting only on publication bias or QRPs, or combining different combinations of number of studies and heterogeneity categories — we provide a Shiny app: https://pbiasmethods.shinyapps.io/main/. We also provide an R package called "pbiasr" that allows users to apply the 21 publication bias tests assessed in this paper to their own datasets. Further details, including instructions on how to install pbiasr can be found here; https://github.com/sho-125/BiasDetectionMethods. For applying pbiasr wisely, we refer to our recommendations based on the results of our simulation study.

Limitations

Although our simulation environment, inspired by Carter et al. (2019), covers a wide range of conditions, the exact magnitude of method performance will depend on the true extent and form of publication bias and QPRs in practice, which likely varies across fields. For example, what our simulation environment defines as "medium" publication bias implies that statistically significant findings are approximately four times more likely to be published than nonsignificant findings, which may be stronger than the selection processes operating in some disciplines. If actual publication bias is weaker than assumed here, the statistical power of bias-detection methods in practice will be lower than what our simulation results suggest. However, the main qualitative findings—particularly the importance of heterogeneity and the relative performance of methods across conditions—are likely to generalize across a wide range of research settings. Accordingly, the decision framework based on heterogeneity and the number of studies should remain broadly applicable.

An important limitation concerns how QRPs are modeled, because different QRPs affect meta-analytic bias in different ways. For instance, optional stopping or data peeking primarily increases the proportion of statistically significant results and concentrates *p*-values near the significance threshold and hardly biases effect sizes (Francis, 2012; Hartgerink et al., 2016; van Aert et al., 2016). This is in contrast with reporting the best result from multiple outcome variables that yields a positive bias in effect size (Maassen et al., 2025; van Aert et al., 2016). As a result, the relative performance of bias-detection methods depends not only on the presence of QRPs, but also on the specific types of QRPs operating in the research environment. Simulation results should therefore be interpreted as conditional on the QRP mechanisms modeled here rather than as universally applicable across all forms of researcher behavior.

It should also be noted that our number of studies and I^2 subgroup boundaries are pragmatic classification devices rather than theoretically derived thresholds. The cutoffs were chosen to yield approximately equal numbers of meta-analyses across subgroups, not because they represent structural breakpoints in statistical behavior. Accordingly, these categories should be interpreted as approximate decision regions rather than strict thresholds. I^2 itself also depends on the sampling variances of the studies included in the meta-analyses (e.g., Rücker et al., 2008), so our categories may not be adequate for meta-analyses with substantially different study sample sizes than we simulated. Moreover, observed I^2 is itself potentially endogenous to bias, as publication bias and QRPs can inflate or otherwise distort heterogeneity (Augusteijn et al., 2019; Jackson, 2007). Thus, the very statistic used to guide method selection may partially reflect the distortions the bias-detection procedures are intended to identify, which introduces an additional layer of uncertainty into condition-based decision rules.

Despite these limitations, the results provide a practical framework for selecting bias-detection methods based on observable characteristics of a meta-analysis, while recognizing that method performance ultimately depends on the underlying bias processes operating in a given research environment.

Conclusion

This study provides a comprehensive comparison of 21 bias-detection methods across 864 simulation conditions incorporating both publication bias and QRPs. By evaluating performance within a realistic data-generating framework and organizing results according to observable characteristics of meta-analyses—specifically, the number of studies (*K*) and

observed heterogeneity (I^2)—we provide condition-specific guidance for applied researchers conducting meta-analyses. More broadly, our results show that the performance of bias-detection methods is systematically related to observable features of meta-analytic datasets, implying that bias detection should be treated as a condition-dependent decision rather than a default procedure.

A primary contribution of this study is the identification of observable meta-analytic conditions under which specific bias-detection methods perform best. Overall, TESS (Stanley et al., 2021) exhibits the most consistent performance across a wide range of realistic meta-analytic settings, particularly in small and medium meta-analyses with low-to-moderate heterogeneity. Under high heterogeneity, alternative procedures—most notably selection models for small and medium meta-analyses and caliper tests for large meta-analyses—often achieve the strongest discrimination, although TESS generally remains competitive in these settings. By contrast, precision–based approaches, including Egger-type regressions, consistently perform poorly due to inflated false positive rates under heterogeneity. Taken together, these findings indicate that no single test is uniformly best; instead, the optimal bias-detection method depends primarily on heterogeneity and, secondarily, on the number of studies, supporting a condition-matched strategy for bias detection rather than reliance on routine default procedures.

Underlying these recommendations is a more general insight: effect size heterogeneity is the primary determinant of relative performance across bias-detection tests. Increases in I^2 systematically reorder performance rankings primarily through their impact on false positive rates. Methods that are insufficiently robust to heterogeneity exhibit inflated Type I error rates, which reduce their ability to distinguish between bias and no-bias conditions as I^2 rises. This result helps explain why commonly used small-study effect tests often perform poorly in practice, where heterogeneity is typically substantial.

The implications of our findings extend beyond individual studies. Wu et al. (2026) report that a large majority of published meta-analyses formally test for publication bias, yet relatively few reject the null hypothesis of no bias. This pattern is consistent with the possibility that widely used bias-detection tests are either underpowered or prone to false positives under realistic conditions, particularly when heterogeneity is present. When bias-detection procedures lack sufficient power or fail to control Type I error under heterogeneity, both over- and under-diagnosis of bias become likely. Consequently, method choice influences not only conclusions within individual meta-analyses but also the collective assessment of bias prevalence across disciplines. More broadly, our results suggest that publication bias detection

should be viewed as a condition-dependent model selection problem rather than a routine diagnostic exercise. Adopting condition-matched bias-detection procedures can therefore improve inference at both the study level and the field level, thereby contributing to more reliable cumulative scientific evidence.

References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLOS ONE*, *12*(3), e0172792. https://doi.org/10.1371/journal.pone.0172792

Agresti, A. (2013). *Categorical data analysis* (Third edition). Wiley-Interscience.

Allum, N., Reid, A., Bidoglia, M., Gaskell, G., Aubert-Bonn, N., Buljan, I., Fuglsang, S., Horbach, S., Kavouras, P., Marušić, A., Mejlgaard, N., Pizzolato, D., Roje, R., Tijdink, J., & Veltri, G. (2023). *Researchers on research integrity: A survey of European and American researchers*. F1000Research. https://doi.org/10.12688/f1000research.128733.1

Andrews, I., & Kasy, M. (2019). Identification of and Correction for Publication Bias. *American Economic Review*, *109*(8), 2766–2794. https://doi.org/10.1257/aer.20180310

Augusteijn, H. E. M., Van Aert, R. C. M., & Van Assen, M. A. L. M. (2019). The effect of publication bias on the Q test and assessment of heterogeneity. *Psychological Methods*, *24*(1), 116–134. https://doi.org/10.1037/met0000197

Bartoš, F., Maier, M., Wagenmakers, E.-J., Nippold, F., Doucouliagos, H., Ioannidis, J. P. A., Otte, W. M., Sladekova, M., Deresssa, T. K., Bruns, S. B., Fanelli, D., & Stanley, T. D. (2024). Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics. *Research Synthesis Methods*, *15*(3), 500–511. https://doi.org/10.1002/jrsm.1703

Begg, C. B., & Mazumdar, M. (1994). Operating Characteristics of a Rank Correlation Test for Publication Bias. *Biometrics*, *50*(4), 1088–1101. https://doi.org/10.2307/2533446

Bom, P. R. D., & Rachinger, H. (2019). A kinked meta-regression model for publication bias correction. *Research Synthesis Methods*, *10*(4), 497–514. https://doi.org/10.1002/jrsm.1352

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis* (1st ed.). Wiley. https://doi.org/10.1002/9780470743386

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. https://doi.org/10.1177/2515245919847196

Deeks, J. J., Macaskill, P., & Irwig, L. (2005). The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology*, *58*(9), 882–893. https://doi.org/10.1016/j.jclinepi.2005.01.016

Dickersin, K. (2005). Publication bias: Recognizing the problem understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (1st ed., pp. 9–33). Wiley. https://doi.org/10.1002/0470870168

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). *Bias in meta-analysis detected by a simple, graphical test*. https://doi.org/10.1136/bmj.315.7109.629

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991. https://doi.org/10.3758/s13423-012-0322-y

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Gerber, A. S., & Malhotra, N. (2008). Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results? *Sociological Methods & Research*, *37*(1), 3–30. https://doi.org/10.1177/0049124108318973

Gopalakrishna, G., Wicherts, J. M., Vink, G., Stoop, I., Van Den Akker, O. R., Ter Riet, G., & Bouter, L. M. (2022). Prevalence of responsible research practices among academics in The Netherlands. *F1000Research*, *11*, 471. https://doi.org/10.12688/f1000research.110664.2

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hartgerink, C., van Aert, R., Nuijten, M., Wicherts, J., & van Assen, M. (2016). *Distributions of p-values smaller than .05 in psychology: What is going on? [PeerJ]*. (4:e1935). https://doi.org/https://doi.org/10.7717/peerj.1935

Higgins, J. P. T., & López-López, J. A. (2025). Reflections on the I-squared index for measuring inconsistency in meta-analysis. *Research Synthesis Methods*, 1–14. https://doi.org/10.1017/rsm.2025.10062

Hong, S., & Reed, W. R. (2021). Using Monte Carlo experiments to select meta-analytic estimators. *Research Synthesis Methods*, *12*(2), 192–215. https://doi.org/10.1002/jrsm.1467

Hong, S., & Reed, W. R. (2026). Is UWLS really better for medical research? *Statistics in Medicine*.

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, *19*(5), 640. https://doi.org/10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245–253. https://doi.org/10.1177/1740774507079441

Irsova, Z., Bom, P. R. D., Havranek, T., & Rachinger, H. (2025). Spurious precision in meta-analysis of observational research. *Nature Communications*, *16*(1), 8454. https://doi.org/10.1038/s41467-025-63261-0

Iyengar, S., & Greenhouse, J. B. (1988). Selection Models and the File Drawer Problem. *Statistical Science*, *3*(1), 109–117. https://www.jstor.org/stable/2245925

Jackson, D. (2007). Assessing the Implications of Publication Bias for Two Popular Estimates of between-Study Variance in Meta-Analysis. *Biometrics*, *63*(1), 187–193. https://doi.org/10.1111/j.1541-0420.2006.00663.x

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. https://doi.org/10.1177/0956797611430953

Kromrey, J. D., & Rendina-Gobioff, G. (2006). On Knowing What We Do Not Know: An Empirical Comparison of Methods to Detect Publication Bias in Meta-Analysis. *Educational and Psychological Measurement*, *66*(3), 357–373. https://doi.org/10.1177/0013164405278585

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*(2), 107–112. https://doi.org/10.1111/j.2044-8317.1978.tb00578.x

Lin, L., & Chu, H. (2018). Quantifying publication bias in meta-analysis. *Biometrics*, *74*(3), 785–794. https://doi.org/10.1111/biom.12817

Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., & Wicherts, J. M. (2025). *The Impact of Publication Bias and Single and Combined p-Hacking Practices on Effect Size and Heterogeneity Estimates in Meta-Analysis* (2uynm_v1). PsyArXiv. https://doi.org/10.31234/osf.io/2uynm_v1

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, *20*(4), 641–654. https://doi.org/10.1002/sim.698

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, *10*(1), 57–71. https://doi.org/10.1002/jrsm.1332

Rabelo, A. L. A., Farias, J. E. M., Sarmet, M. M., Joaquim, T. C. R., Hoersting, R. C., Victorino, L., Modesto, J. G. N., & Pilati, R. (2020). Questionable research practices among Brazilian psychological researchers: Results from a replication study and an international comparison. *International Journal of Psychology*, *55*(4), 674–683. https://doi.org/10.1002/ijop.12632

Renkewitz, F., & Keiner, M. (2019). How to Detect Publication Bias in Psychological Research: A Comparative Evaluation of Six Statistical Methods. *Zeitschrift Für Psychologie*, *227*(4), 261–279. https://doi.org/10.1027/2151-2604/a000386

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (1st ed., pp. 1–7). Wiley. https://doi.org/10.1002/0470870168

Rücker, G., Carpenter, J. R., & Schwarzer, G. (2011). Detecting and adjusting for small-study effects in meta-analysis. *Biometrical Journal*, *53*(2), 351–368. https://doi.org/10.1002/bimj.201000151

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, *8*(1), 79. https://doi.org/10.1186/1471-2288-8-79

Schneck, A. (2017). Examining publication bias—A simulation-based evaluation of statistical tests on publication bias. *PeerJ*, *5*, e4115. https://doi.org/10.7717/peerj.4115

Schwarzer, G., Antes, G., & Schumacher, M. (2002). Inflation of type I error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. *Statistics in Medicine*, *21*(17), 2465–2477. https://doi.org/10.1002/sim.1224

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Stanley, T. D., Doucouliagos, H., Ioannidis, J. P. A., & Carter, E. C. (2021). Detecting publication selection bias through excess statistical significance. *Research Synthesis Methods*, *12*(6), 776–795. https://doi.org/10.1002/jrsm.1512

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*(11), 1119–1129. https://doi.org/10.1016/S0895-4356(00)00242-0

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rucker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T.

(2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*(jul22 1), d4002–d4002. https://doi.org/10.1136/bmj.d4002

Tang, J.-L., & Liu, J. L. (2000). Misleading funnel plot for detection of bias in meta-analysis. *Journal of Clinical Epidemiology*, *53*(5), 477–484. https://doi.org/10.1016/S0895-4356(99)00204-8

Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology*, *58*(9), 894–901. https://doi.org/10.1016/j.jclinepi.2005.01.006

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting Meta-Analyses Based on p Values: Reservations and Recommendations for Applying p-Uniform and p-Curve. *Perspectives on Psychological Science*, *11*(5), 713–729. https://doi.org/10.1177/1745691616650874

Van Assen, M. A. L. M., Van Den Akker, O. R., Augusteijn, H. E. M., Bakker, M., Nuijten, M. B., Olsson-Collentine, A., Stoevenbelt, A. H., Wicherts, J. M., & Van Aert, R. C. M. (2023). The Meta-Plot: A Graphical Tool for Interpreting the Results of a Meta-Analysis. *Zeitschrift Für Psychologie*, *231*(1), 65–78. https://doi.org/10.1027/2151-2604/a000513

van Assen, M., van Aert, R., & Wicherts, J. (2015). Meta-Analysis Using Effect Size Distributions of Only Statistically Significant Studies. *Psychological Methods*, *20*, 293–309. https://doi.org/10.1037/met0000025

van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of Between-Study Heterogeneity for 705 Meta-Analyses Reported in Psychological Bulletin From 1990–2013. *Journal of Open Psychology Data*, *5*(1). https://doi.org/10.5334/jopd.33

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

Vevea, J. L., & Woods, C. M. (2005). Publication Bias in Research Synthesis: Sensitivity Analysis Using A Priori Weight Functions. *Psychological Methods*, *10*(4), 428–443. https://doi.org/10.1037/1082-989X.10.4.428

Wu, W., Duan, J., Reed, W. R., & Tipton, E. (2026). What can we learn from 1,000 meta-analyses across 10 different disciplines? *Research Synthesis Methods*, *17*(1), 123–156. https://doi.org/10.1017/rsm.2025.10035

Wu, W., & Reed, W. R. (2026). Meta-analyses in management and marketing: Current practices and inferential consequences. *Australian Journal of Management*, *51*(2), 327–353. https://doi.org/10.1177/03128962261443442