

Abschlussarbeit
zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.) Psychologie

**Assessing Publication Bias in Meta-Analyses:
A Simulation-Based Estimation Approach Focusing on the Joint
Distribution of Effect Size and Sample Size**

vorgelegt von

Jan Luca Schnatz

Matrikelnummer: 7516898

E-Mail: janluca.schnatz@gmx.de

Goethe-Universität Frankfurt am Main
Faculty 05: Psychology and Sports Sciences
Department of Psychological Methods with Interdisciplinary Focus

Erstgutachter: Prof. Dr. Martin Schultze
Zweitgutachter: Julia Beitner M.Sc.

Abstract

Test

Table of contents

Abstract	ii
Introduction	2
Reasoning of the n-ES correlation	3
Methodological Concerns	3
The Present Study	6
Method	8
The SPEEC Approach	8
Overview	8
Simulation Framework	9
Application of Publication Bias	10
Formulation as an Optimization Problem	11
Algorithmic Parameter Optimization with Differential Evolution	13
Secondary Data Description	13
Statistical Analysis	14
Smallest Effect Size of Interest	15
Results	17
Variable Dispersion of ω_{PBS}	17
Confirmatory Results of the Predictions from the Hypotheses	17
Diagnostic Evaluation of Parameter Estimation in SPEEC	19
Discussion	23
References	24
Appendix	31
Appendix A: Power Analyses determining the SESOIs	31
Appendix B: Research Transparency Statement	32
Appendix C: Deviations of Preregistration	33
Appendix D: Test	33
Appendix E: Regression Tables of Confirmatory Analyses	33

Introduction

Science is commonly conceived as a cumulative enterprise (Cooper et al., 2019) with the overarching goal of attaining robust knowledge about the world (Kitcher, 1993). Within this landscape, researcher often study the same phenomena, driven by the idea that generalizing and synthesizing findings from individual studies contributes to advancement of knowledge. However, this premise hinges on the underlying assumption, that the available scientific literature is representative for all conducted research (Song et al., 2010).

Contrary to this, researchers have pointed out for over half a century that results of published studies differ systematically from unpublished studies (Bakan, 1966; Bozarth & Roberts, 1972; Smart, 1964; Sterling, 1959). This discrepancy arises as the publication of a study often hinges on the direction or strength of its findings (Dickersin, 1990; Dickersin & Min, 1993) and is collectively known as *publication bias*. Especially in a publishing culture that prioritizes novelty and positive results (Nosek et al., 2012), many statistically nonsignificant studies end up in the “file-drawer” and never get published (Rosenthal, 1979).

The ramifications of publication bias are severe, culminating in inflated meta-analytical effect sizes (Franco et al., 2014; Stanley et al., 2021), heightened false-positive rate (Kicinski, 2014; Munafò & Flint, 2010), thereby increasing the risk of erroneous conclusions that may jeopardize the validity of research (Begg, 1994). These ramifications become especially relevant in the light of recent large-scale replication projects providing evidence for non-replicability of many psychological findings (Camerer et al., 2018; Ebersole et al., 2016, 2020; Klein et al., 2014, 2018; Open Science Collaboration, 2015). This underscores why publication bias identified as a major threat to replicable science (Munafò et al., 2017) and thus a considered as a significant contributor to the replication crisis (Renkewitz & Keiner, 2019). Given the myriad of issues associated with publication bias and its widespread impact, there has been considerable attention directed towards investigating methodologies to detect publication bias.

In this regard, there has been a great deal of research on publication bias detection techniques with numerous statistical methods developed over the past 50 years (Marks-Anglin & Chen, 2020). These statistical techniques can generally be classified into methodologies that operate with *p*-values and methodologies that are based on the relationship between effect size and sample size (Vevea et al., 2019). While both categories encompass highly sophisticated statistical techniques (CITATION?), a straightforward and frequently described method, that has been associated with publication bias, involves examining the correlation between effect size and sample size. Additionally, this method encapsulates the central ideas

of other approaches, such as Begg's rank correlation (Begg & Mazumdar, 1994), Egger's regression (Egger et al., 1997), and its proposed variants (for an overview see Song et al., 2010), all rooted in the relationship between effect size and sample size.

Reasoning of the n-ES correlation

The central tenets of the correlation of effect size and sample size as an indicator of publication bias originate from the concepts of the funnel plot and its asymmetry under the influence of publication bias that was introduced by Light & Pillemer (1984). When multiple studies investigate of common underlying effect, the empirical effect sizes (for example Cohen's d or Fisher-z transformed r) follow a normal distribution and fluctuate around the true effect size. Due to sampling error, the lower the sample sizes of individual studies, the less precision they exhibit to estimate the true effect size (i.e., larger standard error), leading to a larger variation around the true effect size. In the absence of publication bias this will result in a symmetric funnel shaped distribution (Light & Pillemer, 1984). However, when the publishing of studies is contingent on their statistical significance, the funnel plot will be asymmetric. As the statistical significance of p -values is jointly determined by the sample size (i.e., standard error of the test statistic) and effect size (i.e., test statistic), larger effect sizes attain statistical significance with smaller sample sizes, while smaller effect sizes necessitate larger sample sizes to be significant. Consequently, the negative correlation between effect size and sample size emerges because the threshold for the smallest effect sizes that is statistically significant decreases with increasing sample size (A. Linden et al., 2024). The correlation between effect size and sample size has been described and attributed to publication bias extensively in various research including psychology (Fritz et al., 2013; Kühberger et al., 2014; Levine et al., 2009), evolutionary biology and ecology (Jennions & Møller, 2002a; Jennions & Møller, 2002b; Møller & Jennions, 2001; Palmer, 1999), political science (Gerber et al., 2001) and educational research (R. E. Slavin et al., 2008; R. Slavin & Smith, 2009). Its prevalence across these disciplines highlights its role as a widely recognized and applied tool for the detection of publication bias.

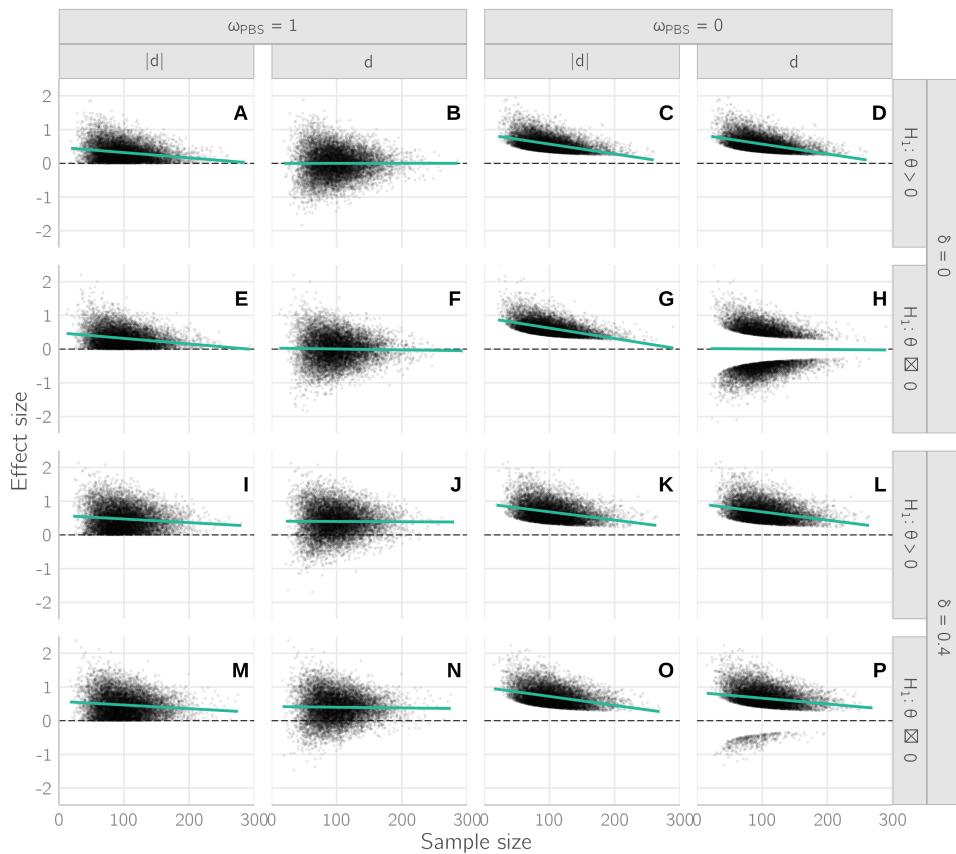
Methodological Concerns

Despite the significant attention and prevalent use of the effect size-sample size correlation in various research fields for detecting publication bias, coupled with its frequent acknowledgment as a valid indicator of such bias, there exist persisting methodological concerns. As I will argue in the next section, these concerns have only been partially discussed and addressed in the existing literature and may compromise the validity of the interpretation

of the correlation as an indicator of publication bias. To illustrate the inherent challenges of the effect size-sample size correlation as an indicator of publication bias, we simulated a set 10000 primary studies¹ on the same effect underlying effect and varied different parameters that contribute to its limitations (see figure 1). This includes the true effect size $\delta : \{0, 0.4\}$, the extent of publication bias $\omega_{PBS} : \{0, 1\}$ or how much less likely studies with non-significant p -values are compared to studies with significant results (in this extreme case either non-significant studies are not published at all, or there are no differences between non-significant and significant studies), the signedness of the effect size (d and $|d|$) and the type of hypothesis (directional $\mathcal{H}_1 : \theta > 0$ and non-directional $\mathcal{H}_1 : \theta \neq 0$)

Figure 1

n-ES Correlation in Simulated Data: Influence of Publication Bias, True Effect Size, and Signed vs. Unsigned Effects



Note:

Firstly, it is common practice to use unsigned effect sizes to estimate the n-es correlation (Kühberger et al., 2014; Levine et al., 2009; e.g., R. Slavin & Smith, 2009; Weinerová et al., 2022). Whilst this is very common, it has only recently been acknowledged that the use

¹The selected number of primary studies aims to visually highlight its inherent limitations. The claims would still hold true even with a significantly lower number of primary studies (more representative of primary studies of meta-analysis e.g., 100), albeit with increased variability.

of unsigned effect sizes can lead to a statistical artefact resulting in a small negative correlation, even in the absence of publication bias (A. Linden et al., 2024). As depicted in figure 1 (leftmost column compared to second leftmost column), the artificial correlation in absence of publication bias is most severe when the true effect is close to zero, as this condition leads to the most sign changes. Especially when considering that effect sizes in psychology are typically smaller than common benchmarks (Lovakov & Agadullina, 2021; Weinerová et al., 2022), and thus it is likely that the true effect sizes of psychological phenomena are often small, this exacerbates the problem of the statistical artifact.

If a negative correlation can emerge even in the absence of publication bias, this raises questions about the appropriate null hypotheses to test against, specifically, what correlation we would expect if publication bias is absent (A. Linden et al., 2024). There has been a long tradition in null hypothesis testing to use the nil null hypothesis (Cohen, 1994), which states that a population parameter is exactly zero. This is also very common in studies that have used the n-es correlation together with unsigned effect sizes (Kühberger et al., 2014; Levine et al., 2009; R. Slavin & Smith, 2009; Weinerová et al., 2022) and underscores a lack of thorough consideration for the potential falseness of this hypothesis in such cases. The determination of an appropriate null hypothesis for testing in these scenarios, however, remains uncertain.

Utilizing *signed* effect sizes may seem like a straightforward solution to the aforementioned problems, however, it introduces its own set of challenges. Especially, when researchers make non-directional hypotheses and where the true effect size is close to zero, the distribution of the signed effect sizes and sample size will be symmetrically hollowed out under the influence of publication bias. This symmetry (see Figure 1 H) will result in the correlation being zero, leading to a false negative - a failure to detect publication bias when it is present.

Apart from these more statistical challenges, there is also a more conceptual challenge.

-> n-es correlation somewhat misses the point of publication bias

- Fails to capture the point of publication bias -> depending on statistical significance -> effect size and sample size correlation only indirectly captures the censorship process of non-significant studies
- as figure shows, the non-linear relationship → critical test statistic value under which $p < \alpha$ nonlinear ->
- Harrer et al. (2021)
- **Questionable linearity assumption**
 - Pearson correlation assumes that under publication bias → linear relationship between effect size and sample size → the higher the effect size the lower the required sample size for the effect to be significant and vice versa may given the

- (false) impression that this assumption holds
- But publication bias operates under statistical significance (which is most dominantly → if p-value smaller than alpha threshold; CITATION) → as figure shows, the non-linear relationship → critical test statistic value under which $p < \alpha$ nonlinear
 - Spearman correlation loosens the assumption of a linear effect in that the relationship has to be only strictly monotonic → but still: this is not how publication bias operates

The Present Study

Research questions: - How does publication bias influence the joint probability distribution of effect size and sample size? - How can the magnitude of publication bias in meta-analyses be estimated and effect sizes under publication bias be corrected from the joint distribution of sample size and effect size

- Introduce novel simulation-based method to estimate publication bias in meta-analyses and correct biased effect sizes under publication bias
- Proof of concept study -> introduce method and assess the method using
- Introduce and explain new framework to estimate publication bias
- Assess the approach using secondary empirical meta-analytical data -> set of predictions/hypotheses that should hold true if the approach works in principle,

Hypothesis III:

Registered reports are an alternative two-stage publishing model, where study protocols are submitted, peer-reviewed and in-principle accepted prior to data collection (Chambers & Tzavella, 2022; Nosek & Lakens, 2014). In-principle accepted studies are then published regardless of the outcome of the study, because the decision to publish was made before the results of the studies are known, which completely eliminates publication bias (Chambers & Tzavella, 2022; Simons et al., 2014). Thus, the effect sizes within these multisite replication projects and registered replication reports that are published within this framework cannot be biased by publication bias. Thus, the third hypothesis \mathcal{H}_3 , the average effect size of a registered replication report $\hat{\delta}$ can be expected to be equivalent to the mean of the Gaussian effect size distribution μ_d that is estimated within the SPEEC approach within specified equivalence bounds.

$$\Delta_{\mu_d} = \hat{\delta} - \hat{\mu}_d$$

$$\mathcal{H}_{01} : \Delta_{\mu_d} \leq \Delta_L \quad \cap \quad \mathcal{H}_{02} : \Delta_{\mu_d} \geq \Delta_U$$

$$\mathcal{H}_1 : \Delta_L > \Delta_{\mu_d} > \Delta_U$$

{eq-h3}

** Hypothesis IV**:

Moreover, when comparing meta-analysis which were published within a traditional publishing system compared to multisite replication studies and registered replication reports, one can also make predictions about the publication bias parameter ω_{PBS}

$$\omega_{PBS}$$

In statistical terms, when the regressor is a binary indicator of the type of research synthesis with the reference level being the publication biased absent multisite replication studies (MR) and the outcome is the estimated publication bias parameter ω_{PBS} , the regression coefficient β_{MR} can be expected to be greater than zero.

$$\mathcal{H}_0 : \beta_{MR} \leq 0$$

$$\mathcal{H}_1 : \beta_{MR} > 0$$

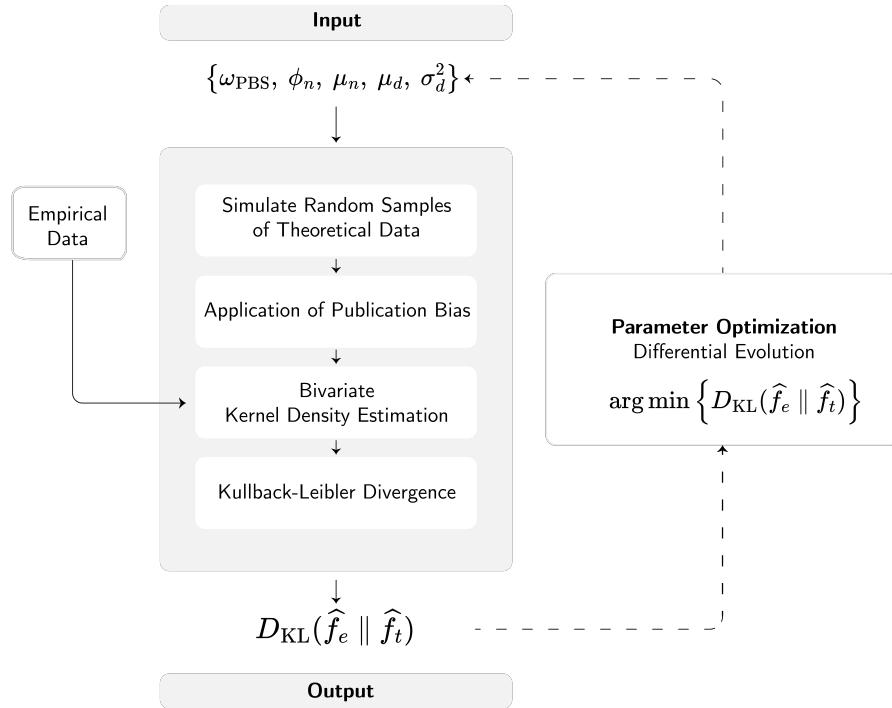
{eq-h4}

Method

The SPEEC Approach

Overview

- Simulation-based approach to estimate publication bias severity and correct potentially biased (inflated) effect sizes under present publication bias based on the joint distribution of effect size and sample size
 - Simulation of theoretical data > joint distribution of effect size and sample size under marginal distributional assumptions > Application of publication bias > empirical kernel density estimation > comparison of empirical and simulated data > loss function
 - Implementation as an open source R package (alpha version) that is already available on GitHub: <https://github.com/jlschnatz/speec>
 - General steps
 - Simulate samples of joint distribution of effect size and sample size from marginal distributional assumptions
 - Application of publication bias
 - KDE for Simulated and Empirical Data
 - Compare Distributions using KL Divergence
 - Parameter Optimization via Simulated Annealing
- (1) Simulation of random samples from joint probability distribution of effect size and sample size
 - (2) Application of publication bias
 - (3) Kernel Density Estimation of drawn theoretical samples and the empirical samples
 - (4) Computation of the divergence between the estimated probability of empirical against theoretical data
 - (5) Algorithmic optimization of bias and distributional parameters via simulated annealing
Simulation of k random samples from the joint distribution of effect size and sample size, subsequent application of publication bias resulting in a subset (d'_i, n'_i) of the initial random samples.

Figure 2*Overview of the SPEEC Approach**Note.* Test.

Simulation Framework

The marginal distribution for the total sample size n should be inherently modeled as a discrete distribution. Count data of this nature are commonly modeled using either a Poisson or Negative-Binomial distribution. In various psychological domains, sample size distributions often exhibit considerable variance and skewness (see for example Cafri et al., 2010; Marszalek et al., 2011; Sassenberg & Ditrich, 2019; Shen et al., 2011; Szucs & Ioannidis, 2017). Considering this variability and skewness we opted for the Negative-Binomial distribution which can account for variance independently of the mean and thus handle overdispersed data effectively. We use the mean-dispersion parametrization where the probability of success p and the target number of successes r are reparametrized to mean $\mu = r(1 - p)/p$ and dispersion $\phi = r$ to model the study-specific total sample sizes n_i .

$$n_1, n_2, \dots, n_k \quad \text{where} \quad N \stackrel{\text{i.i.d.}}{\sim} \mathcal{NB}(\phi_n, \mu_n) \quad \text{for } i = 1, \dots, k \quad (1)$$

Concerning the marginal distribution of the effect size d , we assume a normal distribution with mean μ_d and variance σ_d^2 , where the effect size itself is assumed to originate from a common two-sample independent t -test design. To address the increasing precision in esti-

mating the true effect size mean μ_d as sample size increases, contributing to the characteristic funnel shape of the effect size-sample size distribution, we compute the variance of the mean differences $\bar{x}_{i1} - \bar{x}_{i2}$, from which the effect sizes originate in this type of design. Subsequently, we derive a normalization factor γ_i by dividing each individual variance $\sigma_{\bar{x}_{i1} - \bar{x}_{i2}}^2$ with overall mean of those variances ensuring that $\bar{\gamma} = 1$.

$$\begin{aligned}\sigma_{\bar{x}_{i1} - \bar{x}_{i2}}^2 &= \sigma_d^2/n_i \\ \gamma_i &= \frac{\sigma_{\bar{x}_{i1} - \bar{x}_{i2}}^2}{\sum_{i=1}^k \sigma_{\bar{x}_{i1} - \bar{x}_{i2}}^2/k}\end{aligned}\tag{2}$$

With this normalization factor, the total variance of the individual variances of the individual variances is $\text{Var}(\gamma \cdot \sigma_d^2) = \sigma_d^2$. The study-specific effect sizes d_i are subsequently modeled as

$$d_1, d_2, \dots, d_k \quad \text{where} \quad D \sim \mathcal{N}(\mu_d, \gamma_i \cdot \sigma_d^2) \quad \text{for} \quad i = 1, \dots, k\tag{3}$$

Application of Publication Bias

Following the simulation step of sampling k individual studies from the joint distribution of effect size and sample size conditional on the parameters, the subsequent step entails applying publication bias to these samples. As mentioned in the introduction, we operationalize publication bias in terms of the likelihood of a study being published conditional on the statistical significance of its results. Translated to this simulation setting we can calculate the two-tailed p -value of each individual study i from the random samples of effect size d_i and sample size n_i . We presume that individual studies i originate from a balanced sample size design, where the group sample sizes n_{1i} and n_{2i} are defined as $n_i/2$ when the total sample size is even. If the total sample size is odd, the group sample sizes are determined as the ceilinged $\lceil n_i/2 \rceil$ and floored $\lfloor n_i/2 \rfloor$ values, respectively. To calculate the p -value p_i of each simulated study, derived from t -value t_i

$$t_i = \left| \frac{d_i}{\sqrt{1/n_{1i} + 1/n_{2i}}} \right|\tag{4}$$

$$p_i = 2 \cdot P(t_i \mid df_i)\tag{5}$$

where $P(t_i, df_i)$ is the cumulative central t -distribution with degrees of freedom $df_i = n_{1i} + n_{2i} - 2$. Given each p -value p_i , publication bias is introduced by assigning each study i a weight

$$\omega_{\text{PBS}_i}(p_i) = \begin{cases} \omega_{\text{PBS}} & \text{for } p_i \geq \alpha \\ 1 & \text{otherwise} \end{cases} \quad (6)$$

given $\omega_{\text{PBS}} \in \mathbb{R} : 0 \leq \omega_{\text{PBS}} \leq 1$. This weight denotes the probability of a study i being selected conditional on the p -value and the type I error rate α . If a study i was not statistically significant ($p_i \geq \alpha$), a publication bias weight ω_{PBS} is assigned, else the probability of a study being selected is 1, indicating no publication bias. Thus, the publication bias parameter ω_{PBS} denotes the relative probability of a nonsignificant study being selected relative to a significant study being selected. We assume a fixed type I error rate for all simulated studies at the common threshold of $\alpha = .05$.

Following the computation of publication bias weight ω_{PBS_i} for each study i , the likelihood of a study being selected can be expressed as $\mathbb{P}(S_i = 1) = \omega_{\text{PBS}_i}$. Here, S_i serves as a binary indicator function, signifying whether study i is selected during the publication bias process.

$$S_i = \begin{cases} 0 & \text{study not selected} \\ 1 & \text{study selected} \end{cases} \quad (7)$$

Subsequently the two resulting subsets (d'_i, n'_i) and (d''_i, n''_i) from the initial random sample can be defined as

$$\begin{aligned} (d'_i, n'_i) &= (d_i, n_i \mid S_i = 1) \quad \text{for } i = 1, \dots, k' \quad \text{and} \\ (d''_i, n''_i) &= (d_i, n_i \mid S_i = 0) \quad \text{for } i = 1, \dots, k'' \end{aligned} \quad (8)$$

One challenge lies in the fact that, *ceteris paribus*, as publication bias becomes more severe (i.e., the lower values ω_{PBS}), the lower k' will be in comparison to the original number of simulations k leading to a precision loss for the estimation of the parameters. To mitigate this, the process outlined above (Equation 1 - Equation 8) repeated a second time with an adjusted number of simulations $k'_{\text{adj}} = \lceil k^2/k' \rceil$, ensuring that the number of selected studies k'_{adj} is roughly equal across the entire range of ω_{PBS} .

Formulation as an Optimization Problem

After the selection of the subset (d'_i, n'_i) from the simulated theoretical random samples conditional on the publication bias parameter ω_{PBS} and the parameters of the marginal distributions for the effect size and sample size, the following step involves a statistical comparison between this subset and the empirical meta-analytical data by means of the of the

Kullback-Leibler divergence (Kullback & Leibler, 1951). This assessment aims to quantify how closely the distributions of the theoretical samples align with the empirical data.

In the subsequent phase, the statistical dissimilarity between the empirical meta-analytical data and the selected subset (d'_i, n'_i) of the simulated theoretical samples is evaluated by means of the Kullback-Leibler divergence (Kullback & Leibler, 1951). This assessment aims to quantify how closely the distributions of the theoretical samples align with the empirical data. To achieve this, the joint kernel density is estimated for both theoretical simulated and empirical data is estimated using a bivariate standard Gaussian kernel that is evaluated on a square grid, with a grid size of $n_{\text{grid}} = 2^7 + 1$ equidistant grid points in each dimension. The bounds of the square grid are determined based on the empirical meta-analytical data. To define these bounds the maximum likelihood estimates for the parameters of the marginal distribution of effect size ($\hat{\mu}_d, \hat{\sigma}_d^2$) and sample size ($\hat{\phi}_n, \hat{\mu}_n$) are computed. Then, utilizing these estimates, the quantiles derived from inner 99th percentile ($p_1 = .005, p_2 = .995$) of the cumulative distribution are computed. Subsequently, the bounds are defined as the absolute minimum and maximum of these quantiles and then range of the empirical meta-analytical data, respectively.

$$[Q_n(p_1 | \hat{\phi}_n, \hat{\mu}_n), Q_n(p_2 | \hat{\phi}_n, \hat{\mu}_n)]$$

$$[Q_d(p_1 | \hat{\mu}_d, \hat{\sigma}_d^2), Q_d(p_2 | \hat{\mu}_d, \hat{\sigma}_d^2)]$$

Finally, KL-divergence is calculated from the kernel density estimates for both empirical and simulated data, providing a measure of their statistical distance.

$$D_{\text{KL}}(\widehat{f}_e \| \widehat{f}_t) = \sum_{u=1}^g \sum_{v=1}^g \widehat{f}_e(u, v) \ln \left(\frac{\widehat{f}_e(u, v)}{\widehat{f}_t(u, v)} \right) \quad (9)$$

$$\min_{\mu_d, \sigma_d^2, \mu_n, \phi_n, \omega_{PBS}} \left\{ D_{\text{KL}}(\widehat{f}_e \| \widehat{f}_t) \right\}, \text{ subject to: } \mu_d, \sigma_d^2, \mu_n, \phi_n, \omega_{PBS} \in \mathbb{R}$$

$$-4 \leq \mu_d \leq 4, \quad 0 \leq \sigma_d^2 \leq 6, \quad 10 \leq \mu_n \leq 15000, \quad 0.01 \leq \phi_n \leq 1000, \quad 0 \leq \omega_{PBS} \leq 1$$

constitute an optimization problem, where the aim is to find values for the four distributional parameters and the publication bias parameter such that the kullback-leibler loss function ... the divergence between the estimated joint kernel density of simulated theoretical data from the estimated joint kernel density of the empirical data

Algorithmic Parameter Optimization with Differential Evolution

Objective of finding parameter values for which Kullback-Leibler divergence between the estimated joint kernel density of the simulated theoretical data from the empirical data is minimized, use differential evolution DE (see Storn & Price, 1997), which is a simple metaheuristic algorithm for global optimization (Feoktistov, 2006). DE is an evolutionary algorithm based on principles such as mutation, cross over and selection, and requires in comparison to other optimization algorithms only few control parameters that are generally straightforward to select to achieve favorable outcomes (Storn & Price, 1997). Importantly, all parameters of the DE algorithm including the control parameters, the stopping criterion and boundary constraints of the differential evolution algorithm were defined globally for the parameter estimation of all meta-analyses.

To utilize DE for optimization within the SPEEC approach, the R package *RcppDE* (Edelbuettel, 2022) was employed, implementing the classical algorithm *DE/rand/1* (Storn & Price, 1997). The control parameters for DE were chosen based on the recommendations of Storn & Price (1997) with additional adjustments informed by preliminary testing of simulated data from the simulation framework of SPEEC, setting the population size NP to 150, the mutation constant F to 0.9 and the crossover constant CR to 0.1. In the application of the DE algorithm, we adopted a direct termination criteria approach (Ghoreishi et al., 2017; Jain et al., 2001), with the termination condition being the maximum number of generations. Since there are no universally applicable default values for the maximum number of generations, as it is contingent upon the optimization problem at hand (Jain et al., 2001), the choice for t_{\max} was also informed by preliminary testing of simulated data from the simulation framework of SPEEC. These tests suggested that $t_{\max} = 1000$ is a reasonable decision. In determining the boundaries for the parameter search space, a balance was made between avoiding boundaries that are too wide, which could lead to inefficient exploration of the search space, and ensuring that the boundaries are not too narrow to ensure sufficient coverage of the potential parameters. More specifically, the minima and maxima for all distributional parameters were determined using Maximum Likelihood (see Table XXX), and the boundaries were set slightly above those values to ensure good coverage.

Secondary Data Description

The data used for the assessment of the predictions of the four hypotheses stem from extensive meta-analytical dataset from previous work by A. H. Linden & Hönekopp (2021).

The dataset contains a total of 207 research synthesis on 207 psychological phenomena.

Subdivided into 150 “traditional” meta-analysis which contain 3 times 50 meta-analyses from different subfields of psychology (social psychology, organizational psychology and cognitive psychology).

Additionaly 57 large-scale replication studies and registered reports, which are important for the investigation of \mathcal{H}_3 and \mathcal{H}_4 .

- To investigate predictions of the hypotheses we use secondary data by Linden & Hönekopp (2021)
- Extensive meta-analytical database, containing both traditional meta-analysis as well as multisite replication studies and registered reports (which are absent of publication bias)
- The subset of traditional meta-analysis consists of 150 total meta-analyses
- The subset of traditional meta-analysis from different subfields of psychology (Social psychology, organizational psychology and cognitive psychology; 50 each)
- additionally 57 multisite replication studies and registered reports
- for each meta-analysis information about the total sample size and effect size of each primary study was collected
- Random sampling to determine the meta-analysis for the study (within the inclusion criteria)
- One important inclusions criteria of Linden & Hönekopp (2021), effects had to be reported as standardized mean differences (Cohen’s d or Hedges’g) or as correlations (Pearson’s r or Fisher’s z)
- In instances where a meta-analysis or replication study used a different effect size measure than Cohen’s d, the effect sizes were transformed accordingly

Statistical Analysis

All statistical analyses were performed using R (version 4.4.0, Team, 2023). Data and analysis scripts are made available members of Goethe University on the Local Instrastructure for Open Science (LIFOS) and the publicly on the Open Science Framework (OSF).

Regarding the hypotheses, in which the publication bias parameter ω_{PBS} was the dependent variable (\mathcal{H}_1 , \mathcal{H}_2 , \mathcal{H}_4), beta regression models as implemented in the *betareg* package (Zeileis et al., 2021) was used to analyse the data. This choice is motivated by the restriction of the parameter space for the publication bias to the standard unit interval, whereby non-normality, skewness and heteroscedasticity can anticipated (Cribari-Neto & Zeileis, 2010; Smithson & Verkuilen, 2006). Beta regression is recognized for its adaptability in handling such deviations. We used a logit link for the mean parameter μ and a identity

link for the dispersion parameter that was fixed such that the beta regression model can be described as

$$\begin{aligned}\omega_{\text{PBS}_i} &\sim \mathcal{B}(\mu_i, \phi) \\ \log\left(\frac{\mu_i}{1-\mu_i}\right) &= x_i^\top \beta.\end{aligned}\tag{10}$$

The independent variables for these three hypotheses were as follows: regarding \mathcal{H}_1 , the independent variable was the Fisher z -transformed correlation coefficient of the correlation between effect size and sample size, where the transformation is defined as $z_r = \tanh^{-1}(r)$. The independent variable for \mathcal{H}_2 was the difference $\Delta_{\hat{\mu}_d}$ between the average effect size estimate of each meta-analysis and the estimated mean parameter of the Gaussian effect size distribution from the SPEEC approach. Lastly, the independent variable for \mathcal{H}_4 was a binary indicator specifying the research synthesis type (traditional meta-analysis or multisite replications), with multisite replication studies set as the reference level for regression. The coefficients from the beta-regressions for these hypotheses were estimated using Maximum Likelihood estimation with the BFGS optimizer.

Hypothesis \mathcal{H}_3 was aimed at comparing the estimated means of the Gaussian effect size distribution to the average effect sizes to assess whether the presence of effects in mean differences $\Delta_{\hat{\mu}_d}$ deemed large enough to be considered meaningful, according to specified equivalence bounds Δ_{EQ} , can be rejected (Lakens et al., 2020). For this, we conducted an equivalence test using the Two One-Sided Tests procedure (Schuirmann, 1987) as implemented in the *TOSTER* R package (Lakens & Caldwell, 2023). To perform the TOST procedure, the *TOSTER* R package (Lakens & Caldwell, 2023) was utilized. We employed Welch's two-sample t -tests for dependent samples with corrected degrees of freedom as the Welch's t -test generally offers better control of type I error rates when the samples are heteroscedastic, while maintaining robustness compared to Student's t -test when test assumptions are satisfied (Delacre et al., 2017). Furthermore, the choice of a dependent samples test was necessitated by the dependence between the pairs of samples originating from the same underlying data. The equivalence bounds Δ_{EQ} against which the data was tested were defined by the smallest effect size of interest.

Smallest Effect Size of Interest

For all four hypotheses, we will establish the smallest effect size of interest (SESOI) based on effect sizes that can reliably detected, considering the constraints imposed by the sample size resources available for this secondary data analysis (Lakens, 2014; Lakens et

al., 2018). More specifically, we conducted three simulation-based (\mathcal{H}_1 , \mathcal{H}_3 , \mathcal{H}_4) and one analytical (\mathcal{H}_2) sensitivity power analysis to determine which effect sizes we have at least 80 percent power ($1 - \beta = 0.8$) to detect, taking into account the constraints of the sample size and a fixed type I error rate $\alpha = .05$ (details see Appendix XXX). We specified the SESOI for \mathcal{H}_3 in raw units and all other SESOIs in odds ratios. The SESOI for the equivalence hypothesis will define the equivalence bounds for the TOST procedure ($\Delta_{EQ} = (-0.17, 0.17)$). Table XXX summarises all four SESOIs of the hypotheses.

Table 1

Smallest Effect Sizes of Interest of the Hypotheses

Hypothesis	SESOI	Unit
\mathcal{H}_1	1.28	<i>OR</i>
\mathcal{H}_2	0.59	<i>OR</i>
\mathcal{H}_3	0.17	raw unit
\mathcal{H}_4	1.28	<i>OR</i>

Note. *OR* = Odds ratio

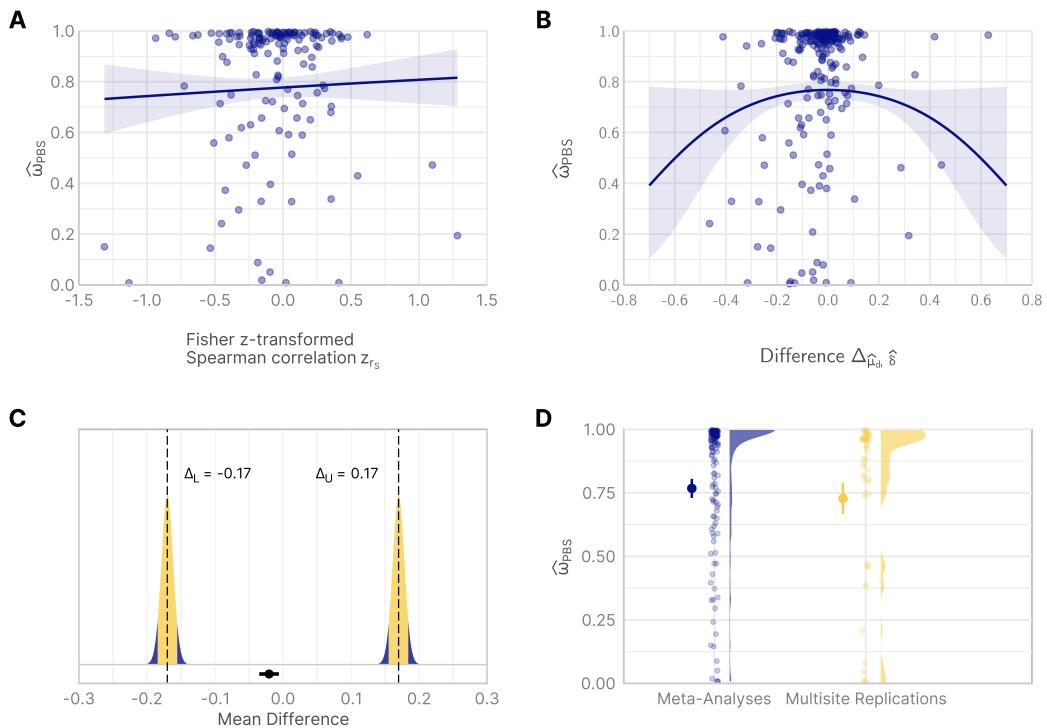
Results

Variable Dispersion of ω_{PBS}

As an initial step, the assumptions made to determine the Smallest Effect Sizes of Interest (SESOI) for the four hypotheses were assessed. In the simulations-based sensitivity power analyses aimed at ascertaining the SESOIs (refer to the Appendix for details), three dispersion parameter conditions $\phi : \{10, 20, 30\}$ for the distribution of the publication bias parameter ω_{PBS} were simulated. Employing an intercept-only beta regression model with the complete dataset, the estimated dispersion parameter was $\hat{\phi} = 1.56$, 95% CI [1.27, 1.84], $SE = 0.15$, $z = 10.72$, $p < .001$. This finding contradicts our initial assumptions regarding the dispersion parameter's magnitude, rendering the interpretation of SESOIs for our hypotheses untenable. Consequently, it is appropriate to refrain from interpreting SESOIs in the subsequent analyses.

Figure 3

Visual Summary of Results from Four Hypotheses



Note. **A.** Estimated publication bias parameter vs. Fisher z-transformed correlation coefficients. Fitted line: regression coefficients with 95

Confirmatory Results of the Predictions from the Hypotheses

```
initial value -53.486433
```

```
iter 10 value -129.284617
final value -129.284618
converged
```

Regarding hypothesis \mathcal{H}_1 , panel A of figure 2 depicts the relationship between estimated publication bias parameter $\hat{\omega}_{\text{PBS}}$ and the Fisher z -transformed Spearman correlation coefficients z_{r_s} of the effect size sample size association in each meta-analysis. The observed slope was marginally positive, the sign of the coefficient was in the direction of the hypothesis, however statistically non-significant $OR = 1.20$, $95\% CI [0.77, \text{Inf}]$, $SE = 0.27$, $z = 0.69$, $p = .245$. This indicates, that lower values of z_{r_s} were not statistically significantly associated with lower publication bias parameter values $\hat{\omega}_{\text{PBS}}$. Additionally, to enhance the interpretability of the regression slope, we refitted the model with standardized values of z_{r_s} and computed the average marginal effects (Arel-Bundock, 2024). On average, for every standard deviation increase in the Fisher z -transformed correlation coefficient z_{r_s} ($SD(z_{r_s}) = 0.31$), the model only predicted an increase of 1.11% in the publication bias parameter $\hat{\omega}_{\text{PBS}}$. In line with this, the general explanatory power of the model as determined by the pseudo R^2 (Ferrari & Cribari-Neto, 2004) was low, $R^2 = 0.003$.

Concerning \mathcal{H}_2 , panel B of figure 1 depicts the relationship between estimated publication bias parameter as a function of the difference between the average effect size $\hat{\delta}$ and the estimated mean parameter of the Gaussian effect size distribution $\hat{\mu}_d$. The corresponding estimated quadratic slope was negative as indicated by the predicted concave inverse u-shaped line and statistically significant at an α -level of 5\%, $OR = 0.04$, $95\% CI [0.00, 0.73]$, $SE = 1.84$, $z = -1.81$, $p = .035$. We again calculated the average marginal effect for improved interpretability. On average, for every standard deviation increase in $\Delta_{\hat{\mu}_d, \hat{\delta}}$ ($SD(\Delta_{\hat{\mu}_d, \hat{\delta}}) = 0.13$), the model only predicted an increase of -0.09% in the publication bias parameter $\hat{\omega}_{\text{PBS}}$. The overall explained variation of ω_{PBS} by $\Delta_{\hat{\mu}_d, \hat{\delta}}$ was low, $R^2_{\text{pseudo}} = 0.026$.

In relation to hypothesis \mathcal{H}_3 , panel C of Figure 2 illustrate the mean difference $\Delta_{\hat{\mu}_d, \hat{\delta}}$ between the estimated mean parameter of the Gaussian effect size distribution $\hat{\mu}_d$ and the average effect size $\hat{\delta}$, along with its corresponding confidence interval. Additionally, the null t -distributions of the Two One-Sided Tests (TOST) against the equivalence bounds $\Delta_{EQ} = (-0.17, 0.17)$ are illustrated. We only report the results of the t-test with the lower t -value in the main results as both tests must be significant to reject the null hypothesis (Lakens, 2017). Both one-sided paired t-tests were statistically significant, $t(56) = 17.3$, $SE = 0.01$, $p < .001$. This is also indicated by 90% confidence interval lying within the equivalence range in panel C of Figure 2. We additionally conducted an exploratory null hypothesis significance test to test the point hypothesis that the true mean difference of $\Delta_{\hat{\mu}_d, \hat{\delta}}$ is exactly zero. The mean

difference significantly deviated from zero $M = -0.02$, 90% CI [-0.03, -0.01], $t(56) = -2.36$, $SE = 0.01$, $p = 0.022$. This indicates that, despite the significant null hypothesis significance test, the difference was too small to be considered meaningful according to the equivalence range $\Delta_{EQ} = (-0.17, 0.17)$ of the equivalence test.

Finally, regarding hypothesis \mathcal{H}_4 , panel D of Figure 2 illustrates the comparison between the estimated publication bias parameters for typical meta-analysis in comparison to multisite replication studies / registered reports. Already descriptively, contrary to our expectation that the estimated publication bias parameters for multisite replication studies (MR) would be greater (i.e., lower publication bias) than for regular meta-analysis (MA), the mean of the estimated publication bias values ω_{PBS} of the regular meta-analysis subset is greater than the mean of the multisite replication subset ($M_{MA} = 0.82$; $M_{MR} = 0.79$). In line with this, slope of the beta regression was non-significant, $OR = 0.81$, 95% CI [0.61, Inf], $SE = 0.18$, $z = -1.17$, $p = .879$, as also indicated by the overlapping confidence interval of the predicted marginal means in panel D. Once more, we computed the average marginal effect to examine how the estimated publication bias parameter $\hat{\omega}_{PBS}$ changes with the discrete shift from the reference level (multisite replication studies / registered reports) to typical meta-analysis, as predicted by the regression model, revealing a change of -3.93% in the opposing direction of the hypothesis.

Diagnostic Evaluation of Parameter Estimation in SPEEC

As this study relies on empirical data to preliminarily assess the proposed SPEEC approach, the true values for the distributional parameters and the publication bias parameter are unknown. However, as discussed previously, publication bias is inherently absent by design in multisite replication studies and registered reports. Thus, the four distributional parameters (μ_d , σ_d^2 , μ_n , ϕ_n) within the SPEEC approach cannot be biased due to publication bias (especially the mean and variance of the effect size distribution). Leveraging this fact, we can use a subset of the data encompassing the multisite replication studies and registered reports for a diagnostic evaluation of the parameter estimation within the SPEEC approach. More specifically, we can derive Maximum Likelihood estimates for the distributional parameters to compare them with the corresponding values estimated by the SPEEC approach, anticipating approximate equivalence between the two approaches. This part was of the analysis was not preregistered and conceived after the confirmatory analyses were conducted. Based on this comparative approach between ML and SPEEC, we formulated five diagnostic questions to assess parameter estimation:

1. To what degree do the estimated distributional parameters differ between SPEEC and

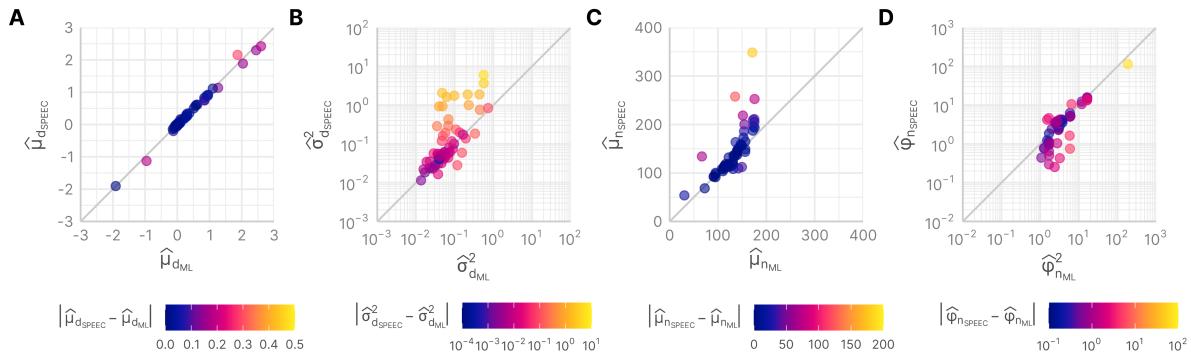
MLE?

2. How are the discrepancies in one parameter associated with those in the other distributional parameters across SPEEC and MLE? Specifically, does a consistency exist in the discrepancies between these parameters?
3. Does the discrepancy between SPEEC and MLE estimates of the distributional parameters correlate with the sample size of the multisite replication studies?
4. Is the discrepancy between SPEEC and MLE in the distributional parameters associated with sample size of multisite replication studies k ?
5. Is the discrepancy between SPEEC and MLE in the distributional parameters associated with the publication bias parameter ω_{PBS} ?

The Maximum Likelihood estimates for the distributional parameters were obtained using the Nelder-Mead optimization algorithm (Nelder & Mead, 1965). Additionally, the mean and median discrepancy between SPEEC and MLE were calculated to descriptively to assess the average difference between the two estimation methods.

Figure 4

Scatter Plot comparing the estimated Distributional Parameters via SPEEC and Maximum Likelihood



Note. **A1.** Comparison of estimated mean parameter μ_d from Gaussian effect size distribution. **A2.** Comparison of estimated variance parameter σ_d^2 of Gaussian effect size distribution. Axes and colorbar are log (base 10) transformed. **B1.** Comparison of mean parameter μ_n of Negative-Binomial sample size distribution **B2.** Comparison of dispersion parameter ϕ_n of Negative-Binomial sample size distribution. Axes and colorbar are log (base 10) transformed.

Regarding question one, Figure 3 provides a visual summary of the analysis comparing the distributional parameters estimated by the SPEEC method against those estimated by Maximum Likelihood Estimation (MLE). The diagonal line signifies perfect alignment between MLE and SPEEC estimates. Values below the diagonal indicate higher values for MLE compared to SPEEC, while values above the diagonal indicate the opposite.

Panel A of figure XX reiterates the findings of the analysis of \mathcal{H}_3 , suggesting a small discrepancy between the two methods in estimating the mean of the Gaussian effect size distribution, $M(\Delta_{\mu_d}) = -0.03$, $Mdn(\Delta_{\mu_d}) = -0.02$. This discrepancy can be deemed practically negligible according to the equivalence test of \mathcal{H}_3 . However, the other panels indicate contrasting outcomes. Panel B reveals a systematic discrepancy in the estimation of the variance parameter of the Gaussian effect size distribution between SPEEC and MLE, $M(\Delta_{\sigma_d^2}) = 0.59$, $Mdn(\Delta_{\sigma_d^2}) = 0.11$. Descriptively, this suggests that on average, the variance was estimated to be greater in the SPEEC approach compared to MLE. Furthermore, this discrepancy increases in exponential trend and displays substantial heteroscedasticity with rising variance estimates from the MLE approach. Similarly, Panel C also illustrates a systematic overestimation of the mean parameter of the Negative-Binomial sample size distribution by SPEEC in comparison to MLE ($M(\Delta_{\mu_n}) = 46.96$, $Mdn(\Delta_{\mu_n}) = 3.52$), which again increases in an exponential trend with higher mean parameter estimates from MLE. Lastly, Panel D shows that the SPEEC approach generally underestimates dispersion parameter of the sample size distribution in comparison to the ML estimate ($M(\Delta_{\phi_n}) = -2.02$, $Mdn(\Delta_{\phi_n}) = -0.19$) and also furthermore indicates a systematic relationship in the discrepancy between the two approaches.

Table 2

Pairwise Pearson Correlations between the Absolute Difference of the Distributional Parameters from SPEEC and ML, Publication Bias Parameter and Meta-Analysis Size

Variable	ω_{PBS}	$ \Delta_{\mu_d} $	$ \Delta_{\sigma_d^2} $	$ \Delta_{\phi_n} $	$ \Delta_{\mu_n} $
ω_{PBS}					
$ \Delta_{\mu_d} $	-0.44*** [-0.61, -0.23]				
$ \Delta_{\sigma_d^2} $	-0.37** [-0.56, -0.14]	0.67*** [0.54, 0.76]			
$ \Delta_{\phi_n} $	0.08 [-0.18, 0.33]	-0.06 [-0.31, 0.2]	-0.05 [-0.3, 0.21]		
$ \Delta_{\mu_n} $	0.03 [-0.23, 0.29]	0.15 [-0.11, 0.39]	0.04 [-0.22, 0.3]	-0.05 [-0.3, 0.21]	
k	-0.14 [-0.38, 0.12]	0.04 [-0.22, 0.29]	0.15 [-0.11, 0.39]	-0.17 [-0.41, 0.09]	0 [-0.26, 0.26]

Note. Computed p -values are corrected for multiple comparison using the correction by Benjamini & Hochberg (1995).

$|\Delta|$ is the absolute difference for each distributional parameter between SPEEC and MLE.

* Significance *** $p < .001$; ** $p < .01$; * $p < .05$

To address the remaining diagnostic questions, we conducted a pairwise correlational analysis between the absolute differences of the parameter estimates derived from the two estimation methods, alongside the publication bias parameter $\hat{\omega}_{PBS}$ and the total number of primary replication studies k within each multisite replication project or registered report.

We used the Pearson correlation coefficient and corrected the obtained p -values for multiple comparisons to control the false discovery rate (Benjamini & Hochberg, 1995).

Regarding the parameters of the Gaussian effect size distribution, a strong positive correlation was observed between the absolute difference in the mean parameter estimates and the variance parameter estimates obtained from ML and SPEEC. This indicates that as the absolute discrepancy between SPEEC and ML increased for the mean parameter μ_d , the absolute discrepancy also increased for the variance parameter σ_d^2 of the effect size distribution. Furthermore, strong negative correlations were found between the publication bias parameter ω_{PBS} and the discrepancy between ML and SPEEC estimates of the mean and variance parameters of the effect size distribution. More specifically, as the absolute discrepancy between both estimation methods increased for both the mean ($|\Delta_{\mu_d}|$) and variance ($|\Delta_{\sigma_d^2}|$), the publication bias parameter ω_{PBS} decreased, signifying more severe predicted publication bias. Notably, the total number of primary replications k was not significantly associated to the divergence of ML and SPEEC of any distributional parameter or the publication bias parameter ω_{PBS} .

Discussion

Test

References

- Allaire, J. J., Teague, C., Scheidegger, C., Xie, Y., & Dervieux, C. (2024). *Quarto*. <https://doi.org/10.5281/zenodo.5960048>
- Arel-Bundock, V. (2024). *Marginal effects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests*. <https://marginaleffects.com/>
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66(6), 423–437. <https://doi.org/10.1037/h0020412>
- Begg, C. B. (1994). Publication bias. In *The handbook of research synthesis* (pp. 399–409). Russell Sage Foundation.
- 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>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300. <https://www.jstor.org/stable/2346101>
- Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist*, 27(8), 774–775. <https://doi.org/10.1037/h0038034>
- Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A Meta-Meta-Analysis: Empirical Review of Statistical Power, Type I Error Rates, Effect Sizes, and Model Selection of Meta-Analyses Published in Psychology. *Multivariate Behavioral Research*, 45(2), 239–270. <https://doi.org/10.1080/00273171003680187>
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., ... Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637–644. <https://doi.org/10.1038/s41562-018-0399-z>
- Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, 6(1), 29–42. <https://doi.org/10.1038/s41562-021-01193-7>
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12), 997–1003. <https://doi.org/10.1037/0003-066X.49.12.997>
- Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 3–16). Russell Sage Foundation. <http://www.scopus.com/inward/record.url?scp=84902712953&partnerID=8YFLLogxK>

- Cribari-Neto, F., & Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, 34, 1–24. <https://doi.org/10.18637/jss.v034.i02>
- Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of student's t-test. *International Review of Social Psychology*, 30(1), 92–101. <https://doi.org/10.5334/irsp.82>
- Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *JAMA*, 263(10), 1385–1389.
- Dickersin, K., & Min, Y.-I. (1993). Publication Bias: The Problem That Won't Go Away. *Annals of the New York Academy of Sciences*, 703(1), 135–148. <https://doi.org/10.1111/j.1749-6632.1993.tb26343.x>
- Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., ... Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68–82. <https://doi.org/10.1016/j.jesp.2015.10.012>
- Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., ... Nosek, B. A. (2020). Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Advances in Methods and Practices in Psychological Science*, 3(3), 309–331. <https://doi.org/10.1177/2515245920958687>
- Edelbuettel, D. (2022). *RcppDE: Global Optimization by Differential Evolution in C++*. <https://cran.r-project.org/web/packages/RcppDE/index.html>
- Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ (Clinical Research Ed.)*, 315(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>
- Feoktistov, V. (2006). *Differential Evolution – In Search of Solutions* (Vol. 5). Springer. <https://doi.org/10.1007/978-0-387-36896-2>
- Ferrari, S., & Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31(7), 799–815. <https://doi.org/10.1080/0266476042000214501>
- 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>

- Fritz, A., Scherndl, T., & Kühberger, A. (2013). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? *Theory & Psychology*, 23(1), 98–122. <https://doi.org/10.1177/0959354312436870>
- Gerber, A. S., Green, D. P., & Nickerson, D. (2001). Testing for Publication Bias in Political Science. *Political Analysis*, 9(4), 385–392. <https://www.jstor.org/stable/25791658>
- Ghoreishi, N., Clausen, A., & Jørgensen, B. N. (2017). Termination Criteria in Evolutionary Algorithms: A Survey. *Proceedings of 9th International Joint Conference on Computational Intelligence*, 1, 373–384. <https://doi.org/10.5220/0006577903730384>
- Harrer, M., Cuijpers, P., A, F. T., & Ebert, D. D. (2021). *Doing Meta-Analysis With R: A Hands-On Guide* (1st ed.). Chapman & Hall/CRC Press.
- Jain, B. J., Pohlheim, H., & Wegener, J. (2001). On termination criteria of evolutionary algorithms. *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, 768–768.
- Jennions, M. D., & Møller, A. P. (2002a). Publication bias in ecology and evolution: An empirical assessment using the “trim and fill” method. *Biological Reviews*, 77(2), 211–222. <https://doi.org/10.1017/S1464793101005875>
- Jennions, M. D., & Møller, A. P. (2002b). Relationships fade with time: A meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, 269(1486), 43–48. <https://doi.org/10.1098/rspb.2001.1832>
- Kicinski, M. (2014). How does under-reporting of negative and inconclusive results affect the false-positive rate in meta-analysis? A simulation study. *BMJ Open*, 4(8), e004831. <https://doi.org/10.1136/bmjopen-2014-004831>
- Kitcher, P. (1993). *The Advancement of Science: Science Without Legend, Objectivity Without Illusions*. Oxford University Press.
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., ... Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45(3), 142–152. <https://doi.org/10.1027/1864-9335/a000178>
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., ... Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443–490. <https://doi.org/10.1177/2515245918810225>

- Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLoS ONE*, 9(9), e105825. <https://doi.org/10.1371/journal.pone.0105825>
- Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86. <https://doi.org/10.1214/aoms/1177729694>
- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7), 701–710. <https://doi.org/10.1002/ejsp.2023>
- Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, 8(4), 355–362. <https://doi.org/10.1177/1948550617697177>
- Lakens, D., & Caldwell, A. (2023). *TOSTER: Two One-Sided Tests (TOST) Equivalence Testing*. <https://cran.r-project.org/web/packages/TOSTER/index.html>
- Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving Inferences About Null Effects With Bayes Factors and Equivalence Tests. *The Journals of Gerontology: Series B*, 75(1), 45–57. <https://doi.org/10.1093/geronb/gby065>
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample Sizes and Effect Sizes are Negatively Correlated in Meta-Analyses: Evidence and Implications of a Publication Bias Against NonSignificant Findings. *Communication Monographs*, 76(3), 286–302. <https://doi.org/10.1080/03637750903074685>
- Light, R., & Pillemer, D. (1984). *Summing up: The science of reviewing research*. Harvard University Press. https://scholars.unh.edu/psych_facpub/194
- Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of Research Results: A New Perspective From Which to Assess and Promote Progress in Psychological Science. *Perspectives on Psychological Science*, 16(2), 358–376. <https://doi.org/10.1177/1745691620964193>
- Linden, A., Pollet, T. V., & Hönekopp, J. (2024). *Publication Bias in Psychology: A Closer Look at the Correlation Between Sample Size and Effect Size*. <https://doi.org/10.31234/osf.io/s4znd>
- Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, 51(3), 485–504. <https://doi.org/10.1002/ejsp.2752>
- Marks-Anglin, A., & Chen, Y. (2020). A historical review of publication bias. *Research Synthesis Methods*, 11(6), 725–742. <https://doi.org/10.1002/jrsm.1452>

- Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, 112(2), 331–348. <https://doi.org/10.2466/03.11.PMS.112.2.331-348>
- Merkel, D. (2014). Docker: Lightweight Linux containers for consistent development and deployment. *Linux Journal*, 2014(239), 2:2.
- Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrew, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). *Sustainable data analysis with Snakemake*. F1000Research. <https://doi.org/10.12688/f1000research.29032.1>
- Møller, A., & Jennions, M. (2001). How important are direct fitness benefits of sexual selection? *Naturwissenschaften*, 88(10), 401–415. <https://doi.org/10.1007/s001140100255>
- Munafò, M. R., & Flint, J. (2010). How reliable are scientific studies? *The British Journal of Psychiatry*, 197(4), 257–258. <https://doi.org/10.1192/bjp.bp.109.069849>
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 1–10. <https://doi.org/10.1038/s41562-016-0021>
- Nelder, J. A., & Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7(4), 308–313. <https://doi.org/10.1093/comjnl/7.4.308>
- Nosek, B. A., & Lakens, D. (2014). Registered Reports: A Method to Increase the Credibility of Published Results. *Social Psychology*, 45(3), 137–141. <https://doi.org/10.1027/1864-9335/a000192>
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, 7(6), 615–631. <https://doi.org/10.1177/1745691612459058>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>
- Palmer, A. R. (1999). Detecting Publication Bias in Meta-analyses: A Case Study of Fluctuating Asymmetry and Sexual Selection. *The American Naturalist*, 154(2), 220–233. <https://doi.org/10.1086/303223>
- Renkewitz, F., & Keiner, M. (2019). How to Detect Publication Bias in Psychological Research. *Zeitschrift Für Psychologie*, 227(4), 261–279. <https://doi.org/10.1027/2151-2604/a000386>
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>

- Sassenberg, K., & Ditrich, L. (2019). Research in Social Psychology Changed Between 2011 and 2016: Larger Sample Sizes, More Self-Report Measures, and More Online Studies. *Advances in Methods and Practices in Psychological Science*, 2(2), 107–114. <https://doi.org/10.1177/2515245919838781>
- Schuirmann, D. J. (1987). A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6), 657–680. <https://doi.org/10.1007/BF01068419>
- Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (2011). Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology*, 96(5), 1055–1064. <https://doi.org/10.1037/a0023322>
- Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An Introduction to Registered Replication Reports at Perspectives on Psychological Science. *Perspectives on Psychological Science*, 9(5), 552–555. <https://www.jstor.org/stable/44290039>
- Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly*, 43(3), 290–322. <https://doi.org/10.1598/RRQ.43.3.4>
- Slavin, R., & Smith, D. (2009). The Relationship Between Sample Sizes and Effect Sizes in Systematic Reviews in Education. *Educational Evaluation and Policy Analysis*, 31(4), 500–506. <https://doi.org/10.3102/0162373709352369>
- Smart, R. G. (1964). The importance of negative results in psychological research. *Canadian Psychologist / Psychologie Canadienne*, 5a(4), 225–232. <https://doi.org/10.1037/h0083036>
- Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1), 54–71. <https://doi.org/10.1037/1082-989X.11.1.54>
- Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., Hing, C., Kwok, C. S., Pang, C., & Harvey, I. (2010). Dissemination and publication of research findings : An updated review of related biases. *Health Technology Assessment*, 14(8), 1–220. <https://doi.org/10.3310/hta14080>
- 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>
- Sterling, T. D. (1959). Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa. *Journal of the American Statistical Association*, 54(285), 30–34. <https://doi.org/10.1080/01621459.1959.10501497>

- Storn, R., & Price, K. (1997). Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, 11(4), 341–359. <https://doi.org/10.1023/A:1008202821328>
- Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, 15(3), e2000797. <https://doi.org/10.1371/journal.pbio.2000797>
- Team, R. C. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org>
- Ushey, K., & Wickham, H. (2024). *Renv: Project Environments*. <https://cran.r-project.org/web/packages/renv/index.html>
- Vevea, J. L., Coburn, K., & Sutton, A. J. (2019). Publication Bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 383–433). Russell Sage Foundation.
- Weinerová, J., Szűcs, D., & Ioannidis, J. P. A. (2022). Published correlational effect sizes in social and developmental psychology. *Royal Society Open Science*, 9(12), 220311. <https://doi.org/10.1098/rsos.220311>
- Zeileis, A., Cribari-Neto, F., Gruen, B., Kosmidis, I., by), A. B. S. (earlier. version, & by), A. V. R. (earlier. version. (2021). *Betareg: Beta Regression*. <https://cran.r-project.org/web/packages/betareg/index.html>

Appendix

Appendix A: Power Analyses determining the SESOIs

The simulated-based sensitivity power analysis targeted a statistical power of $1 - \beta = 0.8$ with a fixed significance level of $\alpha = .05$. The simulated samples sizes were specified according to the hypotheses as follows:

- \mathcal{H}_1 : $n = 150$ (only traditional meta-analyses)
- \mathcal{H}_2 and \mathcal{H}_4 : $n = 207$ (both traditional meta-analyses and multisite replication studies)
- \mathcal{H}_3 : $n = 57$ (only multisite replication studies)

The distributional assumptions for the four sensitivity power analyses were specified as follows:

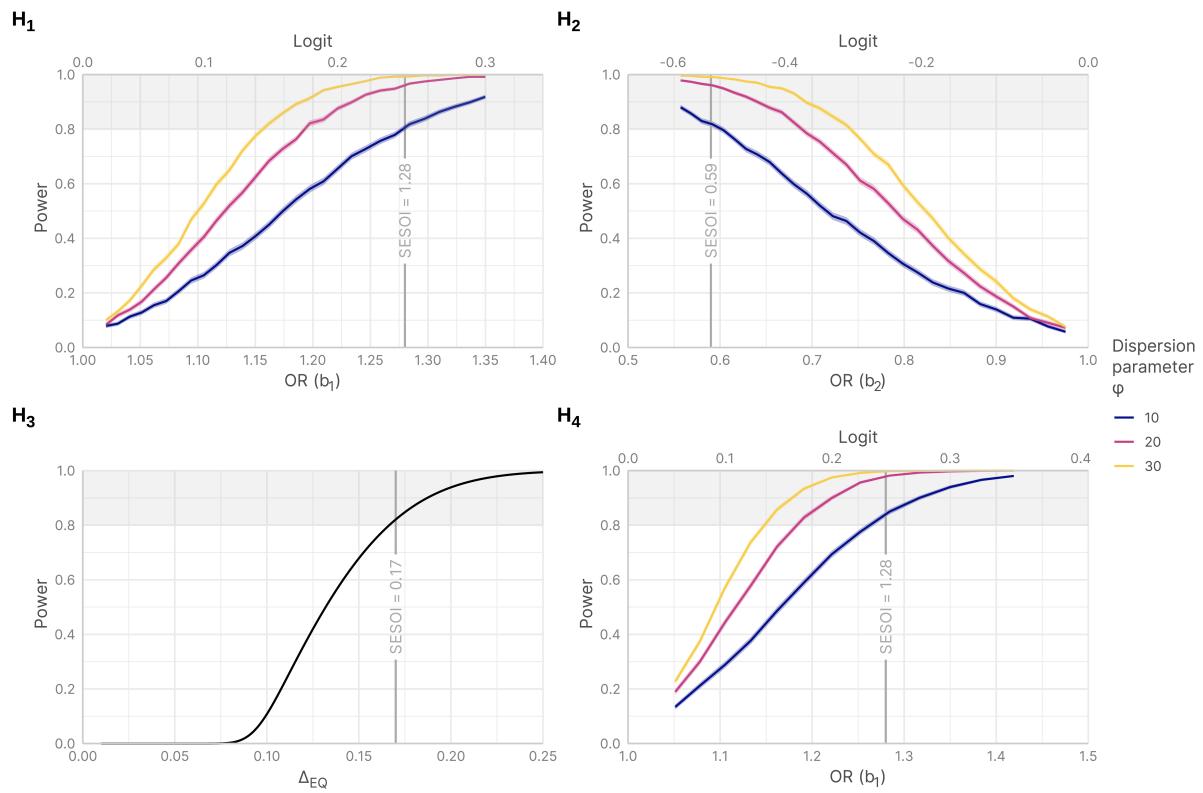
$$\mathcal{H}_1 : z_{r_S} \sim \mathcal{N}(\mu = -0.1, \sigma = 0.5)$$

$$\mathcal{H}_2 \text{ and } \mathcal{H}_3 : \Delta_{\mu_d} \sim \mathcal{N}\left(\mu = 0, \sigma_{diff} = \sqrt{0.3^2 + 0.3^2}\right)$$

For hypothesis four, the proportions of the categorical predictor of the research synthesis type (traditional meta-analysis MA , multisite replication MR) were chosen according the the actual proportions of the data ($n_{MA} = 150, n_{MR} = 57$). For all simulations-based sensitivity power analyses ($\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_4$), the number of simulations was set to $n_{iter} = 5000$. More over, the beta-regressions on ω_{PBS} in $\mathcal{H}_1, \mathcal{H}_2$, and \mathcal{H}_4 involved simulations for different dispersion parameter conditions $\phi = \{10, 20, 30\}$, as lower dispersion parameters result in reduced test power. We set the SESOI for the parameters of interest more conservatively, ensuring a minimum power of 80% for the lowest dispersion parameter $\phi = 10$.

Figure 5

Power Curves of the Sensitivity Power Analyses determining the SESOIs



Note. OR: Odds ratio. Ribbons around the lines represent the 95

Appendix B: Research Transparency Statement

As transparent and reproducible as possible. The preregistration of study, the data needed to reproduce the results are open and data analysis scripts that allow reproducing all reported results are all openly available at URL

- The project is fully st containerized using *Docker* (Merkel, 2014)
- Software and operating system virtualization, fully containerized project using Docker R package dependencies and version management using *renv* (Ushey & Wickham, 2024).
- Data analysis workflow management tool *Snakemake* (Mölder et al., 2021)
- thesis is written using the *Quarto* (Allaire et al., 2024) open-source scientific and technical publishing system

Appendix C: Deviations of Preregistration
Appendix D: Test
Table 3

Estimated Parameters for the Distribution of Effect Size and Sample Size from each Meta-Analysis via ML

Parameter	Minimum	Maximum
Effect Size		
$\hat{\mu}_d$	-1.911	2.599
$\hat{\sigma}_d^2$	0.003	4.349
Sample Size		
$\hat{\phi}_n$	0.042	176.620
$\hat{\mu}_n$	17.365	1438.443

Note. Maximum Likelihood Estimation using the Nelder-Mead optimizer.

Appendix E: Regression Tables of Confirmatory Analyses
Table 4

Beta Regression Results for \mathcal{H}_1

Term	Estimate	CI (95%)	SE	z	p
Mean model component: μ					
Intercept	3.49 ^a	[2.79, 4.36]	0.11	10.93	< .001
z_{r_S}	1.20 ^a	[0.71, 2.04]	0.27	0.69	.245
Precision model component: ϕ					
b_0	5.41 ^b	[3.73, 7.84]	0.19	8.89	< .001

Note. $LL = 129.28$, $MAE = 0.22$, $AIC = -252.57$, $BIC = -243.54$, $R^2 = 0.003$

^a OR

^b Identity coefficient

Table 5*Beta Regression Results for \mathcal{H}_2*

Term	Estimate	CI (95%)	SE	<i>z</i>	<i>p</i>
Mean model component: μ					
Intercept	3.30 ^a	[2.71, 4.03]	0.10	11.79	< .001
Quadratic	0.04 ^a	[9.58e-04, 1.31]	1.84	-1.81	.035
$\Delta_{\hat{\mu}_d, \hat{\delta}}$					
Precision model component: ϕ					
Intercept	4.85 ^b	[3.63, 6.47]	0.15	10.70	< .001

Note. $LL = 164.25$, $MAE = 0.23$, $AIC = -322.51$, $BIC = -312.51$, $R^2 = 0.026$

^a *OR*

^b Identity

Table 6*Two One-Sided Tests Result for \mathcal{H}_3*

Type	<i>t</i>	SE	<i>df</i>	<i>p</i>
NHST	-2.36	0.009	56	.022
TOST $\Delta < \Delta_L$	17.30	0.009	56	< .001
TOST $\Delta > \Delta_L$	-22.02	0.009	56	< .001

Note. NHST: Null Hypothesis Significance Test, TOST: Two One-Sided Test

Table 7*Beta Regression Results for \mathcal{H}_4*

Term	Estimate	CI (95%)	SE	<i>z</i>	<i>p</i>
Mean model component: μ					
Intercept	3.30 ^a	[2.67, 4.08]	0.11	11.07	< .001
Research	0.81 ^a	[0.57, 1.15]	0.18	-1.17	.879
Synthesis Type (MR)					
Precision model component: ϕ					
Intercept	4.79 ^b	[3.59, 6.38]	0.15	10.71	< .001

Note. MR: Multisite Replication; $LL = 163.31$, $MAE = 0.23$, $AIC = -320.62$, $BIC = -310.62$, $R^2 = 0.011$

^a *OR*

^b Identity coefficient