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

Darmstadt, 21. Mai 2024

Goethe-Universität Frankfurt am Main

Fachbereich 5: Institut für Psychologie und Sportwissenschaften

Abteilung für Psychologische Methoden mit Interdisziplinärer Ausrichtung

Erstgutachter: Prof. Dr. Martin Schultze

Zweitgutachter: Julia Beitner, M.Sc.

Abstract

The high prevalence of publication bias has been a long-recognised concern for the validity of meta-analyses and, more recently, has been identified as a major contributor to the replication crisis leading to an overestimation of meta-analytic evidence. Although numerous statistical methods exist to assess publication bias, many have significant limitations by lacking an explicit generative publication bias model and suffering from limited performance under various conditions frequently encountered in empirical meta-analytical data. The present study introduces SPEEC, a flexible and unified simulation-based framework to assess publication bias and estimate bias-corrected effect sizes based on the joint distribution of effect size and sample size. SPEEC integrates explicit assumptions of the generative process of publication bias in its model. Effect size and sample size data are simulated from this model, and their estimated joint kernel density is iteratively compared with empirical data to adjust the model parameters, minimising the Kullback-Leibler divergence as a statistical distance between the simulated and empirical data. The feasibility of SPEEC was evaluated in a proof-of-concept study using empirical data encompassing both classical meta-analyses ($n = 150$) and publication bias absent registered replication reports ($n = 57$). The confirmatory results highlighted potential challenges in the parameter estimation, while additional exploratory results suggested that between-study heterogeneity could be a significant factor contributing to these challenges. The results are discussed in terms of potential adaptations to SPEEC to address these challenges, additional factors that could be considered in the simulation framework to better model publication bias, and recommendations for future more comprehensive assessments of SPEEC.

Keywords: publication bias, meta-analysis, meta-science, simulation

Table of contents

Abstract **ii**

Introduction **1**
- Consequences of Publication Bias 1
- Methods to Assess Publication Bias and Their Limitations 2
- The Present Study 3
- A Primer on SPEEC 4
- Confirmatory Hypotheses 5

Method **8**
- Comprehensive Methodological Description of SPEEC 8
- Secondary Data Description 15
- Statistical Analysis 15
- Smallest Effect Size of Interest 17

Confirmatory Results **18**

Intermediate Discussion **21**

Diagnostic Evaluation of Parameter Estimation in SPEEC **22**

General Discussion **26**
- Extent and Consistency of Misestimation of SPEEC Parameters 27
- Which Factors Drive the Parameter Misestimation in SPEEC? 28
- Limitations 29
- Directions for Future Research 30
- Conclusion 31

References **32**

Appendix **42**
- Appendix A: Research Transparency Statement 42
- Appendix B: Power Analyses determining the SESOIs 42
- Appendix C: MLE Extrema Distributional Parameters 44
- Appendix D: Sensitivity Analyses of Equivalence Test 44
- Appendix E: Regression Tables of Confirmatory Analyses 46

Introduction

Science is commonly conceived as a cumulative endeavour with the overarching goal of establishing robust knowledge about the world upon which the future of scientific inquiry may be constructed (Curran, 2009). As part of this endeavour, the idea of meta-analyses - quantitatively synthesising research examining the same phenomena - has been attributed a pivotal role in contributing to the cumulative advancement of knowledge (Schmidt, 1992, 1996). However, this premise rests on the fundamental assumption that the research published in the scientific literature is representative of all conducted research (Rothstein et al., 2005; Song et al., 2010). Yet scholars have been stressing for over half a century that systematic differences exist between the results of published and unpublished studies (Bakan, 1966; Bozarth & Roberts, 1972; Smart, 1964; Sterling, 1959). In practice, the publication of a study often hinges on the strength or direction of its findings, collectively known as *publication bias* (Dickersin, 1990; Dickersin & Min, 1993). Especially in a publishing culture that incentivises novelty and positive results, alongside the predominant reliance on null hypothesis significance testing, it has become common practice for the nominal false positive rate to serve as a de facto criterion for publication (Nosek et al., 2012). As a consequence, many statistically nonsignificant studies end up in the "file drawer" and never get published (Rosenthal, 1979).

Consequences of Publication Bias

The ramifications of publication bias are severe, leading to an inflation of meta-analytic effect sizes (Franco et al., 2014; Stanley et al., 2021) and an elevated false-positive rate (Ioannidis, 2005; Kicinski, 2014; Munafò & Flint, 2010), thus increasing the risk of erroneous conclusions that may jeopardise the validity of the research (Begg, 1994). Moreover, the prevalence of questionable research practices such as p-hacking (John et al., 2012) exacerbates the issue, as they interact with publication bias to further collectively distort meta-analytical effect sizes (Friese & Frankenbach, 2020). These ramifications become especially relevant in light of recent large-scale replication projects providing evidence for non-replicability of many psychological findings (Camerer et al., 2018; Ebersole et al., 2016; Ebersole et al., 2020; Klein et al., 2014; Klein et al., 2018; Open Science Collaboration, 2015). This underscores why publication bias is identified as a major threat to replicable science (Munafò et al., 2017) and thus is considered a

significant driver of the replication crisis in psychology (Renkewitz & Keiner, 2019). The myriad issues associated with publication bias and its widespread impact have fueled a great deal of research focusing on statistical methods to detect and address publication bias.

Methods to Assess Publication Bias and Their Limitations

This attention has led to the development of numerous statistical methods to detect and address publication bias over the past decades (Marks-Anglin & Chen, 2020). These statistical techniques can generally be classified into methods based on the relationship between effect size and sample size in meta-analyses and those working with p-values (Vevea et al., 2019).

The former class of methods, also coined small-study effects (Sterne et al., 2000), relies on the idea that, in the presence of publication bias, studies with smaller sample sizes (lower precision) necessitate larger effect sizes to attain statistical significance compared to studies with larger effect sizes (higher precision). Consequently, there are disproportionally fewer studies with low sample sizes and low effect sizes because they are statistically nonsignificant. Visualising the distribution of studies' sample sizes and effect sizes in a funnel plot will display asymmetry. This asymmetry is reflected in a correlation between the effect size and a precision measure, which is then analysed using regression-based methods. Such methods include, for example, PET-PEESE (Stanley & Doucouliagos, 2014), Egger's regression (Egger et al., 1997), Begg's rank correlation (Begg & Mazumdar, 1994), or in its most simplistic form, the correlation between effect size and sample size (e.g., Kühberger et al., 2014).

The latter class of methods based on p-values prominently features publication bias selection methods, including earlier selection models (Hedges, 1984; e.g., Hedges, 1992; Iyengar & Greenhouse, 1988), alongside more recent developments such as p-uniform (Van Assen et al., 2015) and p-curve analysis (Simonsohn et al., 2014). Publication bias selection models aim to directly characterise the selective publication process and consider the likelihood of the publication of a study as a function of p-values (Marks-Anglin & Chen, 2020).

Despite the abundance of statistical methods to assess publication bias, some justified criticisms have been discussed in the literature. Firstly, small study effects methods are commonly criticised for their lack of an explicit model for publication

bias (McShane et al., 2016) and for their only "indirect" approach as they omit the true mechanism of publication bias by being driven on effect sizes rather than p-values (Harrer et al., 2021). Additionally, regarding these methods, it has been discussed that publication bias is not the only factor influencing the distribution of effect size and sample size. Researchers may plan sample sizes before performing the study according to anticipated effect sizes (Linden et al., 2024; Schäfer & Schwarz, 2019), which could compromise the validity of the interpretation of such methods.

More broadly, several comprehensive simulation studies have demonstrated that many existing methods perform poorly across a range of scenarios with realistic settings commonly encountered in empirical meta-analyses (Carter et al., 2019; McShane et al., 2016; Renkewitz & Keiner, 2019; Van Aert et al., 2019). Such influencing factors include the extent of effect size heterogeneity, the prevalence of additional p-hacking and other questionable research practices, the limited number of individual studies included in the meta-analysis, and the severity of publication bias. These factors may, in turn, lead to reduced statistical power, elevated false positive rates, convergence issues, and unsatisfactory agreement among different methods.

Considering these limitations, an explicit modelling framework to assess publication bias could, therefore, be valuable, with explicit assumptions within this framework that can be flexibly adapted for different scenarios. This framework should be, in principle, capable of modelling relevant factors previously discussed in the context of modelling publication bias, such as heterogeneity, sample size planning based on anticipated effect sizes, and the potential modelling of p-hacking.

The Present Study

The present study introduces SPEEC (**S**imulation-based **P**ublication bias **E**stimation and **E**ffect size **C**orrection), a novel simulation-based framework to assess the extent of publication bias in meta-analyses and estimate corrected effect sizes in the presence of publication bias based on the joint distribution of effect size and sample size. The thesis has two primary objectives. Firstly, it aims to introduce the reader to the SPEEC method and comprehensively describe its assumptions and procedure. Secondly, it aims to assess the SPEEC method in a proof of concept using secondary empirical meta-analytical data sourced from Linden and Hönekopp (2021) to assess the feasibility of the introduced method. For this purpose, four theoretically

justifiable hypotheses are derived in the section Confirmatory Hypotheses. These involve predictions regarding the estimated parameters of the SPEEC method that should be corroborated by the empirical data. The thesis is structured as follows: In this section, a brief primer on the central ideas of the SPEEC method is provided, followed by a detailed derivation of the hypotheses. Next, a thorough introduction to the SPEEC method itself is offered. This is followed by the empirical analyses of the hypotheses and a discussion and evaluation of the results of the empirical analyses.

A Primer on SPEEC

The fundamental concept underlying the SPEEC approach involves explicitly modelling the generative process of publication bias in a simulation framework, and iteratively comparing how the distribution of effect size and sample size of simulated studies diverges from the actual empirical meta-analytical data to estimate the model's parameters.

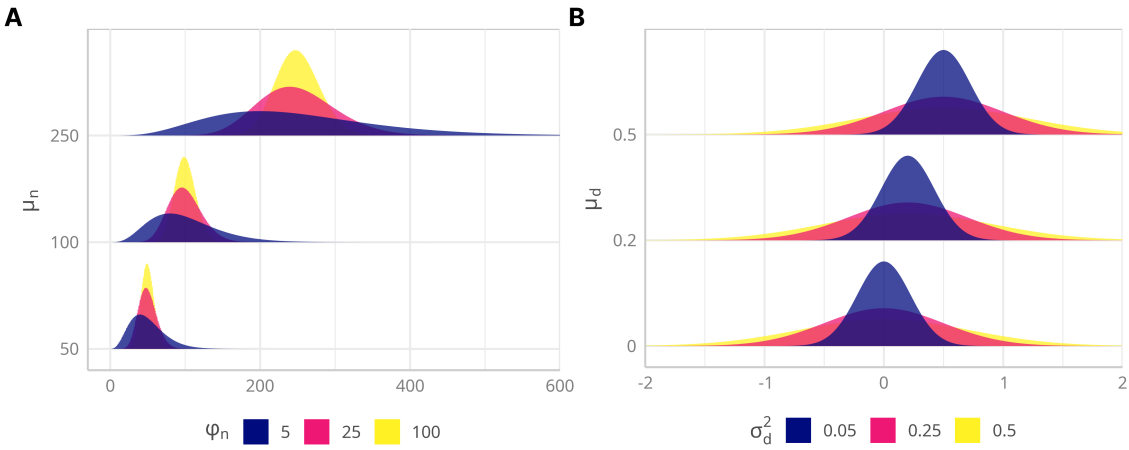
The publication bias model of the simulation framework integrates both assumptions concerning the marginal distribution of effect size and sample size and how publication bias influences their joint distribution. In terms of the former, the sample sizes are modelled by a negative-binomial distribution (with parameters μ_n and ϕ_n), and the effect sizes are modeled by a Gaussian distribution (with parameters μ_d and σ_d^2), where mean parameter μ_d represents the publication bias-corrected effect size estimate. Regarding the latter, the extent of publication bias is modelled by a publication bias parameter, ω_{PBS}, which captures the publication probability of a non-significant study relative to a significant one (for details see section Application of Publication Bias). For instance, a publication bias parameter of $\omega_{\text{PBS}} = 0.5$ would indicate that nonsignificant simulated studies are half as likely to be selected (i.e., published) compared to statistically significant studies. The parameters of SPEEC are further illustrated in *Figure* 1. From this generative publication bias model, individual studies' effect sizes and sample sizes are simulated.

Subsequently, the estimated kernel density distributions of the simulated data from the generative model are compared to those of the empirical meta-analytical data. This comparison serves to quantify the statistical divergence between the data generated by the model and the empirical data, functioning as a loss function. This framework can be conceptualised as an optimisation problem aiming to find values for the distributional

parameters and the publication bias parameter of the generative publication bias model such that the statistical divergence from the empirical data is minimised. Stochastic optimisation algorithms can then be utilized to estimate the parameters of the generative publication bias model.

Figure 1

Distributional Parameters of the SPEEC Method



Note. **A**. Probability mass function of the Negative-Binomial distribution with mean μ_n and dispersion ϕ_n. The dispersion parameter controls how much larger the variance is compared to the mean of the negative-binomial distributed data ($\mathbb{E}[n] = \mu_n, \mathbb{V}[n] = \mu_n + \mu_n^2/\phi_n$). **B**. Probability density function of the normal distribution with mean μ_d and variance σ_d^2.

Confirmatory Hypotheses

To assess the SPEEC method, a set of four theoretical predictions is derived, constituting the hypotheses of this study. These hypotheses serve as benchmarks for assessing the viability of the proposed method and are, therefore, expected to hold true if the approach works in principle. If the predictions fail to be corroborated by the empirical meta-analytical data, this would raise concerns about the viability of the SPEEC method and necessitate a further review of its implementation.

Firstly, regarding $\mathcal{H}^{(1)}$, we conducted a direct comparison between the discussed correlation of effect size and sample size, serving as an alternative indicator of publication bias and the publication bias parameter ω_{PBS} estimated within the SPEEC method. It can be expected that the estimated publication bias parameter $\hat{\omega}_{\text{PBS}}$ is positively associated with the Fisher z-transformed Spearman correlation coefficients of the association between the unsigned effect size and sample size in each meta-analysis z_{r_s}. In other words, when the proposed method estimates high publication bias (i.e., low values for

$\widehat{\omega}_{\text{PBS}}$), the correlation coefficients for each meta-analysis z_{r_s} are expected to be more negative and conversely. In statistical terms, this implies that the regression coefficient $\beta_{z_{r_s}}$ predicting the publication bias parameter ω_{PBS} is expected to be greater than zero.

$$\mathcal{H}_0^{(1)} : \beta_{z_{r_s}} \leq 0 \qquad \mathcal{H}_1^{(1)} : \beta_{z_{r_s}} > 0 \tag{1}$$

In cases where substantial publication bias is present within the scientific literature of a particular research phenomenon, and the true effect size is precisely zero ($\delta = 0$), the distribution of effect size and sample size exhibits increased symmetric sparsity around zero in areas where individual studies would not be statistically significant for a given effect size and sample size (Light & Pillemer, 1984). This is explained by the fact that only studies with either large positive or large negative effects will be statistically significant and consequently have a higher likelihood of being published in the presence of publication bias. Because of this symmetry for a true effect size of zero, the average effect size $\hat{\delta}$ should not be biased since negative and positive effects should, in theory, mutually cancel each other out. Consequently, the difference Δ_{μ_d} between the average effect size $\hat{\delta}$ and the estimated mean parameter $\hat{\mu}_d$ of the effect size distribution from the SPEEC approach should remain invariant, independent of the magnitude of publication bias. However, when the true effect size exceeds zero ($\delta > 0$), publication bias leads to an overestimation of the true effect (i.e. $\hat{\delta} > \delta$), and conversely, overestimation in the opposite direction (i.e., $\hat{\delta} < \delta$) when $\delta < 0$. Suppose, compared to the mean effect size $\hat{\delta}$, the estimated mean parameter $\hat{\mu}_d$ of the Gaussian effect size distribution obtained from the SPEEC approach is a more accurate estimate of the true effect size δ in the presence of publication bias. In that case, it follows from the prior reasoning that a curvilinear, inverted U-shaped pattern can be expected between the difference Δ_{μ_d} of these two parameters and the publication bias parameter ω_{PBS}. In other words, when the mean difference Δ_{μ_d} is approaching zero, publication bias severity is expected to decrease (indicated by larger values for ω_{PBS}). Conversely, when the difference Δ_{μ_d} diverges from zero in both negative and positive directions, publication bias severity is expected to increase (i.e., lower values for ω_{PBS}). In statistical terms, the quadratic regression term $\beta_{\Delta_{\mu_d}}$ is expected to be smaller than zero.

$$\mathcal{H}_0^{(2)} : \beta_{\Delta_{\mu_d}} \geq 0 \qquad \mathcal{H}_1^{(2)} : \beta_{\Delta_{\mu_d}} < 0 \tag{2}$$

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 study‘s outcome because the decision to publish was made before the results of the studies were known, eliminating publication bias (Chambers & Tzavella, 2022; Simons et al., 2014). Consequently, the effect sizes within registered replication projects (RRR) cannot be biased by publication bias. Following this reasoning, for the third hypothesis $\mathcal{H}^{(3)}$, average effect size of a RRR $\hat{\delta}$ can be expected to be equivalent to the mean parameter of the Gaussian effect size distribution μ_d ($\Delta_{\hat{\mu}_d} = \hat{\delta} - \hat{\mu}_d$) that is estimated using the SPEEC approach within specified equivalence bounds $\Delta_{EQ} = \{\Delta_L,\ \Delta_U\}$. The equivalence bounds are determined by the smallest effect size of interest for this study (Lakens et al., 2018) and are defined as $\Delta_{EQ} = \{-0.17, 0.17\}$ (see section Smallest Effect Size of Interest for the rationale of this decision).

$$\begin{gathered} \mathcal{H}_{01}^{(3)} :\ \Delta_{\mu_d} \leq \Delta_L \quad \cap \quad \mathcal{H}_{02}^{(3)} :\ \Delta_{\mu_d} \geq \Delta_U \\ \mathcal{H}_{1}^{(3)} :\ \Delta_L > \Delta_{\mu_d} > \Delta_U \end{gathered} \tag{3}$$

In theory, one could posit that the publication bias parameter ω_{PBS} in RRRs should precisely equal $\omega_{\text{PBS}} = 1$, as individual replication studies are predetermined in the registration of the study and included in the final report independent of their statistical outcomes (Nosek & Lakens, 2014). However, due to the inherent upper limit of $\omega_{\text{PBS}} = 1$, testing this point prediction would not be sensible in a null hypothesis significance testing framework. Instead, a relative comparison between traditional meta-analyses and RRRs allows for making testable predictions. More specifically, it can be expected that the publication bias parameter ω_{PBS} is greater for RRRs than for traditional meta-analyses. This indicates that relative to individual statistically nonsignificant primary studies within traditional meta-analyses, statistically nonsignificant replication studies should have a greater likelihood of being included in the final RRR. In statistical terms, when the regressor is a binary indicator of the type of research synthesis with the reference level being registered replication reports and the outcome is the estimated publication bias parameter ω_{PBS}, the regression coefficient β_{RRR} can be expected to be greater than zero.

$$\mathcal{H}_0^{(4)}: \ \beta_{RRR} \leq 0 \qquad \mathcal{H}_1^{(4)}: \ \beta_{RRR} > 0 \tag{4}$$

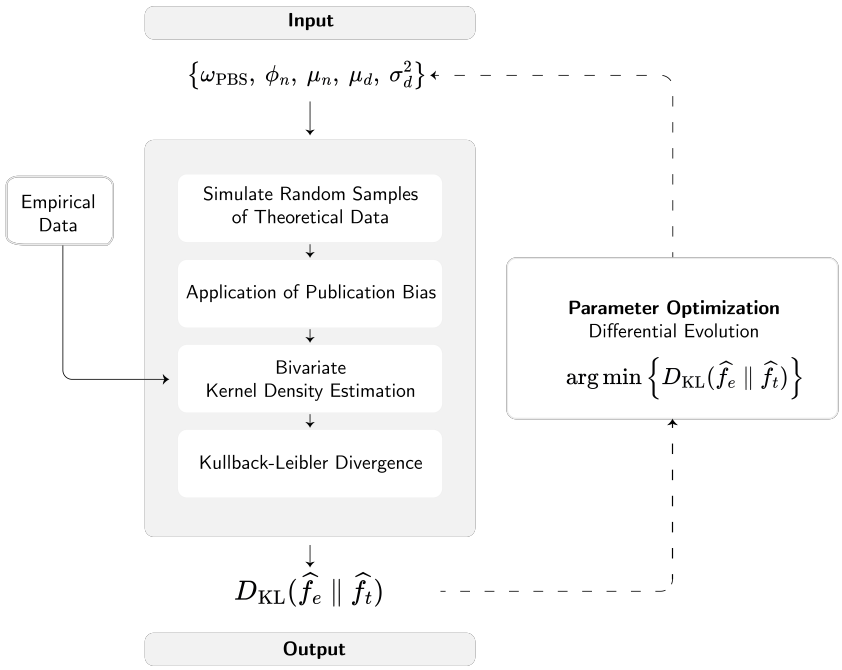
Method

Comprehensive Methodological Description of SPEEC

This section offers a comprehensive description and explanation of the SPEEC framework, which assesses the extent of publication bias and estimates effect sizes in the presence of publication bias in meta-analyses. The flowchart in *Figure* 2 provides an overview of the sequential steps of the SPEEC method. SPEEC is being developed as an open-source R package and is accessible on GitHub at https://github.com/jlschnatz/speec (alpha version).

Figure 2

Flowchart of the SPEEC Method



Note. Flowchart illustrating the steps of the SPEEC method. The simulation framework of the publication bias model is depicted in the left box. Inputs are the publication bias parameter ω_{PBS} and distributional parameters of the effect size and sample size distribution (μ_d, σ_d^2, μ_n, ϕ_n). The output serves as the loss function for optimisation and is the Kullback-Leibler divergence $D_{\text{KL}}(\hat{f}_e \parallel \hat{f}_t)$ between the estimated kernel density of the empirical meta-analytica data $\hat{f}_e$ and theoretical simulated data $\hat{f}_t$ from the publication bias model. Optimisation is performed using differential evolution to the find the minimum of $D_{\text{KL}}(\hat{f}_e \parallel \hat{f}_t)$.

Simulation Framework

The initial step in the SPEEC method entails defining the marginal distributions of effect size and sample size for the generative publication bias model. This requires additional assumptions regarding the type of study design from which the simulated effect sizes and sample sizes originate. For the purpose of this study, Cohen's d is adopted as a widely used effect size measure for mean differences (Lakens, 2013). Accordingly, the simulated effect sizes are assumed to originate from a between-subjects, two-sample t-test study design. However, it is worth noting that the SPEEC framework is adaptable to different study designs where alternative effect size measures are typically employed (e.g., correlational effect sizes or effect sizes derived from proportional data). Depending on the effect size, this would require adapting the marginal distribution of the effect size and how statistical significance is determined based on a different test statistic, which is required for the application of publication bias (see Application of Publication Bias).

The marginal distribution for the total sample size n of a study should be inherently modelled as a discrete distribution. Count data of this nature are commonly modelled using either a Poisson or Negative-Binomial distribution. In various psychological domains, sample size distributions often exhibit considerable variance and skewness (e.g., Cafri et al., 2010; Marszalek et al., 2011; Sassenberg & Ditrich, 2019; Shen et al., 2011; Szucs & Ioannidis, 2017). Considering this, the negative-binomial distribution was chosen because it can model overdispersion more effectively by introducing a second parameter (Lloyd-Smith, 2007). The mean-dispersion parametrisation of the negative-binomial distribution is used where the probability of success p_n and the target number of successes r_n is reparametrised to mean $\mu_n = \frac{r_n \cdot (1-p_n)}{p_n}$ and dispersion $\phi_n = r_n$ to model the study-specific total sample sizes n_i.

$$n_1,\, n_2,\, \ldots,\, n_k \quad \text{where} \quad N \sim \mathcal{NB}(\phi_n, \mu_n) \tag{5}$$

Concerning the marginal distribution of effect size d, it is reasonable to assume a Gaussian distribution with mean μ_d and variance σ_d^2 (e.g., Borenstein et al., 2010). To account for the increasing precision in estimating the true effect size mean μ_d as the sample size increases (i.e., the sampling error), the variance of the mean differences $\bar{x}_{i1} - \bar{x}_{i2}$ are computed, from which the effect sizes are assumed to originate in this type of study design. Subsequently, a normalisation factor γ_i can be derived by scaling each

individual variance $\sigma^2_{\bar{x}_{i1}-\bar{x}_{i2}}$ with the overall mean of those variances such that $\bar{\gamma} = 1$.

$$\begin{aligned} \sigma^2_{\bar{x}_{i1}-\bar{x}_{i2}} &= \sigma^2_d / n_i \\ \gamma_i &= \frac{\sigma^2_{\bar{x}_{i1}-\bar{x}_{i2}}}{\sum_{i=1}^{k} \sigma^2_{\bar{x}_{i1}-\bar{x}_{i2}} / k} \end{aligned} \tag{6}$$

With this normalisation factor, the total variance of the individual variances is $\mathbb{E}(\gamma_i \cdot \sigma^2_d) = \sigma^2_d$. The study-specific effect sizes d_i are subsequently modelled as

$$d_1,\, d_2,\, \dots,\, d_k \quad \text{where} \quad d_i \sim \mathcal{N}(\mu_d, \gamma_i \cdot \sigma^2_d) \quad \text{for} \quad i = 1, \dots, k. \tag{7}$$

Using this definition for the marginal distribution of effect size, the SPEEC method assumes a fixed-effect meta-analytical model, where the only source of variation in effect size is sampling error. However, the publication bias simulation model of the SPEEC method could be further extended to account for effect size heterogeneity by including an additional parameter τ^2 in the simulation framework to account for variability beyond sampling error.

Conditional on the marginal distributions, k studies are drawn from the joint distribution of effect size and sample size, where k is user-defined. An increase in k reduces the uncertainty in the joint distribution but also comes with a trade-off of increased computational complexity. For our analysis, we selected $k = 10^4$ samples for each simulation iteration to analyse the hypotheses.

Application of Publication Bias

Following the simulation step of sampling k studies from the joint distribution of effect size and sample size, the subsequent stage involves the application of publication bias to the studies. As previously discussed, publication bias is operationalized in terms of the likelihood of a study being published conditional on the statistical significance of its results. Translated to this simulation setting, two-tailed p-values for each individual study i can be calculated from the corresponding effect size d_i and sample size n_i. To implement this, it is assumed that the individual studies i originate from a balanced sample size design. When the total sample size n_i is even, the between-subject group sample sizes n_{1i} and n_{2i} are defined as $n_i/2$. Otherwise, when the total sample size n_i is odd, the group sample sizes are determined as the ceilinged $\lceil n_i/2 \rceil$ and floored $\lfloor n_i/2 \rfloor$ values. Subsequently, the p-value p_i of each simulated study can be derived from its

corresponding t-value

$$t_i = \left| \frac{d_i}{\sqrt{1/n_{1i} + 1/n_{2i}}} \right| \qquad p_i = 2 \cdot P(t_i \mid df_i) \tag{8}$$

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

$$\omega_{\text{PBS}_i}(p_i) = \begin{cases} \omega_{\text{PBS}} & \text{for} \quad p_i \geq \alpha \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

with the constraint $\omega_{\text{PBS}} \in \mathbb{R} : 0 \leq \omega_{\text{PBS}} \leq 1$. This weight denotes the probability of a study i being selected conditional on the p-value and the type I error rate α. This definition of the publication bias weight is directly analogous to the step weighting function used in publication bias selection models (see Hedges, 1992; Iyengar & Greenhouse, 1988). If a study i is statistically nonsignificant (i.e., $p_i \geq \alpha$), the publication bias parameter ω_{PBS} is assigned; otherwise the probability of a study being selected is equal to 1, indicating no publication bias. Thus, the publication bias parameter ω_{PBS} denotes the probability of selecting a nonsignificant study relative to a statistically significant one. Importantly, the type I error rate needs to be fixed across all simulated studies and is set to the common nominal value of $\alpha = 0.05$ for the analyses of the hypotheses. Following the computation of the publication bias weight ω_{PBS_i} for each study i, the publication bias selection process can be defined in terms of an indicator function $1_{\omega_{\text{PBS}}}(\omega_{\text{PBS}_i})$ and k draws from a uniform distribution $u_1, u_2, \dots, u_k \sim \mathcal{U}_{[0,1]}$ where

$$1_{\omega_{\text{PBS}}}(\omega_{\text{PBS}_i}) \begin{cases} 1 & \text{if} \quad \omega_{\text{PBS}_i} \geq u_i \quad \text{study selected} \\ 0 & \text{if} \quad \omega_{\text{PBS}_i} < u_i \quad \text{study not selected.} \end{cases} \tag{10}$$

Using this indicator function, the resulting subsets for the selected $\{d_i', n_i'\}$ and non-selected samples $\{d_i'', n_i''\}$ from the initial k simulated studies can be defined as

$$\begin{aligned} \{d_i', n_i'\} &= \{d_i, n_i \mid 1_{\omega_{\text{PBS}}}(\omega_{\text{PBS}_i}) = 1\} \quad \text{for} \quad i = 1, \dots, k' \quad \text{and} \\ \{d_i'', n_i''\} &= \{d_i, n_i \mid 1_{\omega_{\text{PBS}}}(\omega_{\text{PBS}_i}) = 0\} \quad \text{for} \quad i = 1, \dots, k''. \end{aligned} \tag{11}$$

One challenge in simulating a fixed number of k studies is that, ceteris paribus, with increasing publication bias (i.e. lower values for ω_{PBS}), increasingly fewer studies k' remain after the selection process compared to the original number of simulations k. This process would lead to a loss of precision in the parameter estimation for decreasing values of ω_{PBS}. To address this, the steps mentioned above (from Equation 5 to Equation 11) are repeated a second time with an adjusted number of simulations $k_{\text{adj}} = \lceil k^2/k' \rceil$, thereby ensuring that the number of selected studies k'_{adj} is approximately equal over the entire range of ω_{PBS}.

Divergence Between Empirical and Simulated Data from the Publication Bias Model

After the simulation of k_{adj} samples from the publication bias model conditional on the marginal distributional parameters (μ_d, σ_d^2, ϕ_n, μ_n) and the publication bias parameter ω_{PBS} to determine the subset $\{d'_i, n'_i\}$ for $i = 1, \ldots, i = k'_{\text{adj}}$, the following step involves quantifying how closely the simulated data from the publication bias model aligns with the empirical meta-analytical data. More specifically, this step involves measuring the statistical dissimilarity of the bivariate kernel density estimates between the empirical meta-analytical data and the simulated data of the publication bias model. While various measures of statistical dissimilarity between probability distributions exist (for an overview, see Cha, 2007), the Kullback-Leibler divergence (D_{KL}, Kullback & Leibler, 1951) was chosen as a dissimilarity measure for SPEEC. This choice was motivated by superior performance of KLD in capturing the dissimilarity between estimated kernel densities compared to alternative measures (total variation distance and earthmover distance) tested for the SPEEC method, particularly across boundary conditions. In addition, D_{KL} has an intuitive interpretation in this context. It can be interpreted as the expected value of the log-likelihood ratio favouring the true model over a candidate model (Etz, 2018). Here, the estimated joint density of effect size and sample size from the meta-analytical data $\hat{f}_e$ is regarded as the true data generating distribution and the estimated kernel density of the simulated data $\hat{f}_t$ from the publication bias model serves as the approximate candidate distribution.

To implement this step, the *KernSmooth* R package (version 2.23.22, Wand et al., 2023) was used to estimate the joint kernel density distribution for both empirical and simulated data using a bivariate standard Gaussian kernel that is evaluated on

a linearly-binned square grid (Wand, 1994; Wand & Jones, 1994). The grid size was chosen as $n_{\text{grid}} = 2^7 + 1$ equidistant grid points in each dimension such that the total number of cells is $n_{\text{grid}} \times n_{\text{grid}}$, but is user-definable in the R package for SPEEC. The bandwidth of the kernel function was determined using the reliable plug-in method proposed by Sheather and Jones (1991), but the R package also offers other common bandwidth selection methods. To ensure the comparability of the estimated kernel densities between the empirical and simulated data, the bounds of the square grid, $\mathrm{b_n} \times \mathrm{b_d}$, must be identical and are determined from the empirical meta-analytical data.

For defining these bounds, the maximum likelihood values 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 estimated. These estimates are used to determine the quantiles that cover the inner 99% of the cumulative distribution from the quantile functions $Q_d(p \mid \hat{\mu}_d,\ \hat{\sigma}_d^2)$ and $Q_d(p \mid \hat{\phi}_n,\ \hat{\mu}_n)$ for the percentiles $p_1 = 0.5$ and $p_2 = 99.5$. Subsequently, the bounds for the effect size $\mathrm{b_d}$ and sample size $\mathrm{b_n}$ are defined as the minimum and maximum values of these quantiles and the range of the empirical data, respectively.

$$\begin{aligned} \mathrm{b_d} &= \{Q_d(p_1 \mid \hat{\mu}_d,\ \hat{\sigma}_d^2) \wedge \min(d),\ Q_d(p_2 \mid \hat{\mu}_d,\ \hat{\sigma}_d^2) \vee \max(d)\} \\ \mathrm{b_n} &= \left\{Q_n(p_1 \mid \hat{\phi}_n,\ \hat{\mu}_n) \wedge \min(n),\ Q_n(p_2 \mid \hat{\phi}_n,\ \hat{\mu}_n) \vee \max(n)\right\} \end{aligned} \tag{12}$$

This ensures that an adequate range is covered for the kernel density estimation. Finally, D_{KL} is calculated from the binned kernel density estimates of the empirical $\hat{f}_e$ and simulated theoretical $\hat{f}_t$ data.

$$D_{\mathrm{KL}}(\hat{f}_e \parallel \hat{f}_t) = \sum_{u=1}^{n_{\text{grid}}} \sum_{v=1}^{n_{\text{grid}}} \hat{f}_e(u,v) \ln\left(\frac{\hat{f}_e(u,v)}{\hat{f}_t(u,v)}\right) \tag{13}$$

Formulation as an Optimisation Problem

Summarising the previous sections, the simulation framework within the publication bias model of SPEEC requires the distributional parameters for the marginal distributions of effect size and sample size and the publication bias parameter as input parameters. It returns a single scalar value representing the Kullback-Leibler divergence D_{KL} between the estimated joint kernel density of the simulated theoretical data and the empirical meta-analytical data. This framework can be considered as an optimisation problem of finding parameter values such that

$$\begin{gathered}\min_{\mu_d,\ \sigma_d^2,\ \mu_n,\ \phi_n,\ \omega_{PBS}} \Big\{ D_{\mathrm{KL}}(\hat{f}_e \parallel \hat{f}_t) \Big\}, \\ \text{subject to: } \mu_d,\ \sigma_d^2,\ \mu_n,\ \phi_n,\ \omega_{PBS} \in \mathbb{R}, \quad \text{where} \quad 0 \leq \omega_{PBS} \leq 1, \\ -4 \leq \mu_d \leq 4, \quad 0 \leq \sigma_d^2 \leq 6, \quad 10 \leq \mu_n \leq 15000, \quad 0.01 \leq \phi_n \leq 1000.\end{gathered} \tag{14}$$

Parameter Optimisation with Differential Evolution

For this purpose, we use differential evolution (DE, Storn & Price, 1997), a simple global optimisation algorithm (Feoktistov, 2006). DE is an evolutionary meta-heuristic based on principles such as mutation, cross-over, and selection. It only requires a few control parameters that are generally straightforward to select to achieve favourable outcomes compared to other optimisation algorithms (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 apply DE for the optimisation, the R package *RcppDE* (version 0.1.7, Edelbuettel, 2022) was utilised, which implements the classical DE algorithm $DE/rand/1$ (Storn & Price, 1997). The control parameters for DE were chosen based on the recommendations of Storn and 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 cross-over constant CR to 0.1. In applying 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 optimisation problem at hand (Jain et al., 2001), the choice for $t_{\max}$ was also informed by preliminary testing of the simulation framework of SPEEC. These tests suggested that $t_{\max} = 1000$ is a reasonable decision. In establishing the bounds for the parameter search space (see Equation 14), a balance was struck between avoiding bounds that were too wide, which could lead to ineffective exploration of the search space, and ensuring that the bounds were not too narrow to provide sufficient coverage of the potential parameters. More specifically, the minima and maxima for all distributional parameters were determined using maximumm likelihood (see Appendix C *Table C1*), and the

boundaries were set slightly above those values to ensure good coverage.

Secondary Data Description

To examine the confirmatory hypotheses of this study, secondary data sourced from previous research by Linden and Hönekopp (2021) is utilised. The dataset can be accessed both from its source (see https://osf.io/yr3xd/) and through the repositories associated with this project (see section Appendix A). The dataset is comprehensive and includes both classical meta-analyses and registered replication reports for which publication bias is assumably absent. The dataset encompasses a total of 207 research syntheses covering various psychological phenomena. It includes 150 meta-analyses, each subset encompassing 50 meta-analyses from different subfields of psychology (social psychology, organisational psychology, and cognitive psychology). Additionally, the dataset includes 57 registered replication reports, which are particularly relevant for investigating hypotheses $\mathcal{H}^{(3)}$ and $\mathcal{H}^{(4)}$. Information on the total sample size n_i and effect size d_i for each primary study was gathered for each research synthesis. The meta-analyses were selected using random sampling, with predefined inclusion criteria and specified journals from which the data were selected. One important inclusion criterion by Linden and Hönekopp (2021) was that effect sizes must be reported as standardised mean differences (Cohen's d or Hedges' g) or as correlations (Pearson's r or Fisher's z). In cases where a different effect size measure than Cohen's d was used, effect sizes were transformed accordingly (Linden & Hönekopp, 2021).

Statistical Analysis

The confirmatory statistical analyses were preregistered based on an adaption of the secondary data analysis preregistration template by Van den Akker et al. (2021). The preregistration includes additional metadata on the dataset and a transparency statement regarding prior data knowledge. Any supplementary analyses not preregistered and deviations of the preregistered confirmatory analyses are explicitly labelled as such. All data, analysis scripts and the preregistration are made available (see Appendix A). All confirmatory analyses were performed using the R programming language (version 4.4.0 Puppy Cup, R Core Team, 2023).

Regarding the hypotheses, in which ω_{PBS} was the dependent variable ($\mathcal{H}^{(1)}$, $\mathcal{H}^{(2)}$, $\mathcal{H}^{(4)}$), beta regression as implemented in the *betareg* package (version 3.1.4,

Zeileis et al., 2021) was used to analyse the data. This choice was motivated by the restriction of the parameter space for ω_{PBS} to the standard unit interval, whereby non-normality, skewness, and heteroscedasticity can be anticipated (Cribari-Neto & Zeileis, 2010; Smithson & Verkuilen, 2006). Beta regression is recognised for its adaptability in handling such deviations. Because the optimisation approach permits $\hat{\omega}_{\text{PBS}}$ values within the range $0 \leq \hat{\omega}_{\text{PBS}} \leq 1$, and the beta regression model assumes that $0 < \hat{\omega}_{\text{PBS}} < 1$, a common transformation proposed by Smithson and Verkuilen (2006) was employed. This transformation, denoted as $\omega'_{\text{PBS}} = (\omega_{\text{PBS}} \cdot (n-1) + 0.5)/n$, subtly adjusts the bounds to slightly narrow the range between zero and one. We used a logit link for the mean parameter μ and an identity link for the fixed dispersion parameter such that the beta regression model can be described as

$$\begin{aligned} \omega_{\text{PBS}'_i} &\sim \mathcal{B}(\mu_i, \phi) \\ \log\left(\frac{\mu_i}{1-\mu_i}\right) &= \mathbf{X_i}^\top \boldsymbol{\beta}. \end{aligned} \tag{15}$$

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 $\Delta_{\widehat{\mu}_d}$ between the average effect size estimate of each meta-analysis and the mean parameter of the Gaussian effect size distribution estimated with SPEEC. Lastly, the independent variable for $\mathcal{H}^{(4)}$, was a binary indicator specifying the research synthesis type (classical meta-analysis or RRR), with RRR set as the reference level for regression. The coefficients from the beta regressions for these hypotheses were estimated using Maximum Likelihood estimation with the L-BFGS optimiser (Liu & Nocedal, 1989).

Hypothesis $\mathcal{H}^{(3)}$ was formulated to compare the estimated means of the Gaussian effect size distribution to the average effect sizes. This comparison aimed at assessing whether the presence of effects in mean differences $\Delta_{\widehat{\mu}_d}$ deemed large enough to be considered meaningful, according to specified equivalence bounds Δ_{EQ}, can be rejected (Lakens et al., 2020). For this purpose, an equivalence test using the Two One-Sided Tests (TOST) procedure (Schuirmann, 1987) was conducted as implemented in the *TOSTER* R package (version 0.8.2, Lakens & Caldwell, 2023). This involves conducting two one-tailed tests against the upper and lower equivalence bounds, where both must yield significant results to claim statistical equivalence within the equivalence

range Δ_{EQ}. The preregistration specified that the *means* for both variables would be compared for the TOST procedure, implying a Student t-test for dependent samples. However, upon analysis of the assumptions for the Student t-test, it was found that the normality assumption was violated for both distributions. Parametric tests like the Student t-test are generally robust to violations of the normality assumption (Boneau, 1960; Knief & Forstmeier, 2021), so we proceeded as preregistered. However, to assess the robustness of the findings, we also conducted sensitivity analyses using the non-parametric Wilcoxon signed-rank test as part of the TOST procedure (see Appendix D). The choice of a dependent sample test was necessitated by the dependence between the pairs of samples originating from the same underlying data. The calculation of effect size measures is beneficial to obtain additional information about the magnitude of equivalence (not preregistered analyses). However, traditional effect size metrics, such as Cohen‘s d, prove inadequate within equivalence testing as they do not consider the equivalence range. Hence, the Proportional Distance PD, explicitly designed for equivalence testing, was utilised as an effect size metric (see Martinez Gutierrez & Cribbie, 2023). The PD quantifies the proportional distance from the observed effect ($\Delta_{\widehat{\mu}_d}$) to the bound of the equivalence range that is the same sign as the observed effect.

Smallest Effect Size of Interest

For all four hypotheses, the smallest effect size of interest (SESOI) was established based on effect sizes that can be 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, three simulation-based ($\mathcal{H}^{(1)}$, $\mathcal{H}^{(2)}$, $\mathcal{H}^{(4)}$) and one analytical ($\mathcal{H}^{(3)}$) sensitivity power analysis were conducted to determine for which effect sizes we have at least 80% power ($1 - \beta = 0.8$) to detect, taking into account the predetermined sample size and fixed type I error rate $\alpha = .05$ (details see *Appendix B*). We specified the SESOI for $\mathcal{H}^{(3)}$ in raw units and all other SESOIs in odds ratios. The SESOI for the equivalence hypothesis defines the equivalence bounds for the TOST procedure, $\Delta_{EQ} = \{-0.17, 0.17\}$. *Table 1* summarises all four SESOIs of the hypotheses.

Table 1

SESOIs of the Four Hypotheses

Hypothesis	SESOI	Unit
$\mathcal{H}^{(\mathrm{i})}$	1.28	*OR*
$\mathcal{H}^{(\mathrm{ii})}$	0.59	*OR*
$\mathcal{H}^{(\mathrm{iii})}$	0.17	raw unit
$\mathcal{H}^{(\mathrm{iv})}$	1.28	*OR*

Note. Except for $\mathcal{H}^{(3)}$ all SESOIs are defined in terms of odds ratios (OR). The SESOI of $\mathcal{H}^{(3)}$ is defined in raw units.

Confirmatory Results

As an initial step, the assumptions made to determine the SESOIs for the four hypotheses were assessed. In the simulation-based sensitivity power analyses aimed at determining the SESOIs (see Appendix B), 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 much lower estimated dispersion in comparison to the simulated conditions indicates the variability in ω_{PBS} was much higher that expected. In general, the lower the dispersion parameter ϕ for a fixed mean μ, the higher the variance (Ferrari & Cribari-Neto, 2004), as $\mathbb{V}[\omega_{\mathrm{PBS}}] = \mu \cdot (1 - \mu)/(1 + \phi)$. 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 that were determined from the simulation-based sensitivity power analyses in the subsequent analyses.

Regarding hypothesis $\mathcal{H}^{(1)}$, panel A of *Figure* 3 depicts the relationship between the estimated publication bias parameter $\hat{\omega}_{\mathrm{PBS}}$ and the Fisher z-transformed Spearman correlation coefficients z_{r_s} of the correlation of effect size and sample size in each meta-analysis. The observed slope was positive in the direction of the hypothesis and statistically significant $OR = 2.22$, 95% CI [1.38, Inf], $SE = 0.29$, $z = 2.74$, $p = .003$. Lower values for z_{r_s} were significantly associated with lower publication bias parameter values $\hat{\omega}_{\mathrm{PBS}}$. Additionally, to enhance the interpretability of the regression

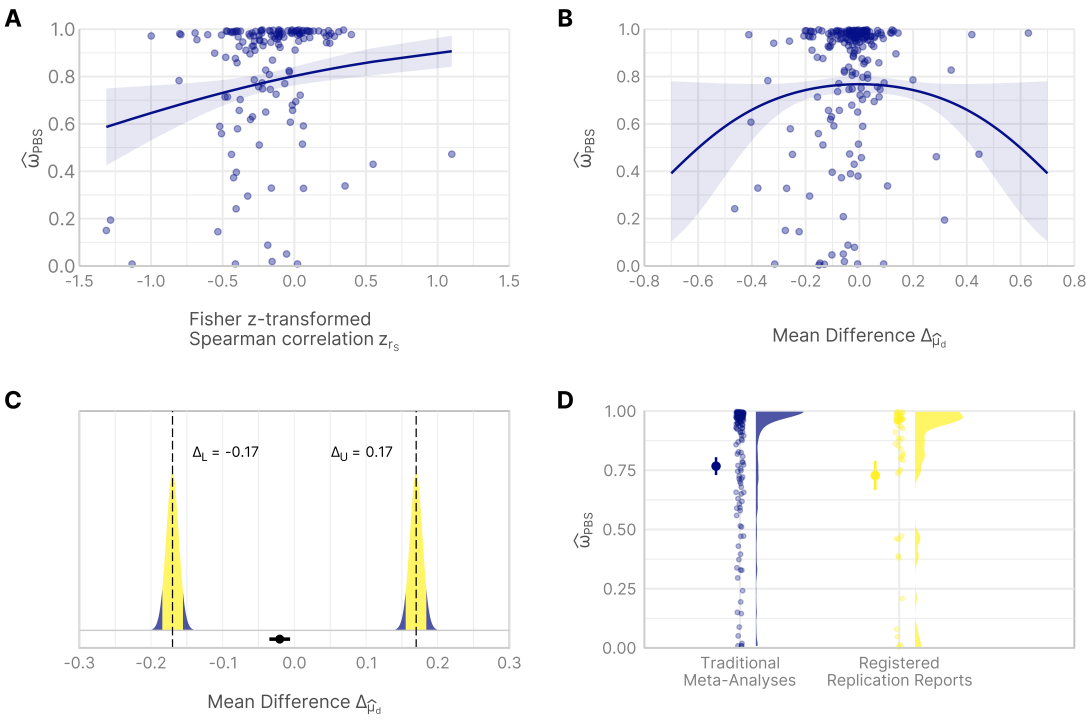
slope, we refitted the model with standardised values of z_{r_s} and computed the average marginal effects using the *marginaleffects* package (version 0.18.0, Arel-Bundock, 2024). On average, for every standard deviation increase in the Fisher z-transformed correlation coefficient, $SD(z_{r_s}) = 0.31$, the model only predicted an increase of 4.35% in the publication bias parameter $\hat{\omega}_{\text{PBS}}$. In line with this, the general explanative power of the model as determined by the pseudo-R^2 (Ferrari & Cribari-Neto, 2004) was low, pseudo-$R^2 = 0.05$. Thus, only 5% of the variance in ω_{PBS} could be explained by the variance of z_{r_s}.

Concerning $\mathcal{H}^{(2)}$, panel B of *Figure* 3 depicts the relationship between the 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, $p = .035$. Again, we calculated the average marginal effect for improved interpretability. On average, for every standard deviation increase in $\Delta_{\hat{\mu}_d}$, $SD(\Delta_{\hat{\mu}_d}) = 0.13$, the model only predicted a decrease of -0.09% in the publication bias parameter $\hat{\omega}_{\text{PBS}}$. The overall explained variation of ω_{PBS} by $\Delta_{\hat{\mu}_d}$ was low, pseudo-$R^2 = 0.03$.

Regarding hypothesis $\mathcal{H}^{(3)}$, panel C of *Figure* 3 illustrates the mean difference $\Delta_{\hat{\mu}_d}$ 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 90% confidence interval. Additionally, the null t-distributions of the TOSTs 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 falling inside the equivalence range in panel C of *Figure* 3. We additionally conducted an exploratory null hypothesis significance test to test the point hypothesis that the true mean difference of Δ_{μ_d} is exactly zero (not preregistered). The mean difference significantly deviated from zero MD = -0.02, 90% CI [-0.03, -0.01], $t(56)$ = -2.36, $SE = 0.01$, $p = 0.022$. Again, this is also illustrated in the *Figure* 3, as the 90% confidence interval does not contain zero. Regarding the additional non-preregistered analyses on the magnitude of equivalence, the proportional distance was PD = -0.12,

Figure 3

Visual Summary of the Confirmatory Results from the Four Hypotheses



Note. **A**. Scatter plot of the Fisher z-transformed correlation coefficients of the effect size sample size correlation predicting estimated publication bias parameters ω_{PBS}. Fitted line and ribbon represents the predicted values and 95% confidence interval from the beta regression model. $n = 150$. **B**. Estimated publication bias parameter ω_{PBS} predicted by the difference Δ_{μ_d} between mean parameter of the effect size distribution of SPEEC μ_d and average effect size $\hat{\delta}$. Fitted line and ribbon represents the predicted values and 95% confidence interval from the beta regression model. $n = 207$. **C**. Mean difference between the between mean parameter of the effect size distribution of SPEEC μ_d and average effect size $\hat{\delta}$ with 90% CI, compared to null t-distribution for the equivalence bounds. $n = 57$. **D**. Comparison of estimated publication bias parameter ω_{PBS} distributions between classical meta-analyses and registered replication reports. The point interval represents the predicted values and 95% confidence interval from the beta regression model. $n = 207$.

95% bootstrapped bias corrected and accelerated (BCa) *CI* [-0.205, -0.009]. This indicates that the observed mean difference $\Delta_{\hat{\mu}_d}$ is considerably distant from the lower equivalence bound (i.e., 12.00% of the distance away from 0 to the lower bound). Put differently, the observed mean difference could have been 8.3 times larger to reach the lower equivalence bound.

Finally, regarding hypothesis $\mathcal{H}^{(4)}$, panel D of *Figure* 3 depicts a comparison of the distributions of the estimated publication bias parameter $\hat{\omega}_{\text{PBS}}$ between the classical meta-analyses and the RRRs. The figure illustrates that there are high-density regions in the distribution of $\hat{\omega}_{\text{PBS}}$ close to one for both classical meta-analyses and RRRs. Moreover, the estimated publication bias parameter values below this high-

density region are more uniformly distributed for the classical meta-analyses than the RRRs. In the distribution of $\widehat{\omega}_{\text{PBS}}$ for RRRs, there are notable outliers with predicted values for $\widehat{\omega}_{\text{PBS}} < 0.5$. Already descriptively, contrary to our expectation that the estimated publication bias parameters for RRRs would be greater (i.e., lower publication bias) than for classical meta-analyses (MA), the mean of the estimated publication bias values ω_{PBS} of the regular meta-analysis subset is greater than the mean of the RRR subset ($M_{\text{MA}} = 0.82$; $M_{\text{RRR}} = 0.79$). In line with this, the slope of the beta regression was non-significant, $OR = 0.81$, 95% CI [0.61, Inf], $SE = 0.18$, z = -1.17, $p = .879$. Once more, we computed the average marginal effect to examine how the estimated publication bias parameter $\widehat{\omega}_{\text{PBS}}$ changes with the discrete shift from the reference level (RRR) to classical meta-analysis. The model predicted a change of -3.93% in $\widehat{\omega}_{\text{PBS}}$ in the opposite direction of the hypothesis.

Intermediate Discussion

In the present study we assessed the introduced SPEEC method in a proof of concept using secondary empirical meta-analytical data. We derived four hypotheses, which should be corroborated by the empirical data, if the method works in principle.

Regarding the results of the hypotheses, it was observed that the empirical data was more extreme under the null hypotheses than the prespecified type I error rate of $\alpha = .05$ for $\mathcal{H}^{(1)}$, $\mathcal{H}^{(2)}$ and $\mathcal{H}^{(4)}$, leading us to reject the null hypotheses for these predictions from a statistical viewpoint. However, it is important to evaluate these results not only in terms of their statistical significance but also in terms of their practical significance by means of the magnitude of the observed effects (LeCroy & Krysik, 2007; Shaver, 1993). In this regard, we found that for $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$, both the explained variance and the average marginal effects were low. This indicates both the effect size-sample size correlation as an alternative indicator of publication bias and the difference Δ_{μ_d} between the average effect size and the estimated effect size mean parameter of the SPEEC method only a weak magnitude of effect on the publication bias parameter ω_{PBS}. Regarding $\mathcal{H}^{(3)}$, the significance for both the equivalence test and the null hypothesis significance test indicated that although the point null hypothesis ($\Delta\mu_d = 0$) can be rejected, the equivalence test indicated the difference was too small to be considered meaningful according to the equivalence range. This means that the estimated mean parameter of the effect size μ_d from the SPEEC method and the average

effect size can be considered equivalent within the equivalence range for the subset of RRRs. In fact, the overall magnitude of the effect for the equivalence test can be considered large in terms of the proportional distance PD results. Most importantly, however, we failed to reject the null hypothesis for $\mathcal{H}^{(4)}$. That is, no evidence was found that the publication bias parameter for RRRs would be greater in comparison to classical meta-analyses, or in other words, that the probability for selection of non-significant studies in comparison to significant studies would be greater for RRRs in comparison to classical meta-analyses. Upon closer examination of the distribution of $\hat{\omega}_{\text{PBS}}$ for the RRR subgroup, it was observed that while most predictions for ω_{PBS} were close to one, there were outliers where the predictions for $\hat{\omega}_{\text{PBS}}$ approached zero. This contradicts our expectations because, as argued in the hypothesis, because puublication bias should be absent in RRRs.

Overall, the absence of evidence regarding $\mathcal{H}^{(4)}$ and the weak magnitude of the effect found for $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ point to potential problems within the SPEEC method itself or within the parameter optimization process using differential evolution. These findings underscore the need for additional exploratory analyses aimed at diagnosing and addressing potential problems within the parameter estimation in SPEEC.

Diagnostic Evaluation of Parameter Estimation in SPEEC

As this study relies on empirical data to 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 RRRs. Thus, the four distributional parameters (μ_d, σ_d^2, μ_n, ϕ_n) within the SPEEC method cannot be biased due to publication bias (especially the mean and variance of the effect size distribution). Leveraging this fact, we can use the subset of the RRRs to diagnose the parameter estimation within SPEEC. More specifically, we can use maximum likelihood estimation (MLE) to estimate the distributional parameters and subsequently compare them with the corresponding values estimated by the SPEEC method, anticipating approximate equivalence between the two approaches. This part of the analysis was not preregistered and conceived after the confirmatory analyses were conducted. Based on this comparative approach between MLE and SPEEC, we formulated multiple diagnostic questions to assess the 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. Is the discrepancy between SPEEC and MLE in the distributional parameters associated with the publication bias parameter ω_{PBS}?

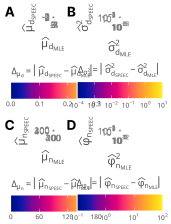
Additionally, we were interested in whether the discrepancy between SPEEC and ML of the distributional parameters and the publication bias parameter ω_{PBS} are associated with potentially other relevant factors that may increase the uncertainty of the parameter estimation. Accordingly, we derived two additional questions:

4. Does the discrepancy between SPEEC and MLE estimates of the distributional parameters correlate with the sample size of the RRRs (k)?
5. Is the discrepancy between SPEEC and MLE in the distributional parameters and the publication bias parameter ω_{PBS} associated with effect size heterogeneity?

The estimates of the MLE for the distributional parameters were obtained using the Nelder-Mead optimizer (Nelder & Mead, 1965). Additionally, the mean and median discrepancy between SPEEC and MLE were calculated to descriptively assess the average difference between the two estimation methods. To estimate the extent to which the true effect sizes vary within an RRR, we used the standard deviation τ of the effect sizes to measure between-study heterogeneity. Heterogeneity was estimated using the DerSimonian-Laird estimator (DerSimonian & Laird, 1986) in the *metafor* package (version 4.6.0, Viechtbauer, 2024).

Figure 4

Divergence Between the Estimated Distributional Parameters of SPEEC and MLE



Note. Comparison of the four estimated distributional parameters between SPEEC and Maximum Likelihood Estimation (MLE). Diagonal line represents perfect match between both methods. Axes of B and D are $\log_{10}$-scaled. Colorbar indicates the absolute divergences between MLE and SPEEC for each parameter.

Regarding the first question, *Figure* 4 provides a visual summary of the analysis comparing the distributional parameters estimated using the SPEEC method against

those estimated using Maximum Likelihood Estimation (MLE). The diagonal line signifies perfect alignment between MLE and SPEEC. Values below the diagonal indicate higher values for MLE than SPEEC, while values above the diagonal indicate the opposite. Panel A of *Figure* 4 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 μ_d, $M(\Delta_{\widehat{\mu}_d})$ = -0.03, $Mdn(\Delta_{\widehat{\mu}_d})$ = -0.02. According to the equivalence test of $\mathcal{H}^{(3)}$, this discrepancy can be deemed negligible. However, the other panels indicate contrasting outcomes. In Panel B, a systematic discrepancy in the estimation of the variance parameter of the effect size distribution σ_d^2 between SPEEC and MLE can be observed, $M(\Delta_{\widehat{\sigma}_d^2}) = 0.59$, $Mdn(\Delta_{\widehat{\sigma}_d^2}) = 0.11$. Descriptively, this suggests that, on average, the variance was greater in the SPEEC approach compared to MLE. Furthermore, this discrepancy increases in a nonlinear trend and displays substantial heteroscedasticity with rising variance estimates from the MLE approach. The greater the estimated variance from MLE, the greater the deviation of SPEEC from MLE, and the larger the variability in the estimation of SPEEC. Similarly, Panel C also illustrates a systematic overestimation of the mean parameter μ_n of the Negative-Binomial sample size distribution by SPEEC in comparison to MLE ($M(\Delta_{\widehat{\mu}_n}) = 46.96$, $Mdn(\Delta_{\widehat{\mu}_n}) = 3.52$), which again increases in a heteroscedastic exponential-appearing trend. Lastly, Panel D shows that SPEEC generally underestimated the dispersion parameter ϕ_n of the sample size distribution in comparison to the MLE estimate ($M(\Delta_{\widehat{\phi}_n})$ = -2.02, $Mdn(\Delta_{\widehat{\phi}_n})$ = -0.19) and furthermore indicates a systematic relationship in the discrepancy between the two approaches. Lastly, Panel D illustrates that SPEEC tends to underestimate the dispersion parameter ϕ_n of the sample size distribution compared to the MLE ($M(\Delta_{\widehat{\phi}_n})$ = -2.02, $Mdn(\Delta_{\widehat{\phi}_n})$ = -0.19) and also furthermore indicates a systematic relationship in the discrepancy between the two approaches.

To address the remaining diagnostic questions, we conducted a pairwise correlational analysis using the Pearson correlation coefficient between the absolute differences of the parameter estimates derived from the two estimation methods, the publication bias parameter $\widehat{\omega}_{\text{PBS}}$, the number of replication studies k within each RRR and the between-study heterogeneity $\widehat{\tau}$. Concerning the multiple testing performed, the question of adjusting p-values in exploratory studies for multiple testing is an ongoing debate (Rubin, 2017). For reasons of transparency, both unadjusted and adjusted p-values are therefore reported following the method of Benjamini and Hochberg (1995).

Table 2 summarises the results of the correlational analysis. Regarding the parameters of the Gaussian effect size distribution (μ_d, σ_d^2), a strong positive correlation was observed between the absolute difference in the mean parameter estimates and the variance parameter estimates obtained from MLE and SPEEC. This indicates that as the absolute differences between SPEEC and MLE increased for the mean parameter $|\Delta_{\widehat{\mu}_d}|$, the absolute discrepancy also increased for the variance parameter $|\Delta_{\widehat{\sigma}_d^2}|$ of the effect size distribution. Furthermore, strong negative correlations were found between the publication bias parameter ω_{PBS} and the discrepancy between MLE and SPEEC estimates of the mean and variance parameters of the effect size distribution. More specifically, an increase in the absolute discrepancy between both estimation methods increased for both the mean ($|\Delta_{\widehat{\mu}_d}|$) and variance ($|\Delta_{\widehat{\sigma}_d^2}|$) was associated with a decrease in the publication bias parameter ω_{PBS}, signifying more severe predicted publication bias. Notably, the total number of primary replications k was not significantly associated with the divergence of MLE and SPEEC of any distributional parameter or the publication bias parameter ω_{PBS}. Moreover, regarding the estimated between-study heterogeneity parameter $\widehat{\tau}$, strong positive correlations were observed for divergences between MLE and SPEEC in both the effect size mean and variance parameter. Thus, an increase in the effect size heterogeneity of the RRRs was associated with an increase in the divergence between ML and SPEEC for both parameters of the effect size distribution. Lastly, the estimated between-study heterogeneity parameter $\widehat{\tau}$ was also negatively associated with the estimated publication bias parameter $\widehat{\omega}_{\text{PBS}}$, such that an increase in heterogeneity predicted a decrease in the publication bias parameter (i.e., more severe publication bias). Notably, this correlation was only significant for the unadjusted p-value.

Table 2

Pairwise Correlations Between Absolute Divergences in Distributional Parameters of SPEEC and MLE, Publication Bias Parameter, Meta-Analysis Size and Heterogeneity

Comparison	r (95% CI)	p	p_{adj}
τ - ω_{PBS}	**-0.30** [-0.52, -0.04]	.024	.085
τ - k	0.22 [-0.04, 0.46]	.097	.292
τ - $\lvert\Delta_{\mu_d}\rvert$	**0.66** [0.48, 0.79]	<.001	<.001
τ - $\lvert\Delta_{\sigma_d^2}\rvert$	**0.51** [0.29, 0.68]	<.001	<.001
τ - $\lvert\Delta_{\phi_n}\rvert$	-0.02 [-0.28, 0.24]	.863	.906
τ - $\lvert\Delta_{\mu_n}\rvert$	-0.06 [-0.32, 0.20]	.635	.883
ω_{PBS} - k	-0.14 [-0.39, 0.13]	.300	.572
ω_{PBS} - $\lvert\Delta_{\mu_d}\rvert$	**-0.44** [-0.63, -0.20]	<.001	.003
ω_{PBS} - $\lvert\Delta_{\sigma_d^2}\rvert$	**-0.37** [-0.57, -0.12]	.005	.021
ω_{PBS} - $\lvert\Delta_{\phi_n}\rvert$	0.08 [-0.18, 0.34]	.533	.883
ω_{PBS} - $\lvert\Delta_{\mu_n}\rvert$	0.03 [-0.23, 0.29]	.799	.883
k - $\lvert\Delta_{\mu_d}\rvert$	0.04 [-0.23, 0.29]	.793	.883
k - $\lvert\Delta_{\sigma_d^2}\rvert$	0.15 [-0.12, 0.39]	.280	.572
k - $\lvert\Delta_{\phi_n}\rvert$	-0.17 [-0.41, 0.09]	.199	.523
k - $\lvert\Delta_{\mu_n}\rvert$	0.00 [-0.26, 0.26]	.983	.983
$\lvert\Delta_{\mu_d}\rvert$ - $\lvert\Delta_{\sigma_d^2}\rvert$	**0.67** [0.49, 0.79]	<.001	<.001
$\lvert\Delta_{\mu_d}\rvert$ - $\lvert\Delta_{\phi_n}\rvert$	-0.06 [-0.31, 0.20]	.662	.883
$\lvert\Delta_{\mu_d}\rvert$ - $\lvert\Delta_{\mu_n}\rvert$	0.15 [-0.12, 0.39]	.275	.572
$\lvert\Delta_{\sigma_d^2}\rvert$ - $\lvert\Delta_{\phi_n}\rvert$	-0.05 [-0.31, 0.21]	.721	.883
$\lvert\Delta_{\sigma_d^2}\rvert$ - $\lvert\Delta_{\mu_n}\rvert$	0.04 [-0.22, 0.30]	.764	.883
$\lvert\Delta_{\phi_n}\rvert$ - $\lvert\Delta_{\mu_n}\rvert$	-0.05 [-0.31, 0.21]	.718	.883

Note. r: Pearson correlation coefficient, CI: confidence interval, p-values are adjusted on based on the correction by Benjamini and Hochberg (1995). Highlighted bold values are statistically significant according to the unadjusted p-values.

General Discussion

The objective of the present thesis was to introduce a flexible simulation-based framework to assess the extent of publication bias and estimate corrected effect sizes based on the joint distribution of effect size and sample size in the presence of publication bias (SPEEC). In a proof-of-concept study, the viability of the SPEEC method was evaluated based on four theoretically derived hypotheses, which should be corroborated by empirical data encompassing both classical meta-analyses and publication bias-free

registered replication reports (RRR) if the method works in principle. The confirmatory analyses revealed that, while three out of the four predictions were statistically significant, the effect sizes for these hypotheses were weak. Additionally, no evidence supported the fourth hypothesis, pointing to potential issues with the parameter estimation in the SPEEC method.

To further diagnose these potential parameter estimation challenges, five exploratory diagnostic questions were derived using the subset of publication bias-free RRRs to compare the distributional parameters estimated using SPEEC to those estimated using MLE. Additionally, it was also examined how these divergences between SPEEC and MLE are associated with the publication bias parameter ω_{PBS} and uncertainty-related factors such as the number of individual replication studies within each RRR k and between-study heterogeneity τ. Overall, this analysis aimed to quantify the extent and consistency of misestimation in SPEEC compared to MLE (question 1-3) and to identify potential root causes of the parameter estimation issues (question 4-5).

Extent and Consistency of Misestimation of SPEEC Parameters

Regarding the first three exploratory diagnostic questions, it was found that, except for the mean μ_d of the effect size distribution, there were systematic divergences between SPEEC and MLE in the parameter estimation. This misestimation of SPEEC from MLE was not constant across the parameter space of the MLE estimates, displaying nonlinear functional forms and substantial heteroscedasticity.

The correlational analyses indicated consistency in the misestimation of SPEEC compared to MLE for the mean $|\Delta_{\widehat{\mu}_d}|$ and variance parameter $|\Delta_{\widehat{\sigma}^2_d}|$ of the effect size distribution. In other words, an increase in the absolute divergence between SPEEC and MLE in one parameter was associated with an increase in the absolute divergence in the other parameter. The absence of any substantial (and significant) correlations between the divergences of SPEEC and MLE for effect size distribution parameters ($|\Delta_{\widehat{\mu}_d}|$, $|\Delta_{\widehat{\sigma}^2_d}|$) and sample size distribution parameters ($|\Delta_{\widehat{\phi}_n}|$, $|\Delta_{\widehat{\mu}_n}|$) suggests that there is no evidence of consistency between the misestimation of SPEEC from MLE regarding effect size and sample size parameters. However, there was a strong and statistically significant negative correlation between the parameters of the effect size distribution ($|\Delta_{\widehat{\mu}_d}|$, $|\Delta_{\widehat{\sigma}^2_d}|$) and the publication bias parameter $\widehat{\omega}_{\text{PBS}}$ such that larger

divergences between MLE and SPEEC of the distributional parameters were associated with lower predicted values of the publication bias parameter. Because, as argued previously, $\hat{\omega}_{\text{PBS}}$ should, in theory, equal exactly 1, this negative correlation can be interpreted as consistency in the misestimation of the effect size parameter of SPEEC from MLE and the misestimation of the publication bias parameter $\hat{\omega}_{\text{PBS}}$.

Which Factors Drive the Parameter Misestimation in SPEEC?

The systematic discrepancy between MLE and SPEEC for the distribution parameters, together with the consistency in the misestimation of the effect size parameters and the publication bias parameter, begs the question of what factors are responsible for the parameter misestimation of the SPEEC method.

In this regard, no evidence was found that the sample size k of the RRRs was associated with the misestimation of the distributional parameters of SPEEC from MLE. However, strong positive correlations were observed between the misestimation of SPEEC from MLE in the effect size distribution parameters ($|\Delta_{\widehat{\mu}d}|$, $|\Delta_{\widehat{\sigma}_d^2}|$) and between-study heterogeneity $\hat{\tau}$. Additionally, a medium negative correlation was observed between the publication bias parameter $\hat{\omega}_{\text{PBS}}$ and between-study heterogeneity $\hat{\tau}$. This indicates that increasing observed heterogeneity $\hat{\tau}$ was associated with a larger misestimation of the distributional parameter of SPEEC from MLE and a larger misestimation of the publication bias parameter $\hat{\omega}_{\text{PBS}}$. These results highlight the potential significance of between-study heterogeneity in influencing parameter misestimation.

This is plausible from the perspective that the simulation framework of SPEEC currently assumes a fixed-effect meta-analytical model, where the only source of effect size variation is sampling error. Both classical meta-analyses and RRRs of the empirical data used for the proof-of-concept displayed moderate to large levels of between-study heterogeneity (Linden & Hönekopp, 2021). Thus, the empirical data‘s additional level of effect size variation could not be adequately modeled in the SPEEC method due to sampling error alone. In particular, because of the definition of the loss function D_{KL} as a statistical distance between the kernel densities from the empirical and simulated data, parameters may be erroneously adjusted in the presence of heterogeneity to minimise the loss function.

This could be one potential explanation for the associated predicted decrease in $\hat{\omega}_{\text{PBS}}$ for high heterogeneity $\hat{\tau}$, as more severe publication bias implies that the non-

significant studies that are likely to have effect sizes close to zero are censored, thus artificially increasing the overall variance to account for heterogeneity in the minimisation of D_{KL}. Furthermore, the same line of reasoning could also account for the positive relationship between $|\Delta_{\widehat{\sigma}_d^2}|$ and σ_d^2, as the overestimation of σ_d^2 by SPEEC compared to MLE increases the overall variability of the effect size distribution to seemingly capture heterogeneity.

In summary, the challenges encountered by the current version of the SPEEC method, which are likely attributable to between-study heterogeneity, are parallel to those faced by other statistical techniques assuming a fixed-effect meta-analytical model, such as p-uniform and p-curve analysis (Simonsohn et al., 2014; Van Assen et al., 2015). These methods tend to perform poorly in the presence of substantial heterogeneity, resulting in an overestimation of the true effect size under such conditions (Carter et al., 2019; McShane et al., 2016; Van Aert et al., 2016). This aligns with the positive correlation observed between $|\Delta_{\widehat{\mu}_d}|$ and $\hat{\tau}$, suggesting that an increase in heterogeneity was associated with a greater misestimation of SPEEC from MLE for the mean parameter μ_d of the effect size distribution.

Limitations

Besides the previously discussed challenges in the parameter estimation of SPEEC, which are likely to result from heterogeneity, the present thesis has additional limitations that should be considered. These may be subdivided into objections against the SPEEC method itself and study design constraints.

The first major limitation of the SPEEC method is the lack of uncertainty quantification in estimating the parameters in SPEEC. This lack of quantification of parameter uncertainty undermines the ability to assess the confidence of these estimates accurately (Lele, 2020). As an illustration, one would have varying degrees of confidence in the estimated extent of publication bias $\widehat{\omega}_{\mathrm{PBS}}$ and the corrected meta-analytical effect size μ_d for a small-scale meta-analysis with a sample size of $k = 20$ in comparison to a large-scale meta-analysis with a sample size of $k = 500$. This inability can be considered inferior to other common methods to assess publication bias that yield confidence estimates.

Secondly, while an explicit generative publication bias model in SPEEC may be seen as an advancement compared to other methods lacking such assumptions, the

model's assumptions remain oversimplistic. Real-world scenarios involve more intricate data models and selection mechanisms. Studies often encompass multiple dependent effects of interest, with selection likely based on properties of these effects taken jointly (McShane et al., 2016). In addition, questionable research practices in existing scientific research can further complicate the data and selection models due to their interaction with publication bias (Friese & Frankenbach, 2020).

The constraints regarding the study design relate to the study‘s conception as proof of concept. Although the investigation of the theoretically derived hypotheses in the confirmatory analyses yielded important insights into potential parameter estimation challenges, and diagnostic exploratory findings underscored the potential significance of heterogeneity in explaining these challenges, the assessment was limited in its capacity to offer a holistic understanding of SPEEC‘s viability across various conditions. For this, large-scale simulation studies are necessary because the parameters of interest are known and manipulable (Morris et al., 2019).

Directions for Future Research

Considering the results of both the confirmatory and the exploratory diagnostic analyses, there is a call for further investigations in at least two directions. These encompass the further development of the SPEEC method and a comprehensive assessment of SPEEC within a simulation setting.

Regarding the former, considering the likely contribution of heterogeneity to the parameter misestimation in the investigated empirical data, as well as its common prevalence across meta-analyses (Erp et al., 2017; Higgins, 2008; Linden & Hönekopp, 2021), highlights the importance of incorporating a heterogeneity parameter in the simulation framework of SPEEC. Failure to account for heterogeneity could otherwise potentially jeopardize the validity of SPEEC in settings where heterogeneity is present. Additionally, as briefly touched upon in the introduction, there exist other factors that influence the joint distribution of effect size and sample size, including *sample size planning*, where researchers plan sample sizes based on predicated effect sizes (Linden et al., 2024) and questionable research practices such as *p*-hacking, which can interact with publication bias (Friese & Frankenbach, 2020). Incorporating these factors into the simulation framework in future research could lead to more accurate modeling of publication bias.

A comprehensive assessment of SPEEC within a simulation setting should focus on evaluating the accuracy of parameter estimation of SPEEC under various conditions commonly encountered in real-world meta-analytical data and that have been considered in previous simulation studies that assessed other publication bias detection methods. Key factors to consider include the extent of publication bias (ω_{PBS}), true effect size (μ_d), amount of heterogeneity (τ), the number of primary studies k within the meta-analysis and questionable research practices such as p-hacking (Carter et al., 2019; McShane et al., 2016; Renkewitz & Keiner, 2019; Schneck, 2017; Van Aert et al., 2019).

Conclusion

In conclusion, this thesis introduced a novel and flexible framework for assessing publication bias and correcting meta-analytic effect sizes in the presence of publication bias based on the joint distribution of effect size and sample size. This framework includes explicit assumptions about the generative processes of publication bias and is adaptable to incorporating various factors that may be valuable for modeling and assessing publication bias in meta-analyses. However, the results of the proof-of-concept study have demonstrated that adjustments to the simulation framework to incorporate heterogeneity are required before using the SPEEC method in real-world applications. Finally, we seek to stimulate further research to advance the development of SPEEC by providing concrete recommendations on potentially valuable adaptations of SPEEC and factors to consider in a comprehensive evaluation of SPEEC.

References

Allaire, J. J., Teague, C., Scheidegger, C., Xie, Y., & Dervieux, C. (2024). *Quarto* (Version 1.4.551) [Computer software]. https://doi.org/10.5281/zenodo.5960048

Arel-Bundock, V. (2024). *Marginaleffects: Predictions, comparisons, slopes, marginal means, and hypothesis tests* (Version 0.18.0.9) [Computer software]. 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 H. Cooper & L. V. Hedges (Eds.), *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://doi.org/https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, *57*(1), 49–64. https://doi.org/10.1037/h0041412

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

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

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

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, *1*(4), 300–307.

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

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

Curran, P. J. (2009). The seemingly quixotic pursuit of a cumulative psychological science: Introduction to the special issue. *Psychological Methods*, *14*(2), 77–80. https://doi.org/10.1037/a0015972

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *JAMA*, *263*(10), 1385–1389. https://doi.org/10.1001/jama.1990.03440100097014

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++* (Version 0.1.7) [Computer software]. 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

Erp, S. J. van, 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

Etz, A. (2018). *Technical notes on Kullback-Leibler divergence.* PsyArxiv Preprints. https://doi.org/10.31234/osf.io/5vhzu

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

Friese, M., & Frankenbach, J. (2020). p-hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, *25*(4), 456–471. https://doi.org/10.1037/met0000246

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., Furukawa, T., & Ebert, D. (2021). *Doing meta-analysis with R: A hands-on guide* (1st ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781003107347

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*(1), 61–85. https://doi.org/10.3102/10769986009001061

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, *7*(2), 246–255. https://doi.org/10.1214/ss/1177011364

Higgins, J. P. T. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, *37*(5), 1158–1160. https://doi.org/10.1093/ije/dyn204

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, *3*(1), 109–117. https://doi.org/10.1214/ss/1177013012

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.

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

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

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

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, *53*(6), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

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. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

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* (Version 0.8.0) [Computer software]. 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

LeCroy, C. W., & Krysik, J. (2007). Understanding and interpreting effect size measures. *Social Work Research*, *31*(4), 243–248. https://www.jstor.org/stable/42659906

Lele, S. R. (2020). How should we quantify uncertainty in statistical inference? *Frontiers in Ecology and Evolution*, *8*. https://doi.org/10.3389/fevo.2020.00035

Levene, H. (1960). Robust tests for equality of variances. In L. Olkin (Ed.), *Contributions to probability and statistics: Essays in honor of harold hotelling* (pp. 278–292).

Stanford University Press.

Light, R., & Pillemer, D. (1984). *Summing up: The science of reviewing research.* Harvard University Press. https://doi.org/10.2307/j.ctvk12px9

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. H., Pollet, T. V., & Hönekopp, J. (2024). Publication bias in psychology: A closer look at the correlation between sample size and effect size. *Plos One*, *19*(2), e0297075.

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, *45*(1), 503–528. https://doi.org/10.1007/BF01589116

Lloyd-Smith, J. O. (2007). Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PLoS ONE*, *2*(2), e180. https://doi.org/10.1371/journal.pone.0000180

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

Martinez Gutierrez, N., & Cribbie, R. (2023). Effect sizes for equivalence testing: Incorporating the equivalence interval. *Methods in Psychology*, *9*, 100127. https://doi.org/10.1016/j.metip.2023.100127

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*(5), 730–749. https://doi.org/10.1177/1745691616662243

Merkel, D. (2014). Docker: Lightweight linux containers for consistent development and deployment. *Linux Journal*, *2014*(239), 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., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). *Sustainable data analysis with*

snakemake [version 2; peer review: 2 approved] (10). F1000Research. https://doi.org/10.12688/f1000research.29032.2

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

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

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.4.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

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

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments.* John Wiley & Sons. https://doi.org/10.1002/0470870168

Rubin, M. (2017). Do p values lose their meaning in exploratory analyses? It depends

how you define the familywise error rate. *Review of General Psychology*, *21*(3), 269–275. https://doi.org/10.1037/gpr0000123

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

Schäfer, T., & Schwarz, M. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, 1–13. https://doi.org/10.3389/fpsyg.2019.00813

Schmidt, F. L. (1992). What do data really mean?: Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173–1181. https://doi.org/10.1037/0003-066X.47.10.1173

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115–129. https://doi.org/10.1037/1082-989X.1.2.115

Schnatz, J. L., Schultze, M., & Beitner, J. (2024). *Assessing publication bias in meta-analyses: A simulation-based estimation approach.* https://doi.org/10.17605/osf.io/87m9k

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

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

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3), 591–611. https://doi.org/10.2307/2333709

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education*, *61*(4), 293–316. https://www.jstor.org/stable/20152383

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(3), 683–690. https://doi.org/10.1111/j.2517-6161.1991.

tb01857.x

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://doi.org/10.1177/1745691614543974

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology. General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

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. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

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

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

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

Ushey, K., & Wickham, H. (2024). *renv: Project environments* (Version 1.0.7) [Computer software]. https://cran.r-project.org/web/packages/renv/index.html

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 Aert, R. C. M., Wicherts, J. M., & Van Assen, M. A. L. M. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLOS ONE*, *14*(4), e0215052. https://doi.org/10.1371/journal.pone.0215052

Van Assen, M. A. L. M., Van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293–309. https://doi.org/10.1037/met0000025

Van den Akker, O., Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P., Hall, A., Kosie, J., Kruse, E., Olsen, J., Ritchie, S., Valentine, K., Van't Veer, A., & Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, *5*(2625). https://doi.org/10.15626/MP.2020.2625

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.

Viechtbauer, W. (2024). *metafor: Meta-analysis package for R* (Version 4.6-0) [Computer software]. https://cran.r-project.org/web/packages/metafor/index.html

Wand, M. P. (1994). Fast computation of multivariate kernel estimators. *Journal of Computational and Graphical Statistics*, *3*(4), 433–445. https://doi.org/10.2307/1390904

Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing* (1st ed.). Chapman; Hall/CRC. https://doi.org/10.1201/b14876

Wand, M. P., Moler, C., & Ripley, B. (2023). *KernSmooth: Functions for kernel*

smoothing supporting Wand & Jones (1995) (Version 2.23-22) [Computer software]. https://cran.r-project.org/web/packages/KernSmooth/index.html

Zeileis, A., Cribari-Neto, F., Gruen, B., Kosmidis, I., Simas, A. B. S., & Rocha, A. V. (2021). *betareg: Beta regression* (Version 3.1-4) [Computer software]. https://cran.r-project.org/web/packages/betareg/index.html

Appendix

Appendix A: Research Transparency Statement

The present thesis project was aimed to be as transparent and reproducible as possible. The preregistration of the study, all data, code, and supplementary files needed to reproduce the results of this study are openly available in a repository on the Open Science Framework via this link https://osf.io/87m9k/ (see, Schnatz et al., 2024) and on GitHub (https://github.com/jlschnatz/bachelor-thesis). To enhance the reproducibility of this project, all analyses, figures, and the thesis manuscript are generated within a containerized software environment using *Docker* (Merkel, 2014). The workflow for the analyses was managed using *Snakemake* (Mölder et al., 2021), and the R packages dependencies and version management used for the statistical analyses were managed using *renv* (Ushey & Wickham, 2024). The thesis manuscript is dynamically generated and reproducible using the *Quarto* publishing system (Allaire et al., 2024).

Appendix B: Power Analyses determining the SESOIs

The simulated-based sensitivity power analysis targeted a statistical power of 80% with a fixed significance level of $\alpha = .05$. The simulated sample sizes were specified as follows:

- $\mathcal{H}^{(1)}$: $n = 150$ (only classical meta-analyses)
- $\mathcal{H}^{(2)}$ and $\mathcal{H}^{(4)}$: $n = 207$ (both classical meta-analyses and registered replication reports)
- $\mathcal{H}^{(3)}$: $n = 57$ (only registered replication reports)

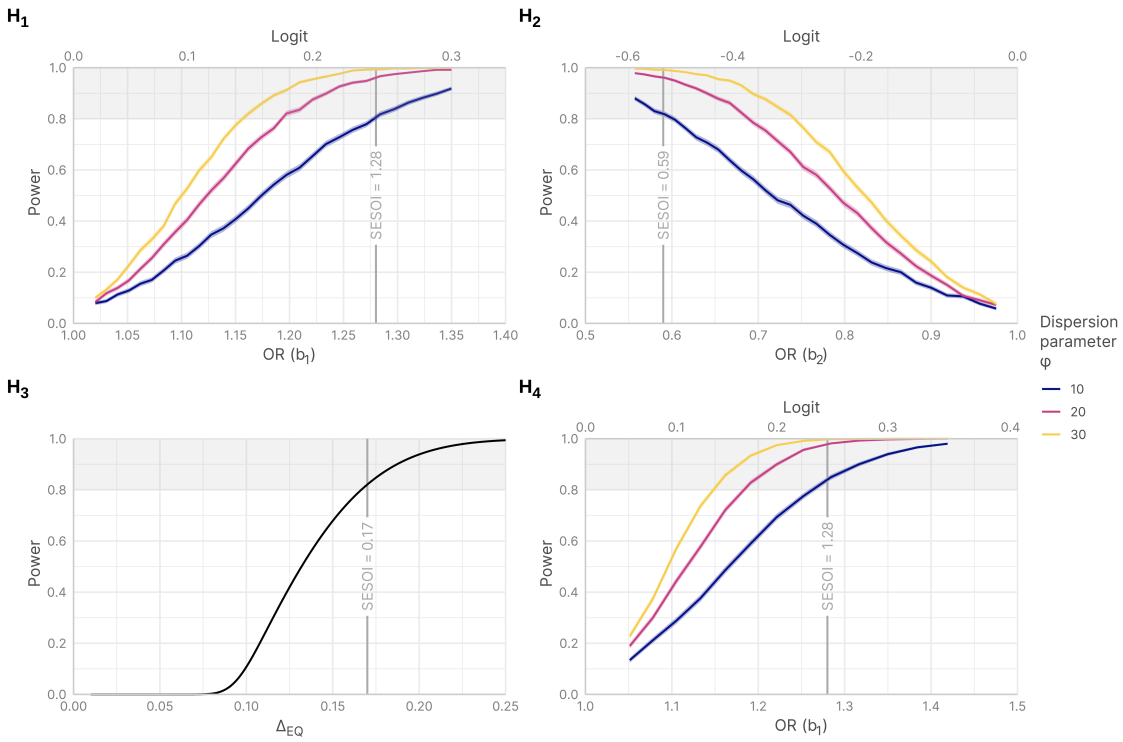
The distributional assumptions were specified as follows:

$$\mathcal{H}^{(1)}: \quad z_{r_S} \sim \mathcal{N}(\mu = -0.1, \sigma = 0.5)$$
$$\mathcal{H}^{(2)} \text{ and } \mathcal{H}^{(3)}: \quad \Delta_{\mu_d} \sim \mathcal{N}\Big(\mu = 0, \sigma_{diff} = \sqrt{0.3^2 + 0.3^2}\Big)$$

For hypothesis four, the proportions of the categorical predictor of the research synthesis type (classical meta-analysis MA, registered replication reports RRR) were chosen according to the actual proportions of the data ($n_{MA} = 150, n_{RRR} = 57$). For all simulations-based sensitivity power analyses ($\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \mathcal{H}^{(4)}$), the number of simulations was set to $n_{iter} = 5000$ and the average power was calculated across all simulations. More over, the the beta-regressions on ω_{PBS} in $\mathcal{H}^{(\text{i})}$, $\mathcal{H}^{(2)}$, and $\mathcal{H}^{(4)}$ involved simulations for different dispersion parameter conditions $\phi = \{10, 20, 30\}$, as lower dispersion parameters result in reduced test power. This is because the dispersion parameter directly influences the variance for a fixed mean. 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 A1

Power Curves of the Sensitivity Power Analyses Determining the SESOIs



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

Appendix C: MLE Extrema Distributional Parameters

Table C1

Maximum Likelihood Estimated Parameters for the Distribution of Effect Size and Sample Size

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 (Nelder & Mead, 1965).

Appendix D: Sensitivity Analyses of Equivalence Test

The homogeneity of variance assumption between the two population means was assessed using a Levene test (Levene, 1960). The outcome was non-significant, indicating a failure to reject the null hypothesis of equal variances ($F = 0.001$, $df_1 = 1$, $df_2 = 112$, $p = 0.972$). Subsequently, the assumption of normality was examined both inferentially and visually through the Shapiro-Wilk test (Shapiro & Wilk, 1965) and quantile-quantile plots, respectively. The inferential outcomes of the Shapiro test revealed deviations from normality for both distributions (see *Table D1*). This finding was supported by the quantile-quantile plot depicted in Figure D1, where the empirical quantiles did not align with the theoretical quantiles expected under a normal distribution.

Table D1

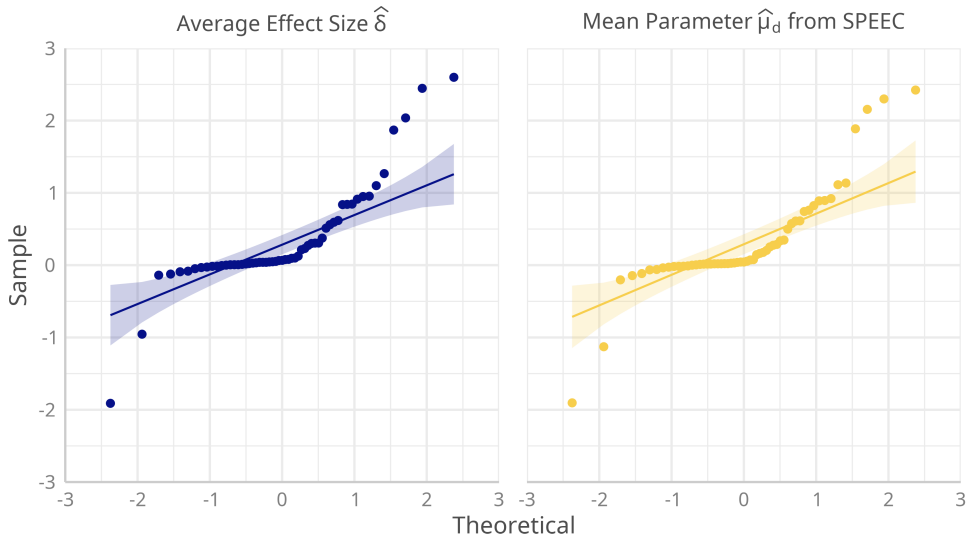
Shapiro-Wilk Test Testing Normality for $\mathcal{H}^{(3)}$

Parameter	W	p
Average effect size $\hat{\delta}$	0.80	$< .001$
Mean parameter $\hat{\mu}_d$ from SPEEC	0.81	$< .001$

Note. W: Shapiro-Wilk test statistic

Figure D1

QQ-Plot Assessing the Normality Assumption for $\mathcal{H}^{(3)}$



Note. Empirical observed quantiles for the average effect size $\hat{\delta}$ and mean parameter of the effect size distribution from SPEEC $\Delta_{\widehat{\mu}_d}$ as a function of theoretical quantiles expected under the normal distribution. Ribbons represent the 95% confidence interval.

Table D2

Sensitivity Analyses for the equivalence test in $\mathcal{H}^{(3)}$ and NHST: Wilxocon Signed-Rank Test

Type	Hypothesis	T^+	μ_{T^+}	σ_{T^+}	z	p
NHST	$\Delta = 0$	418	826.5	125.86	-3.25	.001
TOST	$\Delta < \Delta_L$	1650	826.5	125.86	6.55	$<$ .001
TOST	$\Delta > \Delta_L$	2	826.5	125.86	-6.55	$<$ .001

Note. Continuity correction applied. Approximate Gaussian null distribution used. T^+: positive rank sum (test statistic), μ_{T^+}: mean of positive rank sum under $\mathcal{H}_0$, σ_{T^+}: variance of positive rank sum under $\mathcal{H}_0$

Appendix E: Regression Tables of Confirmatory Analyses

Table E1

Beta Regression Results for $\mathcal{H}^{(1)}$

Term	Estimate	CI (95%)	SE	z	p
Mean model component: μ					
Intercept	4.05^a	[3.16, 5.19]	0.13	11.07	$< .001$
z_{r_s}	2.22^a	$[1.38, \text{Inf}]^c$	0.29	2.74	.003
Precision model component: ϕ					
Intercept	5.84^b	[3.95, 8.63]	0.20	8.85	$< .001$

Note. CI: Confidence interval, SE: standard error, LL = 132.03, MAE = 0.21, AIC = -258.06, BIC = -249.03, R^2 = 0.051

[a] Odds ratio

[b] Raw values

[c] One-sided confidence interval in the direction of the hypothesis

Table E2

Beta Regression Results for $\mathcal{H}^{(2)}$

Term	Estimate	CI (95%)	SE	z	p
Mean model component: μ					
Intercept	3.30^a	[2.71, 4.03]	0.10	11.79	$< .001$
$\Delta^2_{\mu_d}$	0.04^a	$[0.00, 0.73]^c$	1.84	-1.81	.035
Precision model component: ϕ					
Intercept	4.85^b	[3.63, 6.47]	0.15	10.70	$< .001$

Note. CI: Confidence interval, SE: standard error, LL = 164.25, MAE = 0.23, AIC = -322.51, BIC = -312.51, R^2 = 0.026

[a] Odds ratio

[b] Raw values

[c] One-sided confidence interval in direction of the hypothesis

Table E3

Two One-Sided Tests Result using Welch´s Tests regarding $\mathcal{H}^{(3)}$

Type	t	SE	df	p
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 E4

Beta Regression Results for $\mathcal{H}^{(4)}$

Term	Estimate	CI (95%)	SE	z	p
Mean model component: μ					
Intercept	3.30^a	[2.67, 4.08]	0.11	11.07	$<$ .001
RRR	0.81^a	$[0.61, \text{Inf}]^c$	0.18	-1.17	.879
Precision model component: ϕ					
Intercept	4.79^b	[3.59, 6.38]	0.15	10.71	$<$ .001

Note. RRR: registered replication reports, CI: confidence interval. LL = 163.31, MAE = 0.23, AIC = -320.62, BIC = -310.62, R^2 = 0.011

[a] Odds ratio

[b] Raw values

[c] One-sided cnfidence interval in direction of the hypothesis

Statutory Declaration

I herewith declare that I have composed the present thesis myself and without use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The thesis in the same or similar form has not been submitted to any examination body and has not been published. This thesis was not yet, even in part, used in another examination or as a course performance.

Jan Luca Schnatz

Darmstadt, den 21. Mai 2024