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

May, 2024

Goethe-Universität Frankfurt am Main

Faculty 05: Psychology and Sports Sciences

Erstgutachter: Prof. Dr. Martin Schultze

Zweitgutachter: Julia Beitner M.Sc.

Abstract

Test

Table of contents

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 endeavor, the idea of *meta-analyses* - quantitavely synthesizing research findings examining the same phenomena - has been attributed a particularly important role in contributing to the cumulative advancement of knowledge (Schmidt, 1992, 1996). However, this premise rests on the fundamental assumption, that the published research in the scientific literature is representative for all conducted research (Rothstein et al., 2005; Song et al., 2010). Yet, scientists have been pointing out for over half a century that results of published studies differ systematically from those of unpublished studies (Bakan, 1966; Bozarth & Roberts, 1972; Smart, 1964; Sterling, 1959).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 (Dickersin, 1990; Dickersin & Min, 1993) and has be collectively known as *publication bias*. Especially in a publishing culture that incentivizes novelty and positive results, alongside the prevalent use of the null hypothesis significance testing (NHST) framework, 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 non-significant studies end up in the "file-drawer" and never get published (Rosenthal, 1979).

The Impact of Publication Bias

The ramifications of publication bias are severe, culminating in an inflation of meta-analytical effect sizes (Franco et al., 2014; Stanley et al., 2021), heightened false-positive rate (Ioannidis, 2005; Kicinski, 2014; Munafò & Flint, 2010) , thus increasing the risk of errorneous conclusions that may jeopardize the validity of 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 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 identified

as a major threat to replicable science (Munafò et al., 2017) and thus considered as a significant contributor to the replication crisis (Renkewitz & Keiner, 2019). The myriad issues associated with publication bias and its widespread impact have fuelled a great deal of research focusing on statistical methods to detect and address publication bias.

Methods to Assess Publication Bias and their Limitations

To this regard, 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 that are based on the relationship between effect size and sample size in meta-analyses and those that operate with *p*-values (Vevea et al., 2019).

The former class of methods, also coined small-study effects (Sterne et al., 2000), rely 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 less studies with low sample sizes and low effect sizes, because they are statistically nonsignificant. This results in an assymmetrical funnel plot and a correlation between the effect size and a measure of precision (e.g., standard error or sample size), that are then analyzed using regression 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 effect size sample size correlation (e.g., Kühberger et al., 2014).

The latter class of methods 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 (Assen et al., 2015) and *p*-curve analysis (Simonsohn et al., 2014). Publication bias selection models aim to directly characterize 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 abundace of statistical methods to assess publication bias, there are some justified critizisms that 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 may is not the only source that influence the effect size sample size distribution, as researcher may plan sample sizes before performing the study according to anticipated effect sizes (Linden et al., 2024; Schäfer & Schwarz, 2019), which may compromise the validity of the interpretation of such methods. More broadly, several comprehensive simulation studies have demonstrated that many existing methods have poor performance 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 factors of influence include the extent of effect size heterogeneity, the prevalence of additional p-hacking and other questionable research practices, a 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 modeling 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 different factors that are relevant and have been previously discussed in the context for modeling publication bias, such as heterogeneity, sample size planning, and the potential modeling 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 sets out with two primary objectives. Firstly, it aims to introduce the reader to the SPEEC method and provide a comphrensive description of its assumptions and its procedure. Secondly, it aims to assess the SPEEC method in a proof of concept using secondary empirical meta-analytical data sourced from Linden & Hönekopp (2021) to preliminary assess the initial feasability of the introduced approach. For this purpose, four theoretically justifiable hypotheses involving predictions about the estimated parameters of the SPEEC method that should apply to the empirical data

are derived in the section Hypotheses. 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 detailed 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 explicitely modeling the generative process of publication bias in a simulation framwork and iteratively comparing how the distribution of effect size and sample size of simulated studies diverges from the actual empircal meta-analytical data to estimate the parameters of the model. The publication bias model of the simulation framework integrates both asumption concerning the marginal distribution of effect size and sample size, as well as how publication bias influences their joint distribution. In terms of the former, the sample sizes are modeled 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 modeled by a publication bias parameter, ω_{PBS}, which captures the probability of selecting a statistically non-significant simulated study relative to a significant one for publication. From this generative publication bias model, effect sizes and sample sizes of individual studies 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 conceptualized as an optimization problem aiming to find values for the distributional parameters and the publication bias parameter of the generative publication bias model such that the statistical divergences from the empirical data is minimized. Stochastic optimization algorithms can then be employed to estimate the parameters of the generative publication bias model.For this study, the method of choice is differential evolution.

Confirmatory Hypotheses

To assess the SPEEC method, a set of four theoretical predictions are derived, that constitute 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, we conducted a direct comparison between the correlation of effect size and sample size, serving as an 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 unsigned effect size and sample size in each meta-analysis. In other words, when the proposed method estimates high publication bias (i.e., low probabilities for $\hat{\omega}_{\text{PBS}}$) it is expected that the correlation coefficients for each meta-analysis to be more negative and conversely. In statistical terms, this implies that the regression coefficient $\beta_{z_{r_s}}$ is expected to be greater than zero.

$$\begin{aligned} \mathcal{H}_0^{(\text{i})} &: \beta_{z_{r_s}} \leq 0 \\ \mathcal{H}_1^{(\text{i})} &: \beta_{z_{r_s}} > 0 \end{aligned} \tag{1}$$

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 μ_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 $\delta < 0$. If 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 compared to the mean effect size $\hat{\delta}$, 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, for the second hypothesis of this study $\mathcal{H}^{(\text{i})}$, the quadratic regression term $\beta_{\Delta_{\mu_d}}$ is expected to be smaller than zero.

$$\begin{aligned} \mathcal{H}_0^{(\text{i})} &: \quad \beta_{\Delta_{\mu_d}} \geq 0 \\ \mathcal{H}_1^{(\text{i})} &: \quad \beta_{\Delta_{\mu_d}} < 0 \end{aligned} \tag{2}$$

Registered reports are an alternative two-stage publishing model, where study protocols are submitted, peer-reviewed and in-princple accepted prior to data collection (Chambers & Tzavella, 2022; Nosek & Lakens, 2014). In-princple accepted studies are then published regardless of the outcome of the study, because the decision to publish was made before the results of the studies are known, which eliminates publication bias (Chambers & Tzavella, 2022; Simons et al., 2014). Consequently, the effect sizes within these multisite replication projects and registered replication reports that are publishied within this framework cannot be biased by publication bias. Following this reasoning, in the third hypothesis of the present study $\mathcal{H}^{(\text{iii})}$, it can expected that the average effect size of a registered replication report $\hat{\delta}$ to be equivalent to the mean parameter of the Gaussian effect size distribution μ_d that is estimated using the SPEEC approach within specified equivalence bounds $\Delta_{EQ} = \{\Delta_L,\ \Delta_U\}$. The equivalence bounds are defined by the smallest effect size of interest for this study (Lakens et al., 2018) and are set to $\Delta_{EQ} = \{-0.18, 0.18\}$ (see section Smallest Effect Size of Interest for the rationale of this decision).

$$
\begin{aligned}
\Delta_{\mu_d} &= \hat{\delta} - \hat{\mu}_d \\
\mathcal{H}_{01}^{(\text{iii})} : \Delta_{\mu_d} \leq \Delta_L \quad \cap \quad \mathcal{H}_{02}^{(\text{iii})} : \Delta_{\mu_d} &\geq \Delta_U \\
\mathcal{H}_{1}^{(\text{iii})} : \Delta_L > \Delta_{\mu_d} &> \Delta_U
\end{aligned} \tag{3}
$$

In theory, one could posit that the publication bias parameter ω_{PBS} in registered replication reports should precicely equal one, as individual replication studies are predetermined in the registration of the study and included in the final report independent of their statistical outcomes. However, due to the inherent upper limits of one for the publication bias parameter ω_{PBS}, testing this point prediction would not be sensible within a null hypothesis testing framework. Instead, a relative comparison between traditional meta-analyses and registered replication reports allows for making testable predictions. More specifically, it can be expected that the publication bias parameter ω_{PBS} is greater for multisite replication studies in comparison to traditional meta-analysis. This indicates that relative to individual statistically non-significant primary studies within traditional meta-analyses, statistically non-significant individual replication studies should have a greater likelihood of being of being included in the the final registered replication report (and thus being published). In statistical terms, when the regressor is a binary indicator of the type of research synthesis with the reference level being the publication biased absent multisite replication studies (MR) and the outcome is the estimated publication bias parameter ω_{PBS}, the regression coefficient β_{MR} can be expected to be greater than zero.

$$
\begin{aligned}
\mathcal{H}_{0}^{(\text{iv})} : &\quad \beta_{MR} \leq 0 \\
\mathcal{H}_{1}^{(\text{iv})} : &\quad \beta_{MR} > 0
\end{aligned} \tag{4}
$$

Method

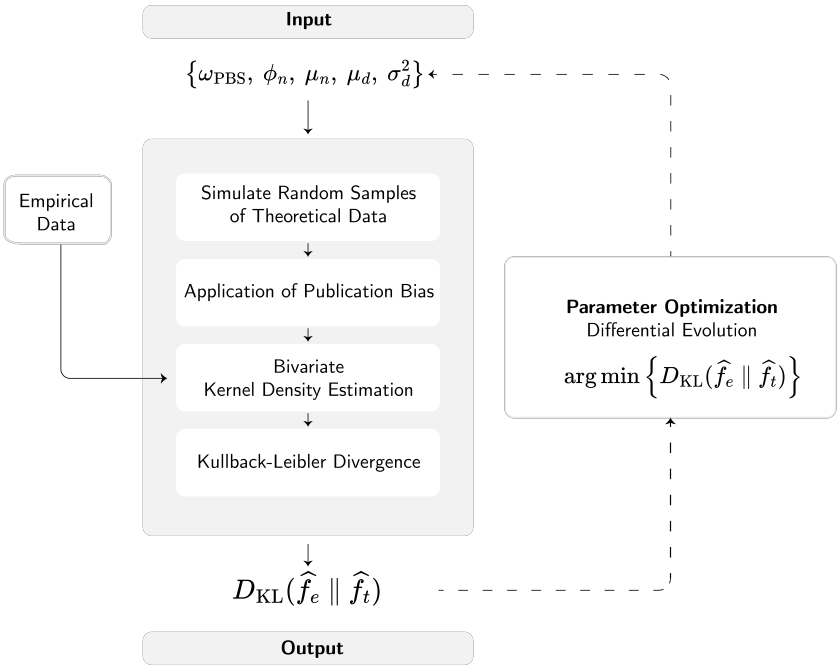
Comprehensive Methodological Description of SPEEC

This section offers a comprehensive description and explanation of the SPEEC framework to assess the extent of publication bias and estimate effect sizes in the presence of publication bias in meta-analyses. The flowchart in Figure 1 provides an overview of the sequentials steps of the SPEEC method. Currently, SPEEC is being developed as an open-source R package and is accessible on GitHub at https:

//github.com/jlschnatz/speec (alpha version).

Figure 1

Overview of the SPEEC Approach



Note. Test.

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 sizes measure for mean differences(Lakens, 2013) as the effect size of interest. Consequently, it is assumed that the simulated effect sizes 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 modeled as a discrete distribution. Count data of this nature are commonly modeled using either a Poisson or Negative-Binomial distribution. In various psychological domains, sample size distributions often exhibit considerable variance and skewness (see for example Cafri et al., 2010; Marszalek et al., 2011; Sassenberg & Ditrich, 2019; Shen et al., 2011; Szucs & Ioannidis, 2017). Considering this, the Negative-Binomial distribution was chosen as it can model overdispersed better effectively by introducing a second parameter (Ismail & Jemain, 2007; Lloyd-Smith, 2007). The mean-dispersion parametrization of the Negative-Binomial distribution is used where probability of sucess p and the target number of sucesses r are reparametrized to mean $\mu = \frac{r \cdot (1-p)}{p}$ and dispersion $\phi = r$ to model the study-specific total sample sizes n_i.

$$n_1, n_2, \ldots, n_k \quad \text{where} \quad N \overset{\text{i.i.d.}}{\sim} \mathcal{NB}(\phi_n, \mu_n) \quad \text{for } i = 1, \ldots, k \tag{5}$$

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. 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), we compute the variance of the mean differences $\bar{x}_{i1} - \bar{x}_{i2}$, from which the effect sizes are assumed to originate in this type of study design. Subsequently, a normalization 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 that $\bar{\gamma} = 1$.

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

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

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

$$d_i = \mu_d + \epsilon_i$$

$$\epsilon_1, \epsilon_2 \ldots \epsilon_k \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, \sigma^2_{\epsilon_i})$$

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

Conditional on the marginal distributions, k individual studies are sampled from the joint distribution of effect size and sample size. The selection of k is user-defined, however, the larger the number of samples k, the lower the uncertainty of the joint distribution of effect size and sample size. In general, there is a trade-off between the increased computational cost and the reduced uncertainty of the joint distribution for increasing values of k. We opted for $k = 10^4$ samples for each simulation iteration for the analyses of 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 samples 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. Statistical significance in traditional null hypothesis significance testing (NHST) contexts is commonly determined using p-values, employing a dichotomous decision rule with conventional type I erro-rates set as cut-off values (Andrade, 2019; McShane et al., 2019). 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. That is, 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, respectively. Subsequently, the p-value p_i of each simulated study can be derived from its corresponding t-value

$$t_i = \Big| \frac{d_i}{\sqrt{1/n_{1i} + 1/n_{2i}}} \Big|. \tag{8}$$

$$p_i = 2 \cdot P(t_i \mid df_i) \tag{9}$$

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

$$\omega_{\text{PBS}_i}(p_i) = \begin{cases} \omega_{\text{PBS}} & \text{for } \; p_i \geq \alpha \\ 1 & \text{otherwise} \end{cases} \tag{10}$$

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 analogue 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, else the probability of a study being selected is equal to one, indicating no publication bias. Thus, the publication bias parameter ω_{PBS} denotes the probability of the selection of a non-significant study relative to a statistiacally significant study. For instance, a publication bias parameter of $\omega_{\text{PBS}} = 0.5$ would indicate that simulated studies that are non-significant are half as likely to be selected (i.e., published) in comparison to studies that are statistically significant. Importantly, the type I error rate needs to be fixed across all simulated studies and is set to the common significance threshold of $\alpha = 0.05$ for the analyses of the hypotheses. Following the computation of publication bias weight ω_{PBS_i} for each study i, the probability of a study being selected can be expressed as $\mathbb{P}(S_i = 1) = \omega_{\text{PBS}_i}$. Here, S_i is a binary indicator function, signifying whether an individual study i is selected during the publication bias selection process. Following the computation of the 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.

$$1_\alpha(\omega_{\text{PBS}_i}) = \begin{cases} 0 & \text{if} \quad \omega_{\text{PBS}_i} \geq \alpha \quad \text{study not selected} \\ 1 & \text{if} \quad \omega_{\text{PBS}_i} < \alpha \quad \text{study selected} \end{cases} \tag{11}$$

Using this indictor 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

$$\begin{aligned} \{d_i', n_i'\} &= \{d_i, n_i \mid 1_\alpha(\omega_{\text{PBS}_i}) = 1\} \quad \text{for} \quad i = 1, \dots, k' \quad \text{and} \\ \{d_i'', n_i''\} &= \{d_i, n_i \mid 1_\alpha(\omega_{\text{PBS}_i}) = 0\} \quad \text{for} \quad i = 1, \dots, k''. \end{aligned} \tag{12}$$

One challenge in simulating a fixed number of k studies lies in the fact that, ceteris paribus, as the degree of publication bias increases (i.e. for lower values for ω_{PBS}), increasingly fewer studies k' remain after the selection process compared to the original number of simulations k. This would lead to a loss of precision in the parameter estimation for decreasing values of ω_{PBS}. To adress this, the aforementioned steps (from Equation 5 - Equation 12) 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}.

Formulation as an Optimization Problem

After the simulation of k_{adj} from the publication bias model conditional of 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 $i = 1, \dots, i = k'_{\text{adj}}$, 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 (KL-divergence, Kullback & Leibler, 1951) was chosen as a dissimlarity measure for SPEEC. This choice was motivated by the KL-divergence's superior performance in capturing 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, the KL divergence has an intuitive interpretation

in this context. It can be interpreted as the expected value of the log-likelihood ratio favoring the ture 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 from the simulated data $\hat{f}_t$ generated by the publication bias model serves as the approximate candidate distribution.

To implement this step, the *KernSmooth* R package (Wand et al., 2023) was used to estimate the joint kernel density distribution for both empirical data and simulated 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 to be $n_{\text{grid}} = 2^7 + 1$ equidistant grid points in each dimension but is user-definable in the *speec* R package. The bandwidth of the kernel function was determined using the reliable plug-in method proposed by Sheather & Jones (1991), but the *speec* 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 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. Using these estimates, the quantiles spanning the inner 99 percentile of the cumulative distribution are obtained 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.

$$\begin{aligned} \mathrm{b_d} &= \left\{Q_d(p_1 \mid \hat{\mu}_d,\, \hat{\sigma}_d^2) \wedge \min(d),\, Q_d(p_2 \mid \hat{\mu}_d,\, \hat{\sigma}_d^2) \vee \max(d)\right\} \\ \mathrm{b_n} &= \left\{Q_n(p_1 \mid \hat{\phi}_n,\, \hat{\mu}_n) \wedge \min(n),\, Q_n(p_2 \mid \hat{\phi}_n,\, \hat{\mu}_n) \vee \max(n)\right\} \end{aligned} \tag{13}$$

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

$$D_{\mathrm{KL}}(\hat{f}_e \parallel \hat{f}_t) = \sum_{u=1}^{n_{\text{grid}}} \sum_{v=1}^{n_{\text{grid}}} \hat{f}_e(u, v) \ln\left(\frac{\hat{f}_e(u, v)}{\hat{f}_t(u, v)}\right) \tag{14}$$

Parameter Optimization with Differential Evolution

Summarizing 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 and returns a single scalar value representing the Kullback-Leibler divergence between the estimated joint kernel density of the simulated theoretical data and the empirical meta-analytical data. This framework can be considered as an optimization problem of finding parameter values for such that

$$\begin{aligned} &\min_{\mu_d,\ \sigma_d^2,\ \mu_n,\ \phi_n,\ \omega_{PBS}} \Big\{ D_{\mathrm{KL}}(\hat{f}_e \parallel \hat{f}_t) \Big\}, \\ &\text{subject to: } \mu_d,\ \sigma_d^2,\ \mu_n,\ \phi_n,\ \omega_{PBS} \in \mathbb{R}, \quad \text{where} \quad 0 \leq \omega_{PBS} \leq 1, \\ &-4 \leq \mu_d \leq 4, \quad 0 \leq \sigma_d^2 \leq 6, \quad 10 \leq \mu_n \leq 15000, \quad 0.01 \leq \phi_n \leq 1000. \end{aligned} \tag{15}$$

For this purpose, we use the differential evolution (DE, Storn & Price, 1997), which is a simple algorithm for global optimization (Feoktistov, 2006). DE is an evolutionary metaheuristic based on principles such as as mutation, cross over and selection, and requires only few control parameters that are generally straightforward to select to achieve favorable outcomes in comparison to other optimization 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 optimization within the SPEEC approach, the R package *RcppDE* (Edelbuettel, 2022) was utilized, 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 & Price (1997) with additional adjustments informed by preliminary testing of simulated data from the simulation framework of SPEEC, setting the population size NP to 150, the mutation constant F to 0.9 and the crossover constant CR to 0.1. In the application of the DE algorithm, we adopted a direct termination criteria approach (Ghoreishi et al., 2017; Jain et al., 2001), with the termination condition being the maximum number of generations. Since there are no universally applicable default values for the maximum number of generations, as it is contingent

upon the optimization problem at hand (Jain et al., 2001), the choice for $t_{\max}$ was also informed by preliminary testing of the simulation framework of SPEEC. These tests suggested that $t_{\max} = 1000$ is a reasonable decision. In determining the boundaries for the parameter search space (see Equation 15), a balance was made between avoiding boundaries that are too wide, which could lead to inefficient exploration of the search space, and ensuring that the boundaries are not too narrow to ensure sufficient coverage of the potential parameters. More specifically, the minima and maxima for all distributional parameters were determined using Maximum Likelihood (see Table XXX), and the boundaries were set slightly above those values to ensure good coverage.

Secondary Data Description

To examine the confirmatory hypotheses of this study aiming to provide a preliminary assessment of the viability of the SPEEC method, we use secondary data sourced from previous research by Linden & Hönekopp (2021). The dataset can be accessed both from its original source (see https://osf.io/yr3xd/) and through the OSF and GitHub repositories associated with this project (see section Appendix B: Research Transparency Statement). Furthermore, detailed metadata of the dataset and a transparency statement regarding prior knowledge of the data are provided in the preregistration of this study. The dataset is comprehensive and includes both "traditional" meta-analyses and registered replication reports for which publication bias is absent. The dataset encompasses a total of 207 research syntheses covering various psychological phenomena. Within the dataset, there are 150 meta-analyses, each subset consisting of 50 meta-analyses from different subfields of psychology (social psychology, organizational psychology, and cognitive psychology). Additionally, the dataset includes 57 registered replication reports, which are particularly relevant for investigating hypotheses $\mathcal{H}^{(\text{iii})}$ and $\mathcal{H}^{(\text{iv})}$. For each research synthesis, information on the total sample size n_i and effect size d_i of each primary study was compiled. 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 & Hönekopp (2021) was that effects must be reported as standardized 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 confiromatory statistical analyses were preregistered and were performed using the R programming language (version 4.4.0 Puppy Cup, R Core Team, 2023). Any supplementary analyses not preregistered and deviations of the preregistered confirmatry analyses are explicitly labeled as such. All data and analysis scripts are made available for this thesis (see section Appendix B: Research Transparency Statement).

Regarding the hypotheses, in which the publication bias parameter ω_{PBS} was the dependent variable ($\mathcal{H}^{(\text{i})}$, $\mathcal{H}^{(\text{ii})}$, $\mathcal{H}^{(\text{iv})}$), beta regression as implemented in the *betareg* package (Zeileis et al., 2021) was used to analyse the data. This choice was motivated by the restriction of the parameter space for the publication bias to the standard unit interval, whereby non-normality, skewness and heteroscedasticity can anticipated (Cribari-Neto & Zeileis, 2010; Smithson & Verkuilen, 2006). Beta regression is recognized for its adaptability in handling such deviations. Because the optimization approach permits $\hat{\omega}_{\text{PBS}}$ values within the range $0 \leq \hat{\omega}_{\text{PBS}} \leq 1$ in principle, and the beta-regression model assumes that $0 < \hat{\omega}_{\text{PBS}} < 1$, a common transformation proposed by Smithson & Verkuilen (2006) was employed. This transfoormation, denoted as $\omega'_{\text{PBS}} = \frac{\omega_{\text{PBS}} \cdot (n-1) + 0.5}{n}$, subtly adjusts the bounds to sligthly narrow the range between zero and one. We used a logit link for the mean parameter μ and a identity link for the dispersion paramater that was fixed such that the beta regression model can be described as

$$\begin{aligned} \omega_{\text{PBS}'_i} &\sim \mathcal{B}(\mu_i, \phi) \\ \log\left(\frac{\mu_i}{1-\mu_i}\right) &= x_i^\top \beta. \end{aligned} \tag{16}$$

The independent variables for these three hypotheses were as follows: regarding $\mathcal{H}^{(\text{i})}$, 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 $\mathcal{H}^{(\text{ii})}$, 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 $\mathcal{H}^{(\text{iv})}$, was a binary indicator specifying the research synthesis type (traditional meta-analysis or registered replication report), with registered replication reports set as the reference level for regression. The coefficients from the beta-regressions for

these hypotheses were estimated using Maximum Likelihood estimation with the BFGS optimizer.

Hypothesis $\mathcal{H}^{(\text{iii})}$ was formulated to comparing 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 procedure (Schuirmann, 1987) was conducted as implemented in the *TOSTER* R package (Lakens & Caldwell, 2023). This involves conducting two one-tailed tests against the upper and lower equivalence bounds, both of which must yield significant results to claim statistical equivalence within the equivalence range Δ_{EQ}. The preregistration specified that the *means* for both variables would 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 normality assumption was violated for both distributions. In general, 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 nonparametric Wilcoxon signed-rank test as part of the TOST procedure (see Appendix XXX). Furthermore, the choice of a dependent samples test was necessitated by the dependence between the pairs of samples originating from the same underlying data. To obtain additional information about the magnitude of equivalence, calculation of effect size measures is beneficial (not preregistered analyses). However, traditional effect sizes metrics, such as Cohen´s d prove to be inadequate within equivalence testing as they don´t consider the equivalence range. Hence, the Proportional Distance PD, explicitely designed for equivalence testing, was utilized 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) are 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 ($\mathcal{H}^{(\text{i})}$, $\mathcal{H}^{(\text{ii})}$, $\mathcal{H}^{(\text{iv})}$) and one analytical ($\mathcal{H}^{(\text{iii})}$)sensitivity power analysis were conducted to determine which effect sizes we have at least 80 percent 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 A: Power Analyses determining the SESOIs). We specified the SESOI for $\mathcal{H}^{(\text{iii})}$ 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 XXX summarises all four SESOIs of the hypotheses.

Table 1

Smallest Effect Sizes of Interest of the Hypotheses

Hypothesis	SESOI	Unit
$\mathcal{H}_1$	1.28	OR
$\mathcal{H}_2$	0.59	OR
$\mathcal{H}_3$	0.17	raw unit
$\mathcal{H}_4$	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.

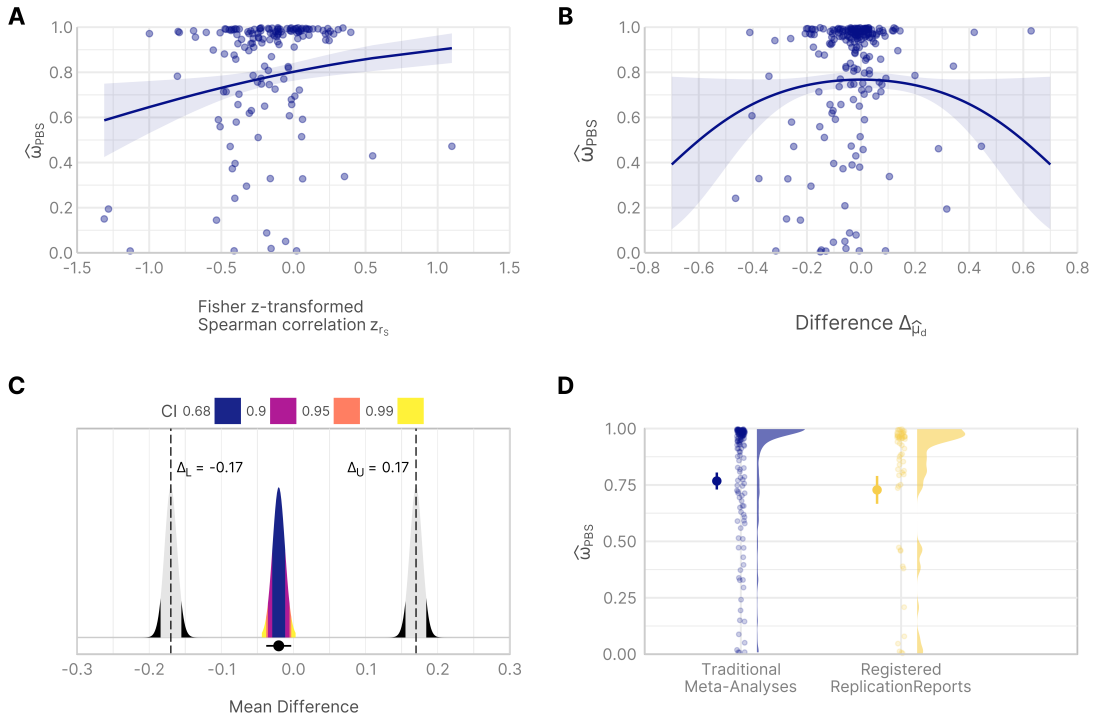
Confirmatory Results Assessing SPEEC

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

Regarding hypothesis $\mathcal{H}_1$, panel A of Figure 2 depicts the relationship between estimated publication bias parameter $\widehat{\omega}_{\text{PBS}}$ and the Fisher z-transformed Spearman correlation coefficients z_{r_s} of the effect size sample size correlation in each meta-analysis. The observed slope was sign was positive in the direction of the hypothesis and statistically signifcant $OR = 2.22$, 95% CI [1.38, Inf], $SE = 0.29$, $z = 2.74$, $p = .003$. We reject null hypothesis $\mathcal{H}_0$, that the beta coefficient is $z_{r_s} \leq 0$. Lower values for z_{r_s} were significantly associated with lower publication bias parameter values $\widehat{\omega}_{\text{PBS}}$. Additionally, to enhance the interpretability of the regression slope, we refitted the model with standardized values of z_{r_s} and computed the average marginal effects using the *marginaleffects* package (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 $\widehat{\omega}_{\text{PBS}}$. 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}.

Figure 2

Visual Summary of the Results from the Four Hypotheses



Note. **A**. Estimated publication bias parameter vs. Fisher z-transformed correlation coefficients. Fitted line: regression coefficients with 95% CI. **B**. Estimated publication bias parameter vs. difference between mean parameter and average effect size. **C**. Mean difference between mean parameter and effect size, with 90% CI, compared to null t-distribution for equivalence bounds. **D**. Comparison of estimated publication bias parameter distribution between normal meta-analyses and multisite replications. Point range: marginal predicted values with 95% CI from regression.

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

In relation to hypothesis $\mathcal{H}^{(iii)}$, panel C of Figure 2 illustrates the mean differ-

ence $\Delta_{\widehat{\mu}_d}$ between the estimated mean parameter of the Gaussian effect size distribution $\widehat{\mu}_d$ and the average effect size $\hat{\delta}$, along with its corresponding 90% confidence interval. Additionally, the null t-distributions of the Two One-Sided Tests (TOST) against the equivalence bounds $\Delta_{EQ} = \{-0.17, 0.17\}$ are illustrated. We only report the results of the t-test with the lower t-value in the main results as both tests must be significant to reject the null hypothesis (Lakens, 2017). Both one-sided paired t-tests were statistically significant, $t(56) = 17.3$, $SE = 0.01$, $p < .001$. This is also indicated by 90% confidence interval falling inside the equivalence range in panel C of Figure 2. We additionally conducted an exploratory null hypothesis significance test to test the point hypothesis that the true mean difference of 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 2, 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, 95% bootstrapped CI_{BCa} [-0.205, -0.016]. This indicates that the observed mean difference $\Delta_{\widehat{\mu}_d}$ 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 equicalence bound.

Finally, regarding hypothesis $\mathcal{H}^{(\text{iv})}$, panel D of Figure 2 depicts a comparison of the distributions of the estimated publication bias parameter $\widehat{\omega}_{\text{PBS}}$ between the traditional meta-analyses and the registered replication reports. The figure illustrates that for both traditional meta-analyses and registered replication reports, there a high density regions in the distribution of $\widehat{\omega}_{\text{PBS}}$ that is close to one. Moreover, the estimated publication bias parameter values below this high density region seem to be more uniformly distributed for the traditional meta-analyses in comparison to the registered replication reports. In the distribution of $\widehat{\omega}_{\text{PBS}}$ for registered replication reports, there are notable outliers with predicted values for $\widehat{\omega}_{\text{PBS}} < 0.5$. Already descriptively, contrary to our expectation that the estimated puhblication bias parameters for registered replication reports (RRRs) would be greater (i.e., lower publication bias) than for traditional 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 multisite replication subset ($M_{\text{MA}} = 0.82$; $M_{\text{RRR}} = 0.79$). In line with this, slope of the beta regression

was non-significant, $OR = 0.81$, 95% CI [0.61, Inf], $SE = 0.18$, z = -1.17, p $p = .879$. 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 (RRRs) to traditional meta-analysis. The model predicted a change of -3.93% in $\widehat{\omega}_{\text{PBS}}$ in the opposing direction of the hypothesis.

Intermediate Discussion of the Confirmatory Results

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 of the hypotheses, it was found that the empirical data was more extreme under the null hypotheses than the prespecified type I error rate of $\alpha = .05$ for $\mathcal{H}^{(\text{i})}$, $\mathcal{H}^{(\text{ii})}$ and $\mathcal{H}^{(\text{iv})}$, leading us to reject the null hypotheses for these predictions from a statistical viewpoint. However, it is important to evaluate these result not only in their statistical signifcance but also in terms of the practical significance by means of the magnitude of the observed effects (LeCroy & Krysik, 2007; Shaver, 1993). In this regard, we found that for $\mathcal{H}^{(\text{i})}$ and $\mathcal{H}^{(\text{ii})}$ 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 $\mathcal{H}^{(\text{iii})}$, 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, but equivalence test indicated the difference was to 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 equivalent range for the subset of RRRs. In fact, the overall magnitude of effect for the equivalence test can be considered large in terms of the results of the proportional distance PD. Most importantly however, we failed to reject to the null hypothesis for $\mathcal{H}^{(\text{iv})}$. That is, no evidence was found that the publication bias parameter for RRRs would be greater in comparison to traditional meta-analyses, or in other words that probability for selection of non-significant studies in comparison

to significant studies greater for RRRs in comparison traditional 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, it is known for a fact that publication bias in RRRs is absent.

Overall, the absence of evidence regarding $\mathcal{H}^{(\text{iv})}$ and the weak magnitude of the of the effect found for $\mathcal{H}^{(\text{i})}$ and $\mathcal{H}^{(\text{ii})}$ 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 preliminarily assess the proposed SPEEC approach, the true values for the distributional parameters and the publication bias parameter are unknown. However, as discussed previously, publication bias is inherently absent by design in registered replication reports. 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 data encompassing the RRRs for a diagnostic evaluation the parameter estimation within SPEEC. More specifically, we can derive Maximum Likelihood estimates for the distributional parameters to compare them with the corresponding values estimated by the SPEEC approach, anticipating approximate equivalence between the two approaches. This part was of the analysis was not preregistered and conceived after the confirmatory analyses were conducted. Based on this comparative approach between ML and SPEEC, we formulated 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}?

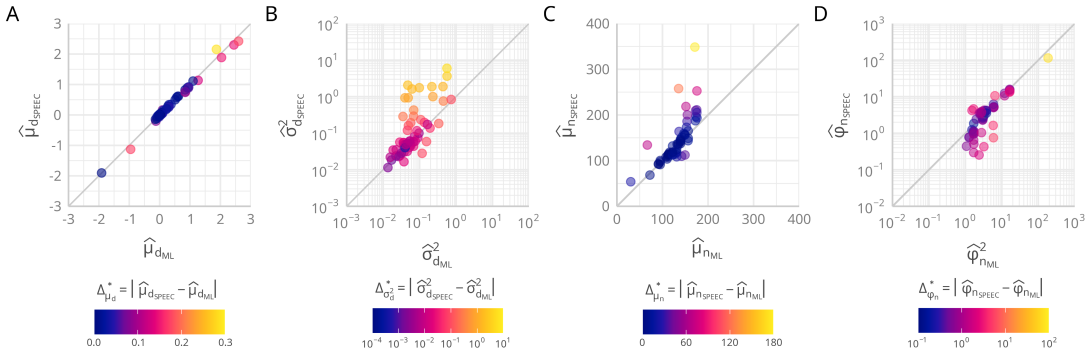
Additionally, we were interested whether the discrepancy between SPEEC and ML of the distributional parameters and the publication bias parameter ω_{PBS} were associated to potentially other relevant factors that may increases the uncertainty of the parameter estimation?

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 discrepeancy between SPEEC and MLE in the distributional parameters and the publication bias parameter ω_{PBS} associated with effect size heterogeneity?

The Maximum Likelihood estimates 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 to assess the average difference between the two estimation methods. To estimate the extent to which the true effect sizes vary within a RRR, we used the standard deviation τ of the effect sizes as a measure of between-study heterogeneity. Heterogeneity was estimated using the DerSimonian-Laird estimator (DerSimonian & Laird, 1986) in the *metafor* package (Viechtbauer, 2024).

Figure 3

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



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

Regarding the first question, Figure 3 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 compared to SPEEC, while values above the diagonal indicate the opposite. Panel A of Figure 3 reiterates the findings of the analysis of $\mathcal{H}^{(\text{iii})}$, suggesting a small discrepancy between the two methods in estimating the mean of the Gaussian effect size distribution μ_d, $M(\Delta_{\mu_d})$ = -0.03, $Mdn(\Delta_{\mu_d})$ = -0.02. This discrepancy can be deemed practically negligible according to the equivalence test of $\mathcal{H}^{(\text{iii})}$. However, the other panels indicate contrasting outcomes. In Panel B a systematic discepancy in the estimation of the variance parameter of the effect size distribution σ_d^2 between SPEEC and MLE can be observed, $M(\Delta_{\sigma_d^2})$ = 0.59, $Mdn(\Delta_{\sigma_d^2})$ = 0.11. Descriptively, this suggests that on average, the variance was estimated to be greater in the SPEEC approach compared to MLE. Furthermore, this discrepancy increases in non-linear trend and displays substantial heteroscedasticity with rising variance estimates from the MLE approach. More concretly, the greater the estimated variance from MLE, the greater the devitation 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 ($M(\Delta_{\mu_n})$ = 46.96, $Mdn(\Delta_{\mu_n})$ = 3.52), which again increases in an heteroscedastic exponential-appearing trend. Lastly, Panel D shows that SPEEC generally understimates the dispersion parameter ϕ_n of the sample size distribution in comparison to the ML estimate ($M(\Delta_{\phi_n})$ = -2.02, $Mdn(\Delta_{\phi_n})$ = -0.19) and also furthermore indicates a systematic relationship in the discrepancy between the two approaches. Lastly, Panel D illustrates that SPEEC tends to underestimate the dispersion parameter ϕ_n of the sample size distribution compared to the ML estimate ($M(\Delta_{\phi_n})$ = -2.02, $Mdn(\Delta_{\phi_n})$ = -0.19) and also furthermore indicates a systematic relationship in the discrepancy between the two approaches. As the estimated dispersion from ML descreases, the disparity between ML and SPEEC becomes progressively greater.

Table 2

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

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. test

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 total number of primary replication studies k within each registered replication report and the estimated between-study heterogeneity $\widehat{\tau}$. With regard to 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 & Hochberg (1995).

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 ML and SPEEC. This indicates that as the absolute differences between SPEEC and ML increased for

the mean parameter μ_d, the absolute discrepancy also increased for the variance parameter σ_d^2 of the effect size distribution. Furthermore, strong negative correlations were found between the publication bias parameter ω_{PBS} and the discrepancy between ML and SPEEC estimates of the mean and variance parameters of the effect size distribution. More specifically, an increases in the absolute discrepancy between both estimation methods increased for both the mean ($|\Delta_{\mu_d}|$) and variance ($|\Delta_{\sigma_d^2}|$), 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 to the divergence of ML and SPEEC of any distributional parameter or the publication bias parameter ω_{PBS}. Moreover, regarding the estimated between-study heterogeneity parameter $\hat{\tau}$, strong positive correlations were observed for divergences between ML and SPEEC both the effect size mean and variance parameter. Thus, an increase in the effect size heterogeneity of the RRRs was associated with an increases in the divergence between ML and SPEEC for both parameters of the effect size distribution. Lastly, estimated between-study heterogeneity parameter $\hat{\tau}$ was also negatively associated with the estimated publication bias parameter $\hat{\omega}_{\text{PBS}}$, such that an increases in heterogeneity predicted a decreases in the publication bias parameter (i.e., more severe publication bias). Notably, this correlation was only significant according to the unadjusted p-value.

General Discussion

- Goal of study: introduction SPEEC method, flexible simulation-based framework, assess SPEEC in proof of concept
- Derivation of Hypotheses for Proof of Concept
- Short summary of main results
- Next, why diagnostic evaluation? -> because of spurious results

Discussion of the Exploratory Results

- Analysis four theoretical predictions that should hold true, if the approach works in principle -> or phrased oppositely, if we don´t find evidence for the hypotheses -> there are problems

  Exploratory findings::

  Q1:

- There are differences between the the ML and SPEEC estimates -> for variance of effect size distribution, and mean and dispersion parameter of sample size distribution
- These differences are not unsystematic across the entire range of the distributional parameters estimated via ML but rather show systematic trends (e.g.,)

Interpretation der Heterogenität und Korrelationen zu Mittelwert, SD und Publikationsbiasparameter:

- Dadurch, dass Meta-analytischen MOdel in SPEEC derzeit Fixed-effects modell ist, kann zusätzliche variabilität über sampling error hinaus nicht modelliert werden
- wann aber in den empirischen daten heterogenität vorliegt, “muss” dadurch die varianz σ_d^2 überschätzt werden, zusätzlich μ_d (ka), aber ω_{PBS} muss kleiner werden (dadurch werden studien mit kleinen effekten und sample size weniger wahrscheinlich) und extremere effekte wahrscheinlicher und dadurch steigt die heterogenität
- Einordnung von Heterogenitätsproblem in Bezug auf andere Publikationsbias Methode -> ist ein häufiges problem
- Und ist auch erwartbar, wenn man das explizit nicht modelliert -> aber kann mit aufgenommen werden :) -> heterogenenitätsparameter τ^2

Limitations

- Limitation des SPEEC approaches an sich
- Limitationen der Studie
  - Proof of concept study -> not an comprehensive assessement of SPEEC appraoch
- Limitations of study also motivates future work

Future Work

- Possibility of including:
  - Heterogeneity parameter τ^2 -> this could be very valuable (discuss problems of current publication bias methods dealing with heterogeneity)
  - Different effect size and thus different marginal distributions (thats the flexible aspect of SPEEC) -> odds ratio (log odds ratio), R2 variance
  - Sample size planning parameter -> publiction bias not the only thing affecting n-es distribution
- Flexibility of SPEEC approach (extending SPEEC):
  - Possibility to include heterogeneity parameter (deviations from the true effect size not only due to sampling error but)
  - Possibility to include other influences on the sample size distribution -> sample size planning
  - Possibility to use other marginal distribution for sample size distribution (depending on the data)

Conclusion

- We did not find working approach for parameter optimization of the study -> this should/could be the focus of future studies
- Framework has several areas which could be responsible for optimization problems
  - The choice of grid size
- Need for larger simulations study
  - Systematically vary: heterogeneity, true effect size, publication bias severity, meta-analysis study size k
  - But also the control parameters of the SPEEC approach could be varied to

see if systematic misestimation lies within here

- For example: grid size choice for KDE has direct influece on the the loss function (KL-divergence) for the optimization -> fine grid size for low sample size studies could be problematic (because of too much uncertainty), while very coarse grid size could be also problematic (because then differences between regarding publication bias parameter cannot be captured anymore)

References

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

Andrade, C. (2019). The p value and statistical significance: Misunderstandings, explanations, challenges, and alternatives. *Indian Journal of Psychological Medicine*, *41*(3), 210–215. https://doi.org/10.4103/IJPSYM.IJPSYM_193_19

Arel-Bundock, V. (2024). *Marginaleffects: Predictions, comparisons, slopes, marginal means, and hypothesis tests* (Version 0.18.0.9) [Computer software]. https://marginaleffects.com/

Assen, M. A. L. M. van, Aert, R. C. M. van, & 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

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. https://doi.org/10.1037/h0020412

Begg, C. B. (1994). Publication bias. In *The handbook of research synthesis* (pp. 399–409). Russell Sage Foundation.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101. https://doi.org/10.2307/2533446

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. https://www.jstor.org/stable/2346101

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

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.

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

Etz, A. (2018). *Technical notes on kullback-leibler divergence*. OSF. 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

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

Ismail, N., & Jemain, A. (2007). *Handling overdispersion with negative binomial and generalized poisson regression models.* https://www.semanticscholar.org/paper/Handling-Overdispersion-with-Negative-Binomial-and-Ismail-Jemain/2791e7be78958751709b7765d92958c0b295597c

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, *3*(1), 109–117. https://www.jstor.org/stable/2245925

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

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://scholars.unh.edu/psych_facpub/194

Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*, *16*(2), 358–376. https://doi.org/10.1177/1745691620964193

Linden, A. H., Pollet, T. V., & Hönekopp, J. (2024). *Publication bias in psychology: A closer look at the correlation between sample size and effect size.* https://doi.org/10.31234/osf.io/s4znd

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

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*, 235–245. https://doi.org/10.1080/00031305.2018.1527253

Merkel, D. (2014). Docker: Lightweight linux containers for consistent development and deployment. *Linux Journal*, *2014*(239), 2:2.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). *Sustainable data analysis with snakemake* (10:33). F1000Research. https://doi.org/10.12688/f1000research.29032.1

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

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://www.jstor.org/stable/2345597

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 regis-

tered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, *9*(5), 552–555. https://www.jstor.org/stable/44290039

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. (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

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*. 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: Power Analyses determining the SESOIs

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

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

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

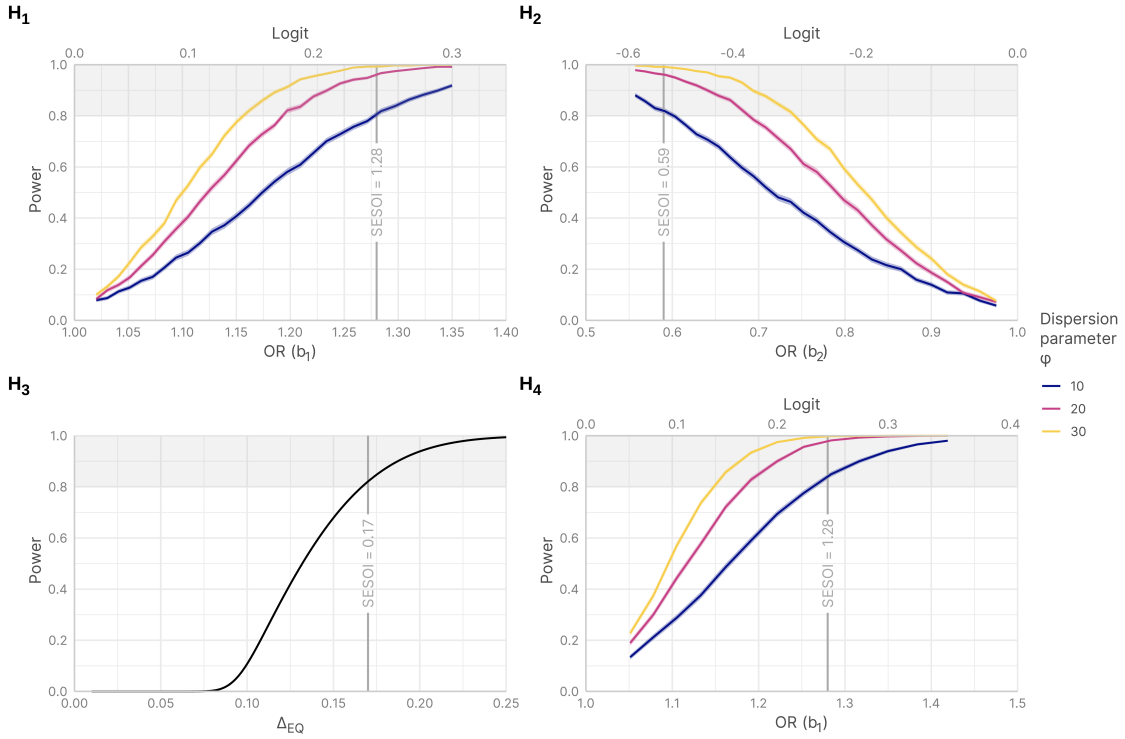
$$\mathcal{H}_1: \quad z_{r_S} \sim \mathcal{N}(\mu = -0.1, \sigma = 0.5)$$

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

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

Figure 4

Power Curves of the Sensitivity Power Analyses determining the SESOIs



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

Appendix B: Research Transparency Statement

The present thesis project was aimed to 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 on the zugehörigen OSF repository of this thesis (https://osf.io/87m9k/?view_only=030c8d1c46474270b5886c4dfc491a78) and on GitHub.To enhance the reproducability of this project, all analyses, figures, the thesis manuscript itself are generated within a containerized software environment using *Docker* (Merkel, 2014). The workflow for the analyses was explicitely management using *Snakemake* (Mölder et al., 2021) and the R packages dependencies and version management used for the statistical analyses were managemed using *renv* (Ushey & Wickham, 2024). The thesis manuscript itself is dynamically generated and reproducible using the *Quarto* publishing system.

- The project is fully st containerized using *Docker* (Merkel, 2014)

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

CITE: All analyses, figures, and the final manuscript were generated using a makefile, bash scripts, and R and Rmarkdown. All relevant metadata, as well as analysis and manuscript code, are available on GitHub: https://github.com/ZimmermanLab/SF-metrosideros-endophytes/ and archived at Zenodo (https://doi.org/10.5281/zenodo.8075450). The complete computational environment for the analyses is documented in a Dockerfile based on rocker containers (Boettiger, 2014) and a renv.lock (Ushey, 2021) file in that same repository.

Appendix C: Deviations of Preregistration

Appendix D: Test

Table 3

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

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

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

Appendix E: Regression Tables of Confirmatory Analyses

Table 4

Beta Regression Results for $\mathcal{H}_1$

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

Note. LL = 132.03, MAE = 0.21, AIC = -258.06, BIC = -249.03, R^2 = 0.051

[a] OR

[b] Identity

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

Table 5

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

[a] OR

[b] Identity

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

Testing the Assumptions of the t-test within the TOST procedure:

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 XXX). This finding was supported by the quantile-quantile plot depicted in Figure Figure 5, where the empirical quantiles did not align with the theoretical quantiles expected under a normal distribution. This was especially the case for the lower and upper quantiles of the empirical distribution.

Table 6

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

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

Note. Test

Figure 5

QQ-Plot assessing the Normality Assumption for Hypothesis 3

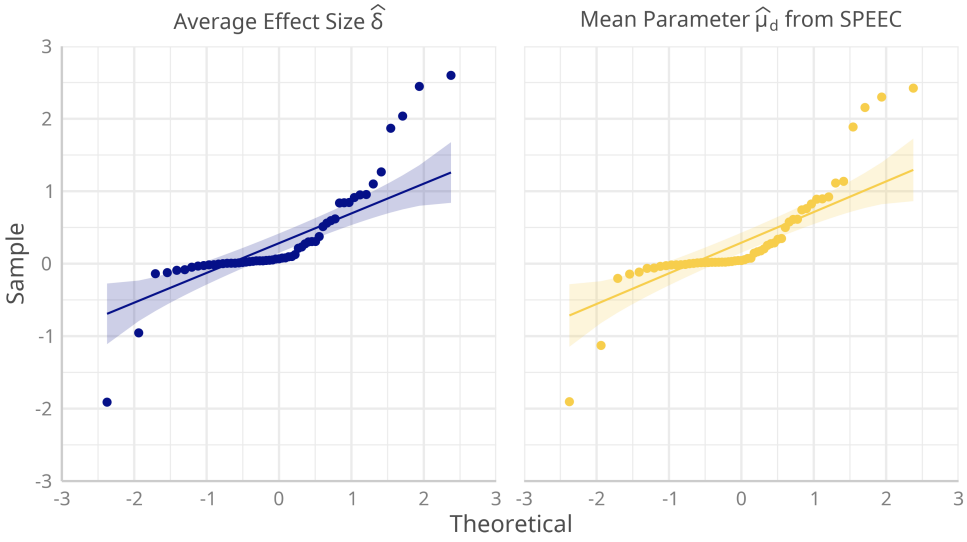


Table 7

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 8

Sensitivity Analyses for Hypotheses III: 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.

Table 9

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	$< .001$
Research Synthesis Type					
Meta-Analyses					
RRR	0.81[a]	$[0.61, Inf]$[c]	0.18	-1.17	.879
Precision model component: ϕ					
Intercept	4.79[b]	[3.59, 6.38]	0.15	10.71	$< .001$

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

[a] OR

[b] Identity

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

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich diese Abschlussarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen der Arbeit, die wortwörtlich oder sinngemäß aus anderen Quellen übernommen wurden, wurden als solche kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen und ist nicht veröffentlicht. Sie wurde nicht, auch nicht auszugsweise, für eine andere Prüfungs- oder Studienleistung verwendet.

Jan Luca Schnatz

Darmstadt, den 12. Mai 2024