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: janlucas.schnatz@gmx.de

May, 2024Y

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 enterprise (Cooper et al., 2019) with the overarching goal of attaining robust knowledge about the world (Kitcher, 1993). Within this landscape, researcher often study the same phenomen, driven by the idea that generalizing and synthesizing findings from individual studies contributes to advancement of knowledge. However, this premise hinges on the underlying assumption, that the available scientific literature is representative for all conducted research (Song et al., 2010). Contrary to this, researchers have pointed out for over half a century that results of published studies differ systematically from unpublished studie (Bakan, 1966; Bozarth & Roberts, 1972; Smart, 1964; Sterling, 1959). This discrepancy arises as the publication of a study often hinges on the direction or strength of its findings (Dickersin, 1990;
Dickersin & Min, 1993) and is collectively known as *publication bias*. Especially in a publishing culture that prioritizes novelty and positive results (Nosek et al., 2012), many statistically nonsignificant studies end up in the "file-drawer" and never get published (Rosenthal, 1979).

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

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

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

Reasoning of the n-ES correlation

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

Methodological Concerns

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

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

Firstly, it is common practice to use unsigned effect sizes to estimate the n-es correlation (e.g., Kühberger et al., 2014; Levine et al., 2009; R. Slavin & Smith, 2009; Weinerová et al., 2022). Whilst this is very common, it has only recently been acknowledged that the use of unsigned effect sizes can lead to a statistical artefact resulting in a small negative correlation, even in the absence of publication bias (Linden et al., 2024). As depicted in figure 1 (leftmost column compared to second leftmost column), the artificial correlation in absence of publication bias is most severe when the true effect is close to zero, as this condition leads to the most sign changes. Especially when considering that effect sizes in psychology are typically smaller than common benchmarks (Lovakov & Agadullina, 2021; Weinerová et al., 2022), and thus it is likely that the true effect sizes of psychological phenomen are often small, this excaberates the problem of the statistical artifact.

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

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

Apart from these more statistical challenges, there is also a more conceptual challenge. -> n-es correlation somewhat misses the point of publication bias

- Fails to capture the point of publication bias -> depending on statistical significance -> effect size and sample size correlation only indirectly captures the censorship process of non-significant studies
- as figure shows, the non-linear relationship → critical test statistic value under which p < alpha nonlinear ->
- Harrer et al. (2021)
- **Questionable linearity assumption**
  - Pearson correlation assumes that under publication bias → linear relationship between effect size and sample size → the higher the effect size the lower the required sample size for the effect to be significant and vice versa may given the (false) impression that this assumptions holds
  - But publication bias operates under statistical significance (which is most dominantly → if p-value smaller than alpha threshold; CITATION) → as figure shows, the non-linear relationship → critical test statistic value under which p < alpha nonlinear
  - Spearman correlation loosens the assumption of a linear effect in that the relationship has to be only strictly monotic → but still: this is not how publication bias operates

Limitations

- Poor performance under effect size heterogeneity
- Oftentimes only indirectly capture publication bias (small-study effects)

The Present Study

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

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 and correct effect sizes under the presence of publication bias.

utilizing/leveraging the joint distribution effect size and sample size, and how publi-

cation bias influence the joint distribution

Two-fold objective of study: first main goal, detailed description and introduction of SPEEC, second goal: proof of concept / feasability study to preliminary assess introduced method using empirical meta-analytical data.

For this, we use existing secondary meta-analytical data that has been collected by Linden and Hönekopp (2021). Includes both traditional meta-analyses as well as registered replication reports.

set of predictions/hypotheses that should hold true if the approach works in principle,

Thesis is strucured the following way, short introduction of the SPEEC method, then explanation of the hypotheses that are tested to assess the SPEEC approach (proof of concept), then detailed explanation of the SPEEC method (as it is and possible extensions), results of the hypotheses

---

Short primer of SPEEC:

The central idea of the SPEEC method is the explicit formulation of a generative model of publication bias that incorporates assumptions about the marginal distributions of effect size and sample size as well as how publication bias impacts the their joint distribution.

The extent of publication bias is modeled by a publication bias parameter ω_{PBS} that captures the probability of a statistically non-significant study being selected (i.e., published relative) relative to a significant study being selected.

From this generative publication bias model, simulation of theoretical data

More specifically, distributional parameters -> marginal distribution of effect size and sample size + publication bias parameter, gibt das ausmaß von publication bias an -> relative likelihood of statistically significant studies

Publication bias parameter

- Central idea: explicitely model generative process of publication bias -> with specific assumption about the marginal distributions of effect size and sample size and how publication bias operates
- Generative model of publication bias -> simulate theoretical data from generative model (this generative model has distributional parameters that reflect the marginal distribution of sample size and effect size as well as publication bias parameter)
- Compare simulared data from generative model to empirical meta-analytical data -> determine the divergence of the estimated bivariate kernel density of the theoretical data from the empirical data (KL-divergence) -> this serves as a loss function -> formulation as an optimization problem

- Find parameters for which the KL-divergence is minimized

Simulation framework: simulation of effect size-sample size data that is sampled from theoretical model conditional on marginal distributional assumptions of effect size (Gaussian distribution) and sample size (Negative_Binomial distribution). Introduction of publication bias parameter, which can be defined as the relative likelihood of a individual statistically non-significant study being published relative to a statistically significant study. Application of publication bias on the simulated joint distribution samples (effect size sample size). Comparison between empirial data and simulated samples from theoretical model in terms of the divergence between their respetice kernel density estimates. Using this framework -> formulation as an optimization problem -> parameter optimization regarding the marginal distributional parameters and the publication bias parameter to optimize the objective function.

The central idea of the SPEEC method involves

simulation of theoretical model that integrates assumptions about the marginal distribution of effect size and sample size and how publication bias influences the joint distribution

simulation of assumption how publication bias operates

simulation of random samples from theoretical

- Simulation-based approach to estimate publication bias severity and correct potentially biased (inflated) effect sizes under present publication bias based on the joint distribution of effect size and sample size
- Simulation of theoretical data > joint distribution of effect size and sample size under marginal distributional assumptions > Application of publication bias > empirical kernel density estimation > comparison of empirical and simulated data > loss function
- Simulation of random samples from theoretical publication bias model which integrates assumptions about the marginal distribution of effect size and sample size and how publication bias influences the joint distribution
- Comparison between random samples for theoretical model to real data → compute loss function as the statistical divergence between the estimated joint density distribution
- Iterative algorithmic optimization (differential evolution) of distributional parameters and publication bias parameter to minimize the loss function

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 $\widehat{\omega}_{\text{PBS}}$ 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 $\widehat{\omega}_{\text{PBS}}$) 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.

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

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).

$$\begin{gathered} \Delta_{\mu_d} = \hat{\delta} - \hat{\mu}_d \\ \mathcal{H}^{\text{iii}}_{01} : \Delta_{\mu_d} \leq \Delta_L \quad \cap \quad \mathcal{H}^{\text{iii}}_{02} : \Delta_{\mu_d} \geq \Delta_U \\ \mathcal{H}^{\text{iii}}_1 : \Delta_L > \Delta_{\mu_d} > \Delta_U \end{gathered} \tag{3}$$

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 almost always 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.

$$\begin{aligned} \mathcal{H}_0^{\text{iv}} : &\quad \beta_{MR} \leq 0 \\ \mathcal{H}_1^{\text{iv}} : &\quad \beta_{MR} > 0 \end{aligned} \tag{4}$$

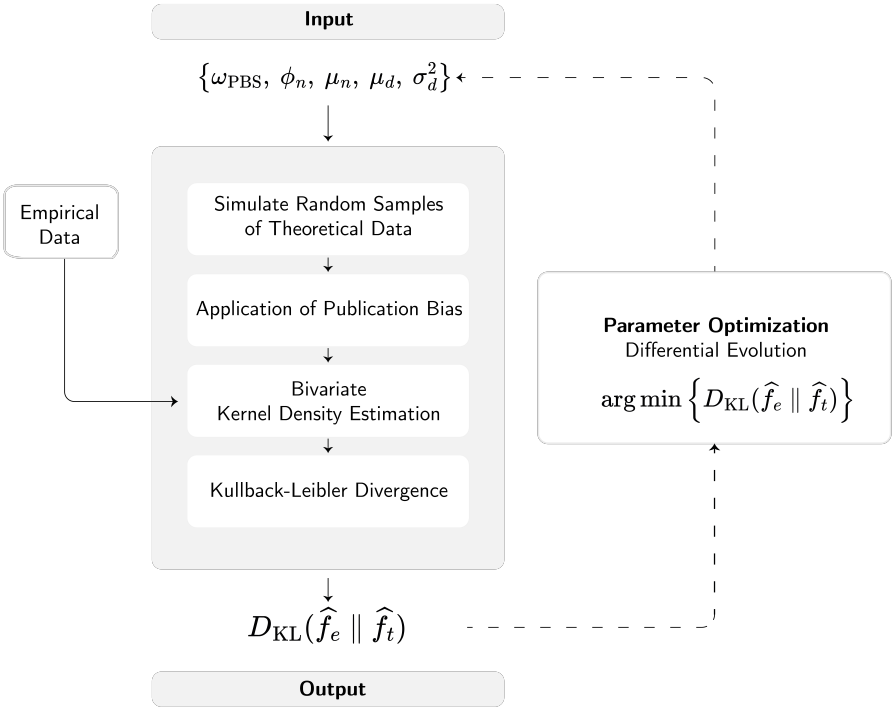
Method

The SPEEC Method

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 this study, Cohen´s d is adopted as a widely used effect sizes measure for mean differences(Lakens, 2013) as the effect size for simulation. For this reason, we assume 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 remains adaptable to different study designs where alternative effect size measures are typically employed (e.g., correlational effect sizes, effect sizes derived from proportional data). Depending on the effect size, this would require a potential adaption of the marginal distribution for the effect size and how statistical significance for each simulated study based on a different effect sizes is determined for the 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 variability and skewness we opted for the Negative-Binomial distribution which can account for variance independently of the mean and thus handle overdispered data effectively. We use the mean-dispersion parametrization where the 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 the 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), that contributes to the characteristic funnel shaped effect size-sample size distribution, we compute the variance of the mean differences $\bar{x}_{i1} - \bar{x}_{i2}$, from which the effect sizes originate in this type of design. Subsequently, a normalization factor γ_i is derived by dividing each individual variance $\sigma^2_{\bar{x}_{i1}-\bar{x}_{i2}}$ with the overall mean of those variances ensuring 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 $\text{Var}(\gamma \cdot \sigma_d^2) = \sigma_d^2$. The study-specific effect sizes d_i are subsequently modeled as

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

Using this definition for the marginal distribution of effect size, the SPEEC approach

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 that goes 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 individual studies from the joint distribution of effect size and sample size, the subsequent stage involves the application of publication bias to the random samples. 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, we assume that the individual studies i in the simulation originate form a balanced sample size design. This means that when the total sample size n_i is even, the group sample sizes n_{1i} and n_{2i} are simply 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 central t-distribution with degrees of freedom $df_i = n_{1i} + n_{2i} - 2$. Given each p-value p_i, publication bias is introduced by assigning each study i a

weight

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

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 not statistically significant (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 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 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 likelihood 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.

$$S_i = \begin{cases} 0 & \text{study not selected} \\ 1 & \text{study selected} \end{cases} \tag{11}$$

Using this binary 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

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

One challenge of simulating a fixed number of k studies is that, all else being equal, as the severity of publication bias increases (i.e., lower values for ω_{PBS}), there are fewer remaining studies k' after the selection process compared to the original number of simulations k. This would result in a loss of precision in the parameter estimation for decreasing values of ω_{PBS}. To address this, the process outlined above (from Equation 5 - Equation 12) is repeated a second time with an adjusted number of simulations $k_{\text{adj}} = \lceil k^2/k' \rceil$, ensuring that the number of selected studies k'_{adj} is roughly equal across the entire range of ω_{PBS}. This adjusted number

of of simulations k_{adj} corresponds to the ceiling of the number of initial simulations k divided by the proportion of "survived" studies k'/k.

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 parameters ω_{PBS} to determine the subset (d_i', n_i') for $i = 1, \ldots, i = k'_{\text{adj}}$, the following step involves quantifying how closely the simulated data from the publication bias model aligns with the empirical meta-analytical data. More specifically, this step involves measuring the statistical dissimilarity of the bivariate kernel density estimates between the empirical meta-analytical data and the simulated data of the publication bias model. While various measures of statistical dissimilarity between probability distributions exist (for an overview, see Cha (2007)), the Kullback-Leibler divergence (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 among estimated kernel densities compared to alternative measures (total variation distance and earthmover distance) tested for the SPEEC method, particularly across various boundary conditions. It is construed as the expected value of the log likelihood ratio favoring the true model over a candidate model (Etz, 2018), where the estimated joint density of effect size and sample size from the meta-analytical data $\widehat{f}_e$ is regarded as the true data generating distribution, and the estimated kernel density from the simulated data $\widehat{f}_t$ 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 (see 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 and Jones (1991), but the *speec* R package also offers other common bandwidth selection methods. The ensure the comparability of the estimated kernel densities between the empirical and simulated data, the bounds of the the square grid, $\text{b}_\text{n} \times \text{b}_\text{d}$, have to be exactly the same and are determined based on the empirical meta-analytical data. To define the bounds, the maximum likelihood values for the parameters of the marginal distribution of effect size ($\widehat{\mu}_d, \widehat{\sigma}_d^2$) and sample size ($\widehat{\phi}_n, \widehat{\mu}_n$) are estimated. Using these estimates, the quantiles spanning the inner 99 percentile of the cumulative distribution function are obtained from the quantile

functions $Q_d(p \mid \hat{\mu}_d,\ \hat{\sigma}_d^2)$ and $Q_d(p \mid \hat{\phi}_n,\ \hat{\mu}_n)$, where $p_1 = 0.005$ and $p_2 = 1-p_1$. Subsequently the bounds for the effect size $\mathrm{b_d}$ and sample size $\mathrm{b_n}$ are defined as the minimum and maximum values of these quantiles and the range of the empirical data, respectively.

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

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

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

Algorithmic Parameter Optimization with Differential Evolution

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

$$-4 \leq \mu_d \leq 4, \quad 0 \leq \sigma_d^2 \leq 6,$$

$$10 \leq \mu_n \leq 15000, \quad 0.01 \leq \phi_n \leq 1000, 0 \leq \omega_{PBS} \leq 1$$

{eq-optim}

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

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

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

Secondary Data Description

To examine the confirmatory hypotheses of this study aiming to provide a preliminary assessment about the viability of the SPEEC method, we use secondary data sourced from previous research by Linden and 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 XXX). Furthermore, detailed metadata about the dataset and a transparency statement regarding prior knowledge of the data are provided in the preregistration of this study. This comprehensive dataset comprises both "traditional" meta-analyses and publication bias free registered replication reports. The dataset encompasses a total of 207 research syntheses covering various psychological phenomena. Within this 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 $\mathcal{H}_3$ and $\mathcal{H}_4$. For each research synthesis, information on the total sample size and effect size of each primary study was compiled. The meta-analyses were selected via random sampling, adhering to predefined inclusion criteria

and prespecified journals from which the data was sampled from. One crucial inclusion criterion by Linden and 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 meta-analysis or replication study employed a different effect size measure than Cohen's d, the effect sizes were transformed accordingly (Linden & Hönekopp, 2021).

Statistical Analysis

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

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

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

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

Hypothesis $\mathcal{H}_3$ was aimed at comparing the estimated means of the Gaussian effect

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

Smallest Effect Size of Interest

For all four hypotheses, we will establish the smallest effect size of interest (SESOI) based on effect sizes that that can reliably detected, considering the constraints imposed by the sample size resources available for this secondary data analysis (Lakens, 2014; Lakens et al., 2018). More specifically, we conducted three simulation-based ($\mathcal{H}_1$, $\mathcal{H}_3$, $\mathcal{H}_4$) and one analytical ($\mathcal{H}_2$) sensitivity power analysis to determine which effect sizes we have at least 80 percent power ($1 - \beta = 0.8$) to detect, taking into account the constraints of the sample size and a fixed type I error rate $\alpha = .05$ (details see Appendix XXX). We specified the SESOI for $\mathcal{H}_3$ in raw units and all other SESOIs in odds ratios. The SESOI for the equivalence hypothesis will define the equivalence bounds for the TOST procedure ($\Delta_{EQ} = (-0.17, 0.17)$). Table XXX summarises all four SESOIs of the hypotheses.

Table 1

Smallest Effect Sizes of Interest of the Hypotheses

Hypothesis	SESOI	Unit
$\mathcal{H}_1$	1.28	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.

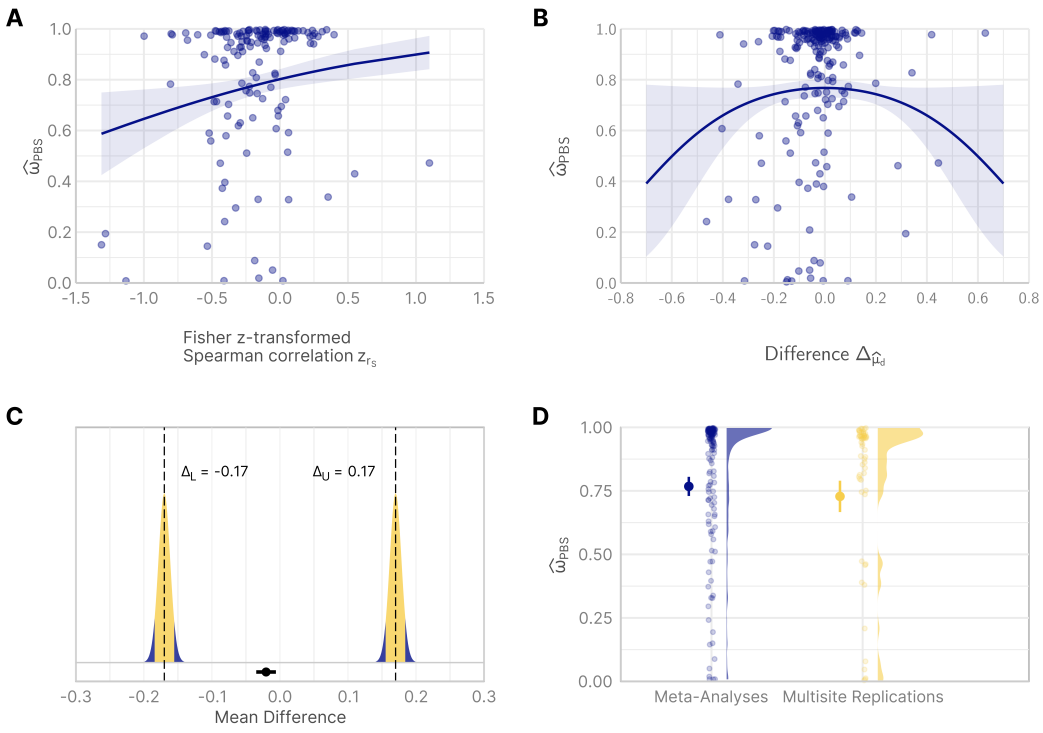
Results

Evaluation the Assumptions for the SESOIs

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.

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

Confirmatory Results of the Predictions from the Hypotheses

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, $R^2 = 0.051$. Thus, only 5.07% of the variance in ω_{PBS} could be explained by the variance of z_{r_s}.

Concerning $\mathcal{H}_2$, panel B of Figure 2 depicts the relationship between estimated publication bias parameter as a function of the difference between the average effect size $\widehat{\delta}$ and the estimated mean parameter of the Gaussian effect size distribution $\widehat{\mu}_d$. 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 $\Delta_{\widehat{\mu}_d,\widehat{\delta}}$ ($SD(\Delta_{\widehat{\mu}_d,\widehat{\delta}})$ = 0.13), the model only predicted an increase of -0.09% in the publication bias parameter $\widehat{\omega}_{\text{PBS}}$. The overall explained variation of ω_{PBS} by $\Delta_{\widehat{\mu}_d,\widehat{\delta}}$ was low, $R^2_{\text{pseudo}} = 0.026$.

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

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

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

Diagnostic Evaluation of Parameter Estimation in SPEEC

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

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

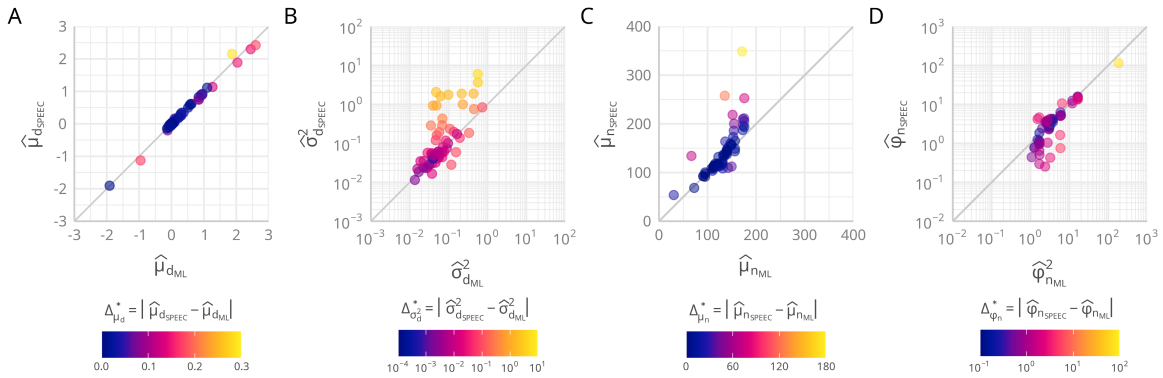
MLE?

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

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

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

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

Table 2

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

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

Note. Computed p-values are corrected for multiple comparison using the correction by Benjamini & Hochberg (1995).
$|\Delta|$ is the absolute difference for each distributional parameter between SPEEC and MLE.

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

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

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

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

Discussion

- 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

  Confirmatory findings::

- Out of the four hypotheses, we found evidence for three predictions (speaking statistically) -> H1, H2, H3
- No evidence that publication bias parameter is greater (less severe publication bias) in registered replication reports in comparison to traditional meta-analyses -> concerning result -> motivated the additional exploratory diagnostic analyses to assess potential problems of the parameter optimization

  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.,)
- Proof of concept study -> not an comprehensive assessement of SPEEC appraoch
- Possibility of including:
  - Heterogeneity parameter τ^2 -> this could be very valuable
  - 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

  “‘

- 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)
- 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)

References

Allaire, J. J., Teague, C., Scheidegger, C., Xie, Y., & Dervieux, C. (2024). *Quarto*. 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). https://marginaleffects.com/

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

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

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

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

Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist*, *27*(8), 774–775. https://doi.org/10.1037/h0038034

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type i error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research*, *45*(2), 239–270. https://doi.org/10.1080/00273171003680187

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

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

Cohen, J. (1994). The earth is round (p extless.05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 3–16). Russell Sage Foundation. http://www.scopus.com/inward/record.url?scp=84902712953&partnerID=8YFLogxK

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in r. *Journal of Statistical Software*, *34*, 1–24. https://doi.org/10.18637/jss.v034.i02

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Psychology*, *30*(1), 92–101. https://doi.org/10.5334/irsp.82

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *JAMA*, *263*(10), 1385–1389.

Dickersin, K., & Min, Y.-I. (1993). Publication bias: The problem that won't go away. *Annals of the New York Academy of Sciences*, *703*(1), 135–148. https://doi.org/10.1111/j.1749-6632.1993.tb26343.x

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Nosek, B. A. (2020). Many labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Edelbuettel, D. (2022, December 20). *RcppDE: Global optimization by differential evolution in c++* (Version 0.1.7). 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, September 6). Technical notes on kullback-leibler divergence. https://doi.org/10.31234/osf.io/5vhzu

Feoktistov, V. (2006, January 1). *Differential evolution – in search of solutions* (Vol. 5). Springer. https://doi.org/10.1007/978-0-387-36896-2

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. https://doi.org/10.1080/0266476042000214501

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

Fritz, A., Scherndl, T., & Kühberger, A. (2013). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? *Theory & Psychology*, *23*(1), 98–122. https://doi.org/10.1177/0959354312436870

Gerber, A. S., Green, D. P., & Nickerson, D. (2001). Testing for publication bias in political science. *Political Analysis*, *9*(4), 385–392. https://www.jstor.org/stable/25791658

Ghoreishi, N., Clausen, A., & Jørgensen, B. N. (2017). Termination criteria in evolutionary algorithms: A survey. *Proceedings of 9th International Joint Conference on Computational Intelligence*, *1*, 373–384. https://doi.org/10.5220/0006577903730384

Harrer, M., Cuijpers, P., A, F. T., & Ebert, D. D. (2021). *Doing meta-analysis with r: A hands-on guide* (1st). Chapman & Hall/CRC Press.

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

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.

Jennions, M. D., & Møller, A. P. (2002a). Publication bias in ecology and evolution: An empirical assessment using the 'trim and fill' method. *Biological Reviews*, *77*(2), 211–222. https://doi.org/10.1017/S1464793101005875

Jennions, M. D., & Møller, A. P. (2002b). Relationships fade with time: A meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, *269*(1486), 43–48. https://doi.org/10.1098/rspb.2001.1832

Kicinski, M. (2014). How does under-reporting of negative and inconclusive results affect the false-positive rate in meta-analysis? a simulation study. *BMJ Open*, *4*(8), e004831. https://doi.org/10.1136/bmjopen-2014-004831

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size (D. Fanelli, Ed.). *PLoS ONE*, *9*(9), e105825. https://doi.org/10.1371/journal.pone.0105825

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D. (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., & Caldwell, A. (2023, September 14). *TOSTER: Two one-sided tests (TOST) equivalence testing* (Version 0.8.0). https://cran.r-project.org/web/packages/TOSTER/index.html

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, *75*(1), 45–57. https://doi.org/10.1093/geronb/gby065

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against NonSignificant findings. *Communication Monographs*, *76*(3), 286–302. https://doi.org/10.1080/03637750903074685

Light, R., & Pillemer, D. (1984, October 1). *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

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

Marks-Anglin, A., & Chen, Y. (2020). A historical review of publication bias. *Research Synthesis Methods*, *11*(6), 725–742. https://doi.org/10.1002/jrsm.1452

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*(2), 331–348. https://doi.org/10.2466/03.11.PMS.112.2.331-348

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, January 18). Sustainable data analysis with snakemake. https://doi.org/10.12688/f1000research.29032.1

Møller, A., & Jennions, M. (2001). How important are direct fitness benefits of sexual selection? *Naturwissenschaften*, *88*(10), 401–415. https://doi.org/10.1007/s001140100255

Munafò, M. R., & Flint, J. (2010). How reliable are scientific studies? *The British Journal of Psychiatry*, *197*(4), 257–258. https://doi.org/10.1192/bjp.bp.109.069849

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–10. https://doi.org/10.1038/s41562-016-0021

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313. https://doi.org/10.1093/comjnl/7.4.308

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Palmer, A. R. (1999). Detecting publication bias in meta-analyses: A case study of fluctuating asymmetry and sexual selection. *The American Naturalist*, *154*(2), 220–233. https://doi.org/10.1086/303223

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.4.0). Vienna, 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

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, *2*(2), 107–114. https://doi.org/10.1177/2515245919838781

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Phar-*

macokinetics and Biopharmaceutics, *15*(6), 657–680. https://doi.org/10.1007/BF01068419

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

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500–506. https://doi.org/10.3102/0162373709352369

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly*, *43*(3), 290–322. https://doi.org/10.1598/RRQ.43.3.4

Smart, R. G. (1964). The importance of negative results in psychological research. *Canadian Psychologist / Psychologie canadienne*, *5a*(4), 225–232. https://doi.org/10.1037/h0083036

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71. https://doi.org/10.1037/1082-989X.11.1.54

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., Hing, C., Kwok, C. S., Pang, C., & Harvey, I. (2010). Dissemination and publication of research findings : An updated review of related biases. *Health Technology Assessment*, *14*(8), 1–220. https://doi.org/10.3310/hta14080

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

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.1080/01621459.1959.10501497

Storn, R., & Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*(4), 341–359. https://doi.org/10.1023/A:1008202821328

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature (E.-J. Wagenmakers, Ed.). *PLOS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Ushey, K., & Wickham, H. (2024, April 11). *Renv: Project environments* (Version 1.0.7). https://cran.r-project.org/web/packages/renv/index.html

Vevea, J. L., Coburn, K., & Sutton, A. J. (2019). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 383–433). Russell Sage Foundation.

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, December 1). *Kernel smoothing*. Chapman; Hall/CRC. https://doi.org/10.1201/b14876

Wand, M. P., Moler, C., & Ripley, B. (2023, July 10). *KernSmooth: Functions for kernel smoothing supporting wand & jones (1995)* (Version 2.23-22). https://cran.r-project.org/web/packages/KernSmooth/index.html

Weinerová, J., Szűcs, D., & Ioannidis, J. P. A. (2022). Published correlational effect sizes in social and developmental psychology. *Royal Society Open Science*, *9*(12), 220311. https://doi.org/10.1098/rsos.220311

Zeileis, A., Cribari-Neto, F., Gruen, B., Kosmidis, I., Simas, A. B. S., & Rocha, A. V. (2021, February 9). *Betareg: Beta regression* (Version 3.1-4). 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:

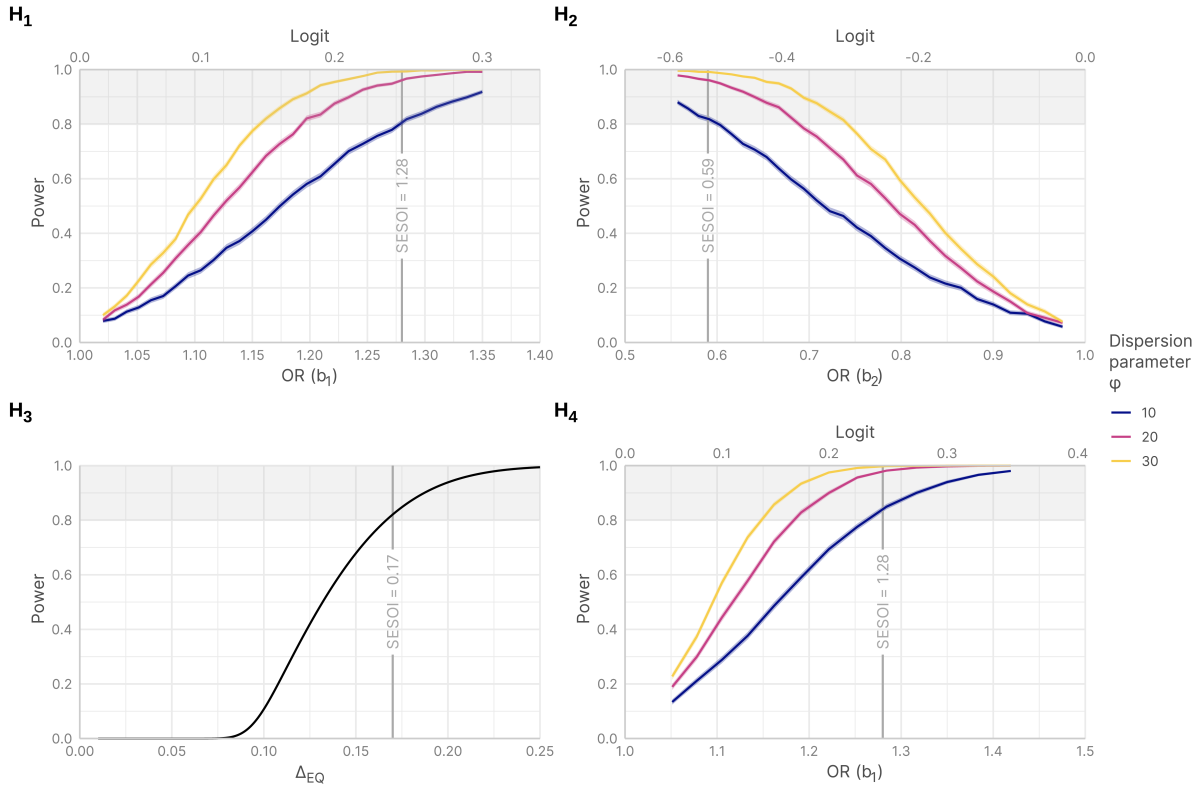
$$\mathcal{H}_1: \quad z_{r_S} \sim \mathcal{N}(\mu = -0.1, \sigma = 0.5)$$

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

For hypothesis four, the proportions of the categorical predictor of the research synthesis type (traditional meta-analysis MA, multisite replication MR) were chosen according the the actual proportions of the data ($n_{MA} = 150, n_{MR} = 57$). For all simulations-based sensitivity power analyses ($\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_4$), the number of simulations was set to $n_{iter} = 5000$. More over, the 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	.035
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

Table 6

Two One-Sided Tests Result for $\mathcal{H}_3$

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 7

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

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst habe. Wörtlich übernommene Sätze oder Satzteile sind als Zitat belegt, andere Anlehnungen, hinsichtlich Aussage und Umfang, unter Quellenangabe kenntlich gemacht. Die Aufgabenstellung habe ich alleine und ohne Besprechung mit anderen bearbeitet.

Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgele-

gen und ist nicht veröffentlicht. Sie wurde nicht, auch nicht auszugsweise, für eine andere Prüfungs- oder Studienleistung verwendet.

Datum: 18. Mai, 2024

Unterschrift: