Effect Size, Heterogeneity, and Power in Meta-Analyses of Direct Replications: A Comprehensive Simulation Study

Abstract

The field of psychology faces certain challenges such as overestimated effect sizes, false positive findings, and replication failures, all stemming from inadequate power, publication bias, and *p*-hacking. Over the past decade, the field of psychology has gone to great lengths to scrutinize well-established findings via large-scale, preregistered multi-lab replication projects. This study primarily investigates effect size variance (i.e., heterogeneity), an often-overlooked contributor for replication failures. We emphasize the importance of understanding and modelling heterogeneity using direct replications. The importance of heterogeneity in power analysis for direct replications is also demonstrated. Overall, given the inevitability of effect size heterogeneity, we highly recommend that researchers take heterogeneity into consideration when planning for future multi-lab replication projects.

Keywords: heterogeneity, effect size, direct replications, meta-analysis

Effect Size, Heterogeneity, and Power in Meta-Analyses of Direct Replications: A Comprehensive Simulation Study

Objectives

The primary objective of our study is to enhance understanding of effect size variability (i.e., heterogeneity) and its relationship to effect size and statistical power of direct replications in psychology. We utilized high-quality meta-analytic data of large-scale multi-lab replication studies to fit appropriate distributions for population effect sizes and their variances, which allowed us to derive unbiased estimates of the overall means of population effect sizes and heterogeneity in psychology. Additionally, we seek to provide insights and recommendations for multi-lab direct replication projects. To do so we will conduct simulations to investigate the impact of the number of direct replications and their sample sizes on the precision of meta-analytic effect sizes and heterogeneity estimates. Last but not least, our study seeks to explain replication failures in direct replication studies from a power analysis perspective. We will compare the power of simulated direct replications using two distinct approaches: the conventional power analysis method and a new method that accounts for heterogeneity. We intend to demonstrate that larger sample sizes may not necessarily guarantee greater power because higher effect sizes are more heterogeneous.

Background and theoretical framework

The field of psychology is currently facing a significant challenge known as the replication crisis. A number of multi-lab replication projects — including the Open Science Collaboration (Open Science Collaboration, 2015); Many Labs projects (e.g., Klein et al., 2014); and the Reproducibility Project: Psychology (RRR) (e.g., Eerland et al., 2016) — replicated a wide range of studies across various psychological subfields and assessed the robustness of their findings. These large-scale replication projects reported low replication rates and decreased effect sizes for a significant portion of tested studies. The far-reaching implications of this predicament question the very nature of psychological research credibility.

Research has identified several factors responsible for the frequent occurrence of replication failures, including low statistical power, publication bias, and *p*-hacking. An often-overlooked factor requiring further attention is effect size variability or true heterogeneity, denoted as τ^2 — in other words, the excess variance in observed effect sizes after accounting for sampling variance (McShane & Böckenholt, 2014; Röver et al., 2023). Considerable heterogeneity is usually seen in conventional meta-analyses, however, it has also been observed in direct replications. While τ^2 is estimated to be small or negligible in 50 to 80% of cases, some effect size estimates exhibit medium to large heterogeneity (Kvarven et al., 2020; Olsson-Collentine et al., 2020). Although certain researchers consider heterogeneity a contributing factor in the hindrance of successful replications, we believe its presence in direct replications affords a unique opportunity to deepen understanding of psychological effects.

The following sections present a conceptual overview of the study's essential concepts.

Direct replications

Direct replications, also known as exact replications and close replications, are designed to faithfully duplicate original studies using standardized procedures and measures to examine and verify prior research findings within the context of newly collected data. Direct replications involve public preregistration of research plans online and are therefore less susceptible to publication bias and *p*-hacking. Furthermore, one can analyze the variability of a true effect size across multiple direct replications and examine its sensitivity to variation in random contextual factors within highly controlled environments.

Meta-analysis

Meta-analysis — a statistical method designed to combine results from multiple studies on the same topic — is naturally suited for combining direct replications. The meta-analytic approach offers a framework along with a multitude of existing meta-analysis methods, enabling accurate estimations of average psychological effects and their variability.

Random-effects model. Many-to-one replications make use of multiple replication studies to verify an initial published finding (such as the Many Labs projects). The random-effects model meta-analysis is preferred to combine such replications since direct replications may possibly be measuring a distribution of effect sizes sharing a common true effect (Maxwell et al., 2015; Simons et al., 2014). The preference is supported by empirical evidence indicating considerable between-study heterogeneity for certain true effect sizes in direct replications (Stanley et al., 2018).

Cochran's Q, H^2, and I^2. To determine the presence of heterogeneity, meta-analysts typically turn to the conventional Cochran's Q test (Cochran, 1954). Relying solely on the Q-test is ill-advised, however, due to its inadequate power to detect low heterogeneity (Chung et al., 2013; Rücker et al., 2008). Heterogeneity is also commonly quantified by the relative index, I^2 (Higgins & Thompson, 2002), but it is important to note that the rules of thumb benchmarks for I^2 [1] only holds true when the within-study error is relatively constant (Borenstein et al., 2017). The less common statistic H^2(Higgins & Thompson, 2002) is able to determine whether the heterogeneity present in observed effects of replication attempts is negligible (i.e., when $H^2 \leq 1.33$, with $H^2 = 1$ indicating homogeneity) (Schauer & Hedges, 2020).

τ^2 estimator. The DerSimonian-Laird (DL) estimator (DerSimonian & Laird, 1986) is arguably the most commonly used method to estimate τ^2. The DL estimator is a method of moments estimator based on the generalized Q-statistic. It is computationally simple and effective when heterogeneity is negligible or small. However, research suggests that the DL estimator has a tendency to produce between-study variance estimates with a downward bias (Langan et al., 2017). Despite this, we deemed it sufficient for fulfilling our study purpose since, in the context of direct replications, most heterogeneity estimates will be small or negligible.

[1] I^2 values of 25%, 50%, and 75% might be considered as small, moderate, and large, respectively.

Power given heterogeneity

Ignoring effect size heterogeneity in power analysis can cause underpowered studies; conventional power calculations assume homogeneity of effect sizes. Kenny and Judd (2019) proposed a method for determining power that explicitly accounts for heterogeneity and is inspired by the Z distribution-based method developed by McShane and Böckenholt (2014). The new method employs a noncentral t-distribution to correct for over- and underestimated power values in the same and opposite direction of a true effect.

Methods

Fitting distributions for true effect sizes and heterogeneity

Data source. Olsson-Collentine et al. (2020) conducted random-effects meta-analyses of direct replications based on Many Labs and RRR projects' data, yielding 43 estimated true effects, denoted as $\hat{\theta}_j$ (Hedge's g), along with associated estimated heterogeneity, $\hat{\tau}_j$, where j represents the jth estimated true effect. These empirical meta-analytic effect sizes are accurate estimations of their respective true effect sizes. We will use these data to a fit the distribution for true effect sizes and fit another distribution for true heterogeneity.

The fitting process. We directly fitted a distribution for true effect sizes, θ_j, whereas for τ_j, we did not fit a distribution directly. Instead, we first calculated a ratio of $\hat{\tau}_j$ to $\hat{\theta}_j$ based on the empirical data, and then we fitted a distribution to the resultant ratios. As θ_j and τ_j are on the same scale, we can conveniently obtain values for τ_j by multiplying the sampled effect sizes by the sampled ratios from their respective fitted distributions.

The first step in fitting distributions is identify potential distributions that may fit the data. This begins with a visual examination of the empirical data's shape (symmetry, skewness, etc.) using a histogram and empirical density. The Cullen and Frey graph helps assess the potential fit of the data in terms of skewness and kurtosis. Its implementation in the R programming software (R Core Team, 2022) recommends a set of potential distributions for us to test their fit.

We tested the following potential distributions for θ_j: normal, gamma, Weibull, Student's t, exponential, Cauchy, log-normal, half-normal, half-t, and generalized inverse gamma distributions. For the ratios, we tested normal, Student's t, exponential, Cauchy, half-normal, and half-t distributions.

The fit was visually assessed using theoretical PDF, theoretical CDF, Q-Q plot, and P-P plot. We conducted goodness of fit tests, including Kolomogorov-Smirnov (KS), Cramer von Mises (CvM) and Anderson-Darling (AD) tests. Lastly, we used Likelihood-Ratio-Test, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to select the best model that fits our data.

Data analysis for fitted data. To estimate the overall average of parameters θ_j and τ_j, we will randomly draw 11,000 values each from the fitted distributions of θ_j and τ_j, repeating this process for 10,000 times. Numerous studies (e.g., Klein et al., 2018) demonstrated a positive correlation between effect sizes and between-study heterogeneity. To reflect such a tendency in the population parameter values, we will first sort the sampled θ_j and τ_j in ascending order following which we will introduce a minor random error that follows $N(0, 0.1)$ to both parameters. Each θ_j is grouped with a corresponding τ_j to constitute a parameter pair. We will exclude τ_j values that are larger than 1, because such values are rarely seen in empirical data and are therefore considered outliers. We will conduct descriptive statistical analyses for the remaining values of θ_j and τ_j in each sample, which includes calculating the means, medians, and interquartile ranges (IQRs). These will be averaged to obtain the final outcomes. We will also investigate the relationship between θ_j and τ_j by computing correlations between the two parameter values. A regression analysis will be conducted to obtain a linear relationship between θ_j and τ_j.

Simulation design

We will explore statistical properties of meta-analysis of direct replications through an extensive simulation study. This involves randomly drawing 1,000 pairs of parameters values from the fitted distributions. For each parameter pair, we will simulate 1,000 random-effects meta-analyses of experimental studies (i.e., two-group between-subject design). We will set the number of studies in a simulated meta-analysis to N = 10, 30, 50, while the total sample size in a primary study is set to n = 50, 100, 500. All studies will maintain a control/treatment allocation ratio to 1:1. Consequently, we will generate 1,000 × 3 × 3 × 1,000 = 9,000,000 meta-analyses. For each simulated meta-analysis, we will record the Q-statistic and its associated p-value, I^2, H^2, and τ estimated by the DL method.

Furthermore, simulations will be conducted to investigate changes in power given heterogeneity across a range of true effect sizes (standardized mean difference). θ_j will be set to 0, 0.2, 0.5, 0.8, corresponding to no, small, medium, and large average effects. We will derive corresponding τ_j based on the obtained linear relationship between θ_j and τ_j. Total sample sizes are set to 100, 200, and 500 with α set at 0.05. For each simulation setting above, we will calculate power given heterogeneity 10,000 times and take the average of the calculated power values.

Preliminary results

A brief summary of preliminary results is provided here. Detailed results will be provided in our extended paper.

We find that the lognormal distribution fits θ_j the best, and the half-t distribution fits the ratios the best. The overall average of psychological effects is estimated to be approximately 0.16, and the median is about 0.11. The overall average of effect size standard deviations is estimated to be approximately 0.1, and the median is about 0.04. These simulated results are consistent with empirical findings.

The remaining simulations and analyses are expected to be completed during Fall 2023.

Scientific significance

This study seeks to contribute to academia's understanding of effect size heterogeneity by focusing on three distinct facets. First, to the best of our knowledge, our study is the first simulation study that models plausible distributions for true effect sizes and heterogeneity on the standardized mean difference scale (Hedges' g) in psychology. Our estimation takes into account the positive (and oft-neglected) relationship between effect size and heterogeneity to present a quantitative depiction of psychological effects that's notable for its increased precision. Second, our study investigates how the positive association of effect size and heterogeneity can compromise a replication attempt's statistical power and, in the process, provide a novel perspective on the commonality of replication failure events. Third, the study highlights how power calculated conventionally differs from power which accounts for heterogeneity when making its determination. The latter method emphasizes the importance of reevaluating power calculations for direct replication attempts.

References

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research synthesis methods*, *8*(1), 5–18.

Chung, Y., Rabe-Hesketh, S., & Choi, I.-H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in medicine*, *32*(23), 4071–4089.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*(1), 101–129.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, *7*(3), 177–188.

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., Berger, S. A., Birt, A. R., Capezza, N., Carlucci, M., et al. (2016). Registered replication report: Hart & albarracın (2011). *Perspectives on Psychological Science*, *11*(1), 158–171.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, *21*(11), 1539–1558.

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological methods*, *24*(5), 578.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., et al. (2014). Investigating variation in replicability. *Social psychology.*

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490.

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434.

Langan, D., Higgins, J. P., & Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. *Research synthesis methods*, *8*(2), 181–198.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487.

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, *9*(6), 612–625.

Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*(10), 922.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

R Core Team. (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org

Röver, C., Sturtz, S., Lilienthal, J., Bender, R., & Friede, T. (2023). Summarizing empirical information on between-study heterogeneity for bayesian random-effects meta-analysis. *Statistics in Medicine.*

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on i 2 in assessing heterogeneity may mislead. *BMC medical research methodology*, *8*, 1–9.

Schauer, J. M., & Hedges, L. V. (2020). Assessing heterogeneity and power in replications of psychological experiments. *Psychological bulletin*, *146*(8), 701.

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.

Stanley, Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological bulletin*, *144*(12), 1325.