Bachelor's thesis

What makes a replication successful?
An investigation of frequentist and Bayesian criteria to assess replication success

---

Institut für Statistik
Ludwig-Maximilians-Universität München

Author	Stephanie Armbruster
Supervisor	Dr. Sabine Hoffmann
Place, time	München, August 1, 2021

Abstract

Over the past decade, the scientific community has experienced the so called replication crisis. The replication crisis follows from the discovery that many scientific studies lack replicability. This finding caused great havoc since non-replicability threatens the quality and validity of science in a philosophical as well as practical sense. Replicability attributes a scientific study its fundamental purpose to gather evidence for the designated research hypothesis. To ascertain the replicability of a certain study, replication studies are conducted. Despite the fundamental relevance and importance, the definitions for replication study and replicability are controversially discussed. So far, no standard statistical method exists to evaluate whether a replication study either corroborates or falsifies its original study.

This thesis explores some of the controversies surrounding replicability, the reasons and concerns of the replication crisis and its effects. It focuses on investigating popular frequentist, Bayesian and interval-based criteria to determine replication success. The criteria are assessed on three multi-lab examples and a simulation study with three different scenarios. They are influenced by distorting factors to varying degrees. The selected criteria differ in their meaningfulness as well as in their ability to detect distortions and handle variation in treatment effect.

The complexity of replicability requires a wholesome approach. Therefore, the thesis recommends the usage of replication reports which incorporate a wider range of criteria to assess replication success. It closes by emphasizing the importance of a heightened awareness surrounding replicability and the increased popularity of proper statistical evaluation of replication studies in order to regain public trust in science and guarantee well founded scientific progress.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the past decade, the scientific community has faced a fundamental crisis - the replication crisis. The crisis was triggered when the first sentence in Ioannidis, 2005 read: "It can be proven that most claimed research findings are false." (p. 696)
From then onward, the awareness of the fundamental importance of replicability and the rate of active implementation of replication studies have steadily increased (e.g. Pashler and Wagenmakers, 2012, Zwaan et al., 2018). Nevertheless, a lack of replicability continues to loom over Psychology as much as Social Science, Economics and Cancer biology (Pashler and Wagenmakers, 2012). It therefore remains a problem affecting science as a whole.

But what exactly does the replication crisis entail and replicability mean? And why all the excitement around it?

The replication crisis describes a deep and troublesome dilemma in which the scientific community finds itself - many scientific studies praised as advancement of science cannot be reconstructed in their findings. As a consequence to non-replicability, their validity is highly questionable (Nelson et al., 2018). Replicability - meaning the ability to obtain similar results from a different sample analyzed with the same methods - is a cornerstone of science (Zwaan et al., 2018). Without it, a study cannot serve as adequate proof to a scientific hypothesis and is thus rendered useless (Dunlap, 1926). This explains the gravity of the replication crisis - it calls the quality and validity of science into question (German Research Foundation, 2017).

Performing replication studies has been widely adopted as an effective remedy for a loss of trust in scientific findings (e.g. Open Science Collaboration, 2015, Hunter, 2001). It implies that studies are repeatably conducted - oftentimes by other research group - as similar to the original set-up as possible and analyzed according to the same methods.

If we accept the importance and relevance of replication studies for scientific discovery, we remain with the practical question of determining whether a replication study is successful or unsuccessful? How can we ascertain whether a replication study corroborates or falsifies the effect size estimated in the original study?

There is a broad spectrum of methods available on how to define replication success, ranging from simple frequentist p-value calculations (e.g. Klein et al., 2014) to complex Bayesian hierarchical models (e.g. Marsman et al., 2017).

This thesis strives to give a short introduction into the replication crisis, its reasons, relevance and consequences. It continues to identify and explain the most promising and popular criteria to determine replica-

tion success. It focuses on investigating the performance of the selected criteria based on multi-lab examples and simulation data with a special focus on possible distortions through publication bias and selective reporting. It concludes with some additional remarks on measures taken to increase replicability and on cultural challenges to overcome the replication crisis as well as concerns surrounding replication studies.

Chapter 2

Background

2.1 The replication crisis in science

Ioannidis, 2018 recounts the anecdote that in the early years of the Royal Society, researchers declaring to have made a scientific discovery would repeat the experiment in the presence of their colleagues to prove their claim. They basically conducted a replication study. More recently, in 1998, the FDA introduced the two trial rule which requires a minimum of two independent trials with a positive outcome for the registration of a new treatment. Hence they actively test for the replicability of significant findings (Rosenkranz, 2021).

Obviously, the awareness around the importance of replication and replicability is not an invention of current times (Nosek and Lakens, 2014). However, those early precautions taken to guarantee replicability were the exception, not the rule. In general, over the course of time, the notion of external accountability for scientific findings has increasingly vanished - only to be re-established as a consequence to the replication crisis now (e.g. Zwaan et al., 2018).

The long lasting neglect of replication studies and the necessity of explicit checks for replicability are puzzling since it is self explanatory that scientific claims can only gain credibility if they can be replicated and result in similar effect estimates for each replication study (Nosek and Errington, 2017).

2.1.1 Philosophy of Science and replication

From a more philosophical perspective, it can be argued that only replicability turns scientific findings into proper evidence.

Karl Popper created a Demarcation Criterion between Science and Non-Science (Musgrave and Pigden, 2021, Zwaan et al., 2018, S. Schmidt, 2009). According to this criterion, science is defined by the ability to empirically falsify its claims. This indicates that empirical observations must have the potential to prove a scientific hypothesis wrong. Replicability - the ability to replicate - is an essential property for any scientific study since it is a core method for scientific falsification (S. Schmidt, 2009). In keeping with statements by Popper and Dunlap, Zwaan et al., 2018 declares that “a finding needs to be repeatable to count as scientific discovery”. (p. 2)

Ideally, as a hypothesis cycles through multiple rounds of replications, it becomes either more robust and refined or looses credibility and vanishes (Zwaan et al., 2018). According to the theory of sophisticated falsificationism, this iterated verification process is called a research program. It can be either progressive - in

case of repeated affirmation - or degenerative - in case of repeated refutation (Zwaan et al., 2018, Musgrave and Pigden, 2021).
A progressive research program serves as proof of existing replicability. After having withstood several refutation attempts under the condition of being potentially refutable, a theory and its experiment are deemed good science and probably true (Musgrave and Pigden, 2021).

Failure to reproduce does not necessarily mean the entire theory behind the hypothesis is false. We rather have to test for additional essential characteristics to the study which might have been neglected in the replication study but were present in the original study. Hence, with every iteration, the degenerative research program becomes more specific until it reaches a certain threshold of irrelevance (Musgrave and Pigden, 2021). A degenerative research program that remains degenerative throughout the iterative testing serves as (indirect) falsification of the underlying hypothesis (Zwaan et al., 2018). Alternatively, a degenerative research program might become progressive if through further specification, studies become replicable. Consequently, replications studies contributed to developing a theory in more detail (Nosek and Lakens, 2014).

2.1.2 Reasons for non-replicability of studies

The replication crisis does not root in one cause but stems from a broad range of problematic yet established practices and complex interaction between them. Next to outright fraud, questionable research practices, methodological flexibility and the procedure of scientific publication contribute to a serious lack of replicability (Laraway et al., 2019).

Researcher's degree of freedom and methodological flexibility demonstrate a high distorting potential (Ferguson and Heene, 2012, Hoffmann et al., 2020). Researchers can choose analysis techniques and statistical models freely to best account for the study data and come to the desired conclusion. All distorting factors play their part in establishing overconfidence in an effect as well as in overestimating the effect size (Hoffmann et al., 2020).

Three major contributors to non-replicability and symptoms of methodological flexibility are publication bias, P-hacking and HARKing.

While some distorting factors have been addressed, others have remained until today (Nosek and Lakens, 2014). Overcoming the replication crisis is still very much an ongoing mission.

Publication bias

Publication bias describes a phenomenon with many nuances and a long history. Rosenthal, 1979 already identified its risks and invented a name for it - the file drawer problem. Zwaan et al., 2018 defines publication bias as "the process by which research findings are selected based on the extent to which they provide support for a hypothesis" (p. 2). Publication bias implies that studies with proof of a significant effect are published with higher probability than those with a non-significant effect - the latter are relegated into the researcher's drawer. This also extends to novel research findings as well. New results are regarded as more worthy of publishing than replicated and already explored results (Ioannidis, 2018, Nosek and Lakens, 2014).

As a direct consequence, published studies do not correctly represent all conducted studies but are a sample skewed toward successful studies with significant and / or new effects (F. Schmidt and Oh, 2016). Publication bias automatically leads to a greater rate of false positive effect estimates than expected.
Locating the fault solely with the publishing companies and their publishing policies, however, is not cor-

rect. Publication bias can be as much internally as externally motivated. Researchers refrain from submitting non-significant findings either because they are convinced of the unpublishable quality of their results or due to loyalty to a certain hypothesis, discredited by the study (Ferguson and Heene, 2012, Zwaan et al., 2018, Rosenthal, 1979).

P-hacking

P-hacking is also known under multiple alternative names, such as fishing for significance or researcher degrees of freedom. According to Zwaan et al., 2018, it describes a scientific practice in which researchers perform as many statistical tests as necessary until they achieve a significant effect. While with every test the likelihood of finding a significant effect increases, the power and replicability of such test steadily decrease. P-hacking is usually no intentional choice but a result of ignorance paired with confirmation bias. The researcher is led to believe the test resulting in a significant effect was the best and most appropriate from the beginning (Zwaan et al., 2018, Nelson et al., 2018). According to Nelson et al., 2018, P-hacking poses "the biggest threat to the integrity" of science (p. 517).

It is deeply linked to publication bias - significant results mean likely publication, mean higher record of papers and thus greater scientific reputation.

Closely related to P-hacking is fishing for the most surprising effect. It denotes the phenomenon in which the effect size calculated based on the original sample is notably greater than the effect size estimated by the replication studies due to actively constructing a surprisingly big and conclusive effect (Nelson et al., 2018).

HARKing

HARKing describes researchers formulating or changing their hypotheses post-hoc after seeing the data. The term and concept were first introduced by Kerr, 1998. HARKing is short for Hypothesizing after the results are known. It stands in stark contrast to the usual (and desirable) hypothetico-deductive approach. The latter characterizes a process in which researchers determine their research hypothesis based on some a priori knowledge and subsequently design an experiment to test their hypothesis. The potential scientific meaning of a certain study is defined before the data collection. HARKing inverts this process. Researchers perform HARKing when they phrase their hypothesis a posteriori after having detected certain patterns in the data. However, they still pretend to have determined it a priori (Kerr, 1998).

HARKing comes in different versions and can be deemed a consequence of publication bias. If a researcher adapts the hypothesis based on the data, the data will by definition lead to a significant effect and thus be more likely to be published. A simple remedy for HARKing are replication studies. They assess if one is "HARKing to explain an illusion" (Kerr, 1998, p. 207) or if the effect actually exists in reality, i.e. is replicable. For more detailed insight into the phenomenon of HARKing and reasons why HARKing is such a wide spread research practice, the reader might refer to Kerr, 1998.

2.2 Replication studies

2.2.1 Definition of replicability and replication study

We have ascertained that while replicability is a central features to science it is non-existent in many studies for multiple reasons. But what exactly defines replicability? When can we justifiably term a study as a replication study? So far, we have not encountered a formal definition. Deploying one is challenging because replicability "is a confused terminology" (Plesser, 2018, p. 1).

Reproducibility vs. replicability
The first step in clarifying the definition of replicability and replication studies is distinguishing between reproducibility and replicability.

Reproducibility describes a preliminary stage to replicability. If a study is reproducible, the original experiment in itself is conclusive and correctly performed, e.g. rerunning the code generates the same outcome as given in the original paper (Plesser, 2018). Reproducibility is determined based on the original data. Replicability relies on new data gathered according to the set-up of the original study.

Despite or maybe because of its relevance, replication or replicability are described by a myriad of different definitions. Gundersen, 2021 for example includes 16 potential understandings of replicability.

For the purpose of this thesis, we limit ourselves to two terms and respective explanations.

1. *replication (study)*: both process and outcome of repeating a study, i.e. to draw a new sample according to the original settings

2. *replicability*: the ability to perform a successful replication, i.e. from new data collected in a replication study obtain a result matching the original output (Goodman et al., 2016)

2.2.2 Purpose of replication studies

The purpose of a replication study is to assess the replicability of a study and its inherent truth. "[I]f a finding can be reliably repeated, it is likely to be true, and if it cannot be, its truth is in question." (Goodman et al., 2016, p. 4) In some sense, replicability therefore allows a mathematical capturing of reality (Goodman et al., 2016).

Thus, the goal is a sort of binary classification. We classify a replication study as a success or a failure and determine whether it corroborates or falsifies the original study.

However, the question remains - what does success in replication actually imply? According to Bayarri and Mayoral, 2002, a replication is successful when it achieves the goals it was designated for. The goals originate from a wide range of possibilities with the most popular ones including,

- Reduction of error: exact replications are conducted to gather more data and decrease the random error in estimates.

- Validation of findings: independent direct replications validate the original findings by leading to similar conclusions (frequently non-zero effect size).

- Extension of conclusions: conceptual replications assess how changes in experimental set-up influence the experimental output and to which extent the original hypothesis can be generalized (frequently on different sub-populations).

- Bias detection: replication studies which are conducted so they are not influenced by any distorting factors assess whether the original findings were likely subjected to any distorting factors (e.g. publication bias).

2.2.3 Types of replication studies

In order to fulfill the various demands placed on replication studies, there are multiple different types of replication studies which vary in design and analysis.

Direct replications

The understanding of replication in this thesis corresponds to the definition of a direct replication in the literature (Nosek and Errington, 2017). A direct replication can be understood as a study in which all relevant settings of the original study are recreated. Relevant means essential in order to capture the same information in the replication study as it has been in the original study (Zwaan et al., 2018, Nosek and Errington, 2017, Open Science Collaboration, 2015, Klein et al., 2014).

Given that studies are experimental in their nature, determining which features are essential or irrelevant for the effect proves to be a challenge. The big difficulty is to distinguish whether the reason for non-replicability can be found in the effect itself rather than in a change of experimental setting which left out an unknown but apparently relevant feature to reproduce the original study (Zwaan et al., 2018)? This question captures the main argument leveraged to dismiss the suitability of direct replications to assess replicability (Stroebe and Strack, 2014).

Simons, 2014 argues that the distinction of relevant and irrelevant experimental conditions comes in accordance with a certain degree of generalization of scientific findings and its underlying theory. Deferring differences between original and replication study solely to hidden mediators renders a theory simultaneously unfalsifiable and unprovable. "Direct replication is the only way to make sure our theories are accounting for signal and not noise." (Simons, 2014, p. 79)

Notwithstanding, replicability in the sense of successful direct replications, whatever its exact definition, cannot be equated with validity (Nosek and Errington, 2017, German Research Foundation, 2017). It merely increases the probability that a certain hypothesis is true based on one specific methodology.

Conceptual replications

In order to verify the existence without methodological condition, the hypothesis requires further verification independent of the statistical approach. This is achieved by conducting the higher acknowledged version of replication studies - conceptual replications. The primary goal of conceptual replication is to deepen theoretical understanding. It specifies the research theory, but rarely inquires whether the theory in its fundamental outline is actually correct (Nosek and Lakens, 2014).

Conceptual replications test a hypothesis with different methods to prove that the underlying theory is correct - and the effect does not happen to be an artefact from applying a certain statistical approach to the data (Nosek and Errington, 2017). This allows the conceptual replication to deliver evidence for an "extension of the theory to a new context" (Zwaan et al., 2018 p. 4).

In combination, direct and conceptual replication serve two purposes: it provides evidence for the replicability of an original finding and verifies the explanation of such finding (Nosek and Errington, 2017, Goodman et al., 2016).

The scientific community has not yet agreed upon one default method on how to determine success in replication studies (e.g. Schweinsberg et al., 2016, Gundersen, 2021). This of course leaves great room for heated debate and scientific controversy (Goodman et al., 2016). The currently applied calculation criteria range from simple p-value comparisons (e.g. Schweinsberg et al., 2016, Open Science Collaboration, 2015) to complex Bayesian random effect meta-analysis and Bayes factor calculations (e.g. Marsman et al., 2017, Ly et al., 2019, Verhagen and Wagenmakers, 2014).

In the framework of this thesis, successful replication means that the significant effect measured in the original study corresponds to reality and is not a false-positive or overestimated effect due to distorting factors such as publication bias, p-hacking or HARKing.

Chapter 3

Methodology

3.1 Mathematical background

3.1.1 Notation

Throughout the thesis, the theory behind the different criteria is presented in a consistent notation. Arbitrary parameters of interest are denoted by γ. θ stands for the raw mean difference. Its empirical estimate is written as $\widehat{\theta}$. The standardized mean difference, also called Cohen's d, is indicated by δ and the corresponding empirical estimate by d. μ traditionally stands for the mean of a variable and is empirically estimated by $\overline{x}$.

σ^2 represents the variation within each sample, while τ^2 stands for the heterogeneity or variation in treatment effect between studies. The empirical estimates are denoted by s^2 and $\widehat{\tau}^2$ respectively.

For the sake of completeness, we provide the empirical estimators in appendix A.

3.1.2 Statistical fundamentals

While the most central or less known statistical concepts are explained in this section, more popular methods are elaborated in appendix A.

One-sided t-test

In the context of the replication crisis, the primary interest lies in testing a one-sided hypothesis. The one-sided t-test is a basic statistical test used to determine if a parameter is smaller or greater than a certain value μ_0 (Fahrmeir et al., 2016).

Hence, the hypothesis can be expressed as
Version 1:

$$H_0 : \mu \leq \mu_0 \tag{3.1}$$

$$H_1 : \mu > \mu_0 \tag{3.2}$$

or alternatively,

Version 2:

$$H_0 : \mu \geq \mu_0 \tag{3.3}$$

$$H_1 : \mu < \mu_0 \tag{3.4}$$

The main conditions for using a one-sided t-test are:

1. The data points x_i originate from a normal distribution $N(\mu, \sigma^2)$.
2. The data points x_i are sampled independently of each other (iid).

If these conditions are fulfilled the t-test is an exact test.

The test is applicable for situations in which mean μ and variance σ^2 are unknown and therefore have to be estimated by the empirical mean $\overline{x}$ and variance s^2.

The t-test is calculated as

$$t = \sqrt{n}\,\frac{\overline{x} - \mu}{\sqrt{s^2}} \tag{3.5}$$

Under the null hypothesis H_0 the test statistic t is distributed according to the t-distribution with $n - 1$ degrees of freedom.

$$t \overset{H_0}{\sim} t(n-1)$$

Inserting data into formula 3.5 leads to t_0.
The p-value is determined depending on the null hypothesis.
Version 1:

$$p = P(t \geq t_0 | H_0) = 1 - T(t_0) \tag{3.6}$$

Version 2:

$$p = P(t \leq t_0 | H_0) = T(t_0) \tag{3.7}$$

where T is the cumulative t-distribution function. It can be approximated by Φ, the cumulative normal distribution function, if the sample size n is greater or equal to 30.
The effect is significant on a significance level of α if $p \leq \alpha$. The most common choice for α is 0.05 (Fahrmeir et al., 2016).

Apart from the degrees of freedom, the t-statistic has another parameter which characterizes its appearance - the non-centrality parameter c (Smithson, 2003, Bayarri and Mayoral, 2002). It has the form

$$c = \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \delta\sqrt{n} \tag{3.8}$$

with μ the actual expectation of the observed variable and μ_0 the assumed expectation under the null hypothesis.
It is easy to see why the non-centrality is neglected during classical hypothesis testing. If the null hypothesis holds, the non-centrality parameter c is equal to zero and the t-distribution becomes central. If the null hypothesis, however, does not hold, the non-centrality parameter has to be acknowledged in order to

construct an accurate confidence interval (Smithson, 2003). This applies in chapter 3.2.5 when non-central confidence intervals of the effect size are used to determine replication success.

The application of a t-test can be expanded from one sample to two samples, resulting in the two-group t-test with null hypothesis $H_0 : \theta \leq 0$ or $H_0 : \theta \geq 0$.
θ denotes the difference between the expected values of the two groups under the null hypothesis, $\theta = \mu_1 - \mu_2$.

The conditions for the one sample t-test have to be equally fulfilled in the two sample case. Additionally, we assume that the two samples are independent of each other and have the same variance $\sigma_1^2 = \sigma_2^2$. The formula for the two-group t-test starkly resembles formula 3.5

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - \theta}{\sqrt{s_p^2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.9}$$

where s_p^2 is the pooled variance.

Under the null hypothesis t is distributed according to a central t-distribution with $n_1 + n_2 - 2$ degrees of freedom, $t \overset{H_0}{\sim} t(n_1 + n_2 - 2)$ (Kelley et al., 2007).

Validity of normality assumption for big sample sizes
We can assume that - given a sufficiently large sample size - the mean $\overline{x}$ is normally distributed and independent of the distribution of the individual observations x_i due to the central limit theorem. Consequently, the t-test for one and two samples can be applied as an approximate test even if the random variable does not fulfill the normal distribution condition (Fahrmeir et al., 2016, Ghasemi and Zahediasl, 2012).

Cohen's d

Cohen's d is a popular method to measure effect size. It calculates the population standardized mean difference between two groups with expected value μ_1 and μ_2 respectively and standard deviation σ for the population (Kenny and Judd, 2019 Cohen, 2013).

$$\delta = \frac{\mu_1 - \mu_1}{\sigma} = \frac{\theta}{\sigma} \tag{3.10}$$

Cohen's d is approximated by standardizing the empirical mean difference between two groups with the empirical estimate of the pooled standard deviation, s_p , (Welkowitz et al., 2006)

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_p} = \frac{\widehat{\theta}}{s_p} \tag{3.11}$$

The main benefit of Cohen's d is comparability across studies since it is scaled. This is especially relevant for meta-analysis in which studies can be summarized despite having measured the same dependent variable on different scales (Diener, 2010).

The effect sizes within this thesis are frequently calculated as Cohen's d. Some hypotheses in chapter 3.2 are formulated based on δ, while others are based on the raw mean difference between condition and control group, denoted by θ. Due to the direct relationship between both effect measures, the hypotheses can be easily converted.

Bayes factor

The Bayes factor serves as a primary Bayesian tool in statistical inference and as a main alternative to frequentist hypothesis testing (Ly et al., 2019). Several methods presented in chapter 3.2 utilize the Bayes factor to determine replication success.

The Bayes factor calculates the factor by which the data is more likely to have been drawn from the null hypothesis than from the alternative hypothesis while including prior knowledge and expertise (Goodman, 1999, Scheibehenne et al., 2016).

The null and alternative hypothesis are defined as

$$H_0 : \gamma \in \Theta_0 \tag{3.12}$$

$$H_1 : \gamma \in \Theta_1 \tag{3.13}$$

where Θ_0 and Θ_1 are some disjunct partitions of $\mathbb{R}$.
For the Bayesian criteria concerned with establishing the existence or non-existence of an effect in this thesis, we set $\Theta_0 = 0$.

As aforementioned, the basic idea behind the Bayes factor is to decide for the hypothesis for which the collected data speaks the most (Marsman et al., 2017). This notion is mathematically expressed by the posterior odds of the probability for each hypothesis conditional on the data x.
The Bayes factor BF_{01} denotes the factor by which the prior odds, $odds_{priori} = \frac{P(H_0)}{P(H_1)}$, is updated to obtain the posterior odds (Bolstad and Curran, 2016)

$$odds_{post} = \frac{P(H_0|x)}{P(H_1|x)} = odds_{priori} \cdot BF_{01} \tag{3.14}$$

with

$$BF_{01} = \frac{f\left(x|H_0\right)}{f\left(x|H_1\right)} \tag{3.15}$$

In case no prior probability of either hypothesis exists, one might choose to assume the null and alternative hypothesis to be equally likely. Mathematically, this means selecting an uninformative prior distribution, a Bernoulli distribution with $p = \frac{1}{2}$. This leads to $P(H_0) = P(H_1) = \frac{1}{2}$ and $\frac{P(H_0)}{P(H_1)} = 1$ (Rouder et al., 2009, van Ravenzwaaij and Etz, 2021).
This a-priori choice is rather common and regarded as objective (Berger and Delampady, 1987). It allows a simplification of the formula 3.14 to

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{f\left(x|H_0\right)}{f\left(x|H_1\right)} \tag{3.16}$$

Consequently, we can justify drawing conclusion regarding the posterior odds based on the Bayes factor (Ly et al., 2019).

The Bayes factor is calculated as an average likelihood ratio. The likelihood is determined by the density function f of X and the prior density p_1 and p_2 for parameter γ depending on the hypothesis (Kauermann

and Hothorn, 2020, Verhagen and Wagenmakers, 2014, Morey and Rouder, 2011)

$$BF_{01} = \frac{f\left(x|\Theta_0\right)}{f\left(x|H_1\right)} = \frac{\int_{\gamma\in\Theta_0} f\left(x|\gamma\right) p_1(\gamma)\, d\gamma}{\int_{\gamma\in\Theta_1} f\left(x|\gamma\right) p_2(\gamma)\, d\gamma} \tag{3.17}$$

$f\left(x|\Theta_0\right)$ and $f\left(x|\Theta_1\right)$ are called marginal likelihood (Verhagen and Wagenmakers, 2014, Ly et al., 2019, Morey and Rouder, 2011). The null and alternative hypothesis differ in their subjective belief in the parameter γ - and are consequently assigned different prior distributions, p_1 and p_2.

When the null hypothesis entails a point estimate γ_0, the entire probability mass of the prior distribution is concentrated on γ_0 (Wagenmakers, Verhagen, et al., 2016)

$$p_1(\gamma) = \begin{cases} 1 & \text{if } \gamma = \gamma_0 \\ 0 & \text{else} \end{cases} \tag{3.18}$$

This kind of distributions is called a one-point density.

The alternative hypothesis could e.g. suspect positive values for γ, resulting in some one-sided prior distribution with its probability mass scattered across a range of positive values.

The interpretation of the Bayes factor is rather intuitive. The data x is by the factor BF_{01} more likely or unlikely, depending on $BF_{01} > 1$ (likely) or $BF_{01} < 1$ (unlikely), to have been gathered under assumption H_0 than H_1 (Rouder et al., 2009, Verhagen and Wagenmakers, 2014, Scheibehenne et al., 2016, Ly et al., 2019).

The relationship can easily be inverted by

$$BF_{10} = \frac{f\left(x|\Theta_1\right)}{f\left(x|\Theta_0\right)} = \frac{1}{BF_{01}} \tag{3.19}$$

BF_{10} subsequently tells us by which factor the data is more likely or unlikely to stem from the alternative hypothesis compared to the null hypothesis.

Advantages of the Bayes factor
The Bayes factor has advantages over the ordinary t-test because it

- allows accepting the null hypothesis, compared to only being able to not reject it in standard null hypothesis significance testing (Verhagen and Wagenmakers, 2014).
- is a more precise and realistic measurement of certainty, compared to p-values which tend to overstate the evidence against the null hypothesis (Rouder et al., 2009).

Its interpretation is also "straightforward and natural" (Rouder et al., 2009, p. 228). Ly et al., 2019 add that the Bayes factor allows considering evidence in sub stages while the data is still in the process of collection. The Bayes factor also incorporates Ockham's razor, i.e. it accounts for the model complexity additional to its predictive performance. This results since the prior probability mass is distributed across a bigger parameter space for complex assumptions than it is for simpler ones, resulting in lower densities at individual values including the true parameter value (Wagenmakers et al., 2010).

Classification of the Bayes factor
Jeffreys, 1961 suggests a classification of the Bayes factor according to its amount of evidence for the null hypothesis according to its magnitude.

$BF_{01} > 3$: some evidence for H_0

$BF_{01} > 10$: strong evidence for H_0

$BF_{01} > 30$: very strong evidence for H_0

Jeffreys, 1961 also proposes redeeming BF_{01} as insufficient evidence if $BF_{01} \in [\frac{1}{3}, 3]$.
Hoijtink et al., 2016 criticize this general classification approach given that the probability of the boundary values differ across different observed empirical estimates of the target variable. They propose a more customized approach to classifying the strength of evidence.

In this thesis, however, we hold on to the more generic classification of Bayes factor values.

Intervals as null hypothesis
"It is rare, and perhaps impossible, to have a null hypothesis that can be exactly modeled as $\gamma = \gamma_0$" (Berger and Delampady, 1987, p. 320). When interested in effect sizes, a null effect is rarely realistic and frequently too restrictive (Bayarri and Mayoral, 2002). Effects can either be influenced by random error or exhibit a non-zero effect size which is minuscule and irrelevant. Consequently, a null hypothesis should not be limited to a zero point estimate but incorporate a zero-centered interval with a small ϵ

$$H_0 : \gamma \in \Theta_0 = [\gamma_0 - \epsilon; \gamma_0 + \epsilon] \tag{3.20}$$

$$H_1 : \gamma \in \Theta \setminus \Theta_0 \tag{3.21}$$

Berger and Delampady, 1987 prove that the Bayes factor for the composite null and alternative hypothesis can be approximated rather well by an ordinary Bayes factor with point null hypothesis.

In this thesis, we rely on such approximation when the null hypothesis is indicated as a point null effect size.

3.2 Overview of methods: definition, explanation and criticism

3.2.1 Selection of frequentist and Bayesian criteria

The statistical assessment of replication success is subject to much controversy in the scientific community (Gundersen, 2021). Researchers agree to disagree on how to defined a successful replication (e.g. Zwaan et al., 2018, Verhagen and Wagenmakers, 2014, Bayarri and Mayoral, 2002, Held, 2020). Several different approaches to either classify replications as success or failure or to quantify the evidence for or against a null hypothesis have been proposed in recent years (e.g. Van Aert and Van Assen, 2018, Held, 2020, Simonsohn, 2015). Some are frequentist, some Bayesian and some a mixture of both.

In this thesis, we will examine 14 methods applicable for either one or multiple replication studies. The methods were selected according to a variety of criteria, from popularity of the criterion (how often were the papers cited and the methods referred to in other papers?) over topicality (how new is the approach?) and usability (is there an easy way to implement the method in R, e.g. existing code or a package?) to criticism (how well perceived is the method?) and complexity (how complex is the method?).

3.2.2 Criteria and their assumptions

Normal distribution

All criteria applied in this thesis assume the effect size to follow a normal distribution. For the sceptical p-value by Held, 2020 and the snapshot hybrid method by Van Aert and Van Assen, 2017, the normal

assumption of the effect size is a hard requirement. Hence, the normality is guaranteed by measuring the effect size as correlation and applying the Fisher transformation.

For the remaining criteria, we test for the normal distribution of $\widehat{\theta}$ visually through plots and by calculating a test statistic.

Q-Q plot
A popular and highly effective approach to check the normality assumption is the Q-Q plot. It compares the quantiles of a normal distribution to the quantiles in the observed data. The higher the congruence between the two, i.e. the closer the data points lie to the bisecting line, the more convincing the normality assumption (Ghasemi and Zahediasl, 2012).

Shapiro Wilks test
The Shapiro Wilks test is considered "the best choice for testing the normality of data" (Ghasemi and Zahediasl, 2012, p. 487). It relies on the correlation between the data and normal scores. Hence, for a test value close to 0, the null hypothesis and consequently the normality assumption are rejected (Ghasemi and Zahediasl, 2012).

For more detailed information on the mathematical formula and proofs underlying the Shapiro Wilks test, the interested reader might refer to Razali, Wah, et al., 2011.

Variance homogeneity

Across all hypotheses, the variance between condition and control group is considered homogeneous. This influences the calculation of the t-test statistic and is an essential assumption in constructing the Bayes factor.

To which extent each hypothesis fulfills this condition is established visually through plotting the difference in variance between condition and control group. The more equal the difference scatters around zero, the more substantial the evidence for variance homogeneity.

To assess homogeneity of variance statistically, the Fligner Killeen test (FK) is applied. The FK test is a non-parametric test constructed based on ranks of the difference between a data point and the median. Hence, it is robust against deviations from the normal distribution which the observations are assumed to follow (Donnelly and Kramer, 1999). The FK test evaluates

$$H_0 : \sigma^2_{cond} = \sigma^2_{crt} \tag{3.22}$$

against

$$H_1 : \sigma^2_{cond} \neq \sigma^2_{crt} \tag{3.23}$$

Heterogeneity - Variation in treatment effects

All effects included in this paper, both in the multi-lab examples (see section 3.3) and simulation data (see section 3.4), are considered variable and heterogeneous between studies. The explicit value of the effect depends on the specific sample and varies due to some unknown factors present in the sample population. The term context dependency is coined by Gollwitzer and Schwabe, 2020 to describe this dependency of effects on certain factors underlying the experimental set-up of different replication studies.
The influencing factors can be either known or unknown - so called known or hidden mediators (Bonett, 2009). While we can account for the known mediators, the number of potential hidden mediators is big,

maybe even "ultimately unknowable" (Kenny and Judd, 2019, p. 7) and "some hidden moderators will always remain hidden." (Kenny and Judd, 2019, p. 22).

Despite its common occurrence, heterogeneity is frequently neglected in statistical analysis (Kenny and Judd, 2019). So far, effects have been primarily regarded as fixed facts which either exist of a certain size across all studies or do not exist at all (Gelman, 2015a). This might be sufficiently accurate for hard sciences (e.g. physics) in which effects can be measured with a high degree of certainty and influencing factors are known and controllable. In social sciences, on the contrary, assuming fixed effects generalizes too much and disregards the prevalent uncertainty, leading to wrong conclusions and calculations (Gelman, 2015a).

In mathematical terms, heterogeneity can be modelled by drawing the study specific parameter θ_i from a normal distribution

$$\theta_i|\theta \sim N(\theta, \tau^2) \tag{3.24}$$

It equals a fixed average effect size θ plus or minus some random variability which is determined by either hidden or known mediators. θ_i characterizes the distribution of which the study observations are sampled with some within study variation σ_i^2. The estimate $\widehat{\theta}_i$ consists of two parts, the true parameter θ_i and the residual or random error $\epsilon_i = \theta_i - \widehat{\theta}_i$ (Kenny and Judd, 2019).

The variability of effect estimates $\widehat{\theta}_i$ across studies is two fold. The variance sources from variability due to sampling error, σ_i^2, which is unwanted but unavoidable as well as from the natural variability in treatment effects due to some mediators which might be explained and be of interest, τ^2 (Gelman, 2015a, Borenstein et al., 2021, Held et al., 2020). The challenge is to distinguish between the two.

Mandel-Paul or Empirical Bayes estimator

There are several different techniques to calculating heterogeneity between studies (e.g. Kenny and Judd, 2019) - a popular choice being the Mandel-Paul or Empirical Bayes (EB) estimator (Sidik and Jonkman, 2019). The EB estimator has gained a reputation of being the best overall estimator of heterogeneity.

While the study variance σ_i^2 is considered known, both hyperparameters θ and τ^2 are unknown. If unknown as well, σ_i^2 can be approximated by the standard empirical variance estimator, s_i^2.

The EB estimator considers the observations and the individual parameters for each study to be normally distributed

$$X_i|\theta_i \sim N(\sqrt{n}\,\sigma\,\theta_i, \sigma_i^2) \tag{3.25}$$

$$\theta_i \sim N(\theta, \tau^2) \tag{3.26}$$

The heterogeneity τ^2 can be estimated by the iterative estimating equation

$$\widehat{\tau}^2 = \frac{\sum_{i=1}^{k} \widehat{w}_i (k(k-1)^{-1}(d_i - \widehat{\theta}_{\widehat{w}})^2 - \widehat{\sigma}_i^2)}{\sum_{i=1}^{k} \widehat{w}_i} \tag{3.27}$$

where $\widehat{\theta}_{\widehat{w}} = \sum_{i=1}^{k} \widehat{w}_i \widehat{\theta}_i \left(\sum_{i=1}^{k} \widehat{w}_i\right)^{-1}$ and $\widehat{w}_i = \left(\widehat{\sigma}_i^2 + \widehat{\tau}^2\right)^{-1}$. In the first step, $\widehat{\tau}^2$ is set to an initial value and iteratively updated with a restriction on positive values, $\widehat{\tau}^2 \geq 0$ (Sidik and Jonkman, 2019).

3.2.3 General remarks on meta-analysis

Meta-analysis is a very powerful tool accumulating data across studies and rendering more robust estimations. F. Schmidt and Oh, 2016 argue that the only way to go forward to obtain more reliable scientific discoveries is meta-analysis. Anderson and Maxwell, 2016 even go so far as to claim that "[i]f reproducibility is the gold standard of science, then meta-analysis may be considered the gold-standard of reproducibility." (p. 7). Scientific knowledge is a cumulative and combined effort by many researchers - and conclusions by one researcher regarding one isolated study should not be attempted.

There are multiple approaches to leveraging meta-analysis to calculate replication success (e.g. Scheibehenne et al., 2016, Wagenmakers, Beek, et al., 2016, Verhagen and Wagenmakers, 2014, Schweinsberg et al., 2016). They include two different kinds of meta-analysis (Bonett, 2009).

- fixed-effect meta-analysis: the samples across all replication studies originate from an identical distribution with one communal true effect size
- random-effect meta-analysis: the parameters vary across all replication studies due to some heterogeneity

Both types of meta-analysis are present in this thesis.

3.2.4 Methods for one replication

Frequentist criteria

Classic frequentist methods are used almost exclusively to perform hypothesis tests and are well established in the scientific community as ways to gather valid evidence from scientific studies. All frequentist criteria encountered in this thesis rely on hypothesis testing and p-value calculation.

Comparison of effect size orientation

One of the most simplistic methods to examine replication success is comparing the effect estimates for each individual replication study to the original effect size with respect to their sign. In this context, the replication study is deemed successful if the original and replication effect size point into the same direction i.e. have the same sign (Schweinsberg et al., 2016).

Significance of replication p-value

Historically in the years prior to the replication crisis, hypothesis testing with t-tests and corresponding p-values has represented the almost sole criterion leveraged to determine replication success. Until today, the t-test statistic remains one of the main tools in data analysis (Simonsohn, 2015). It is widely taught to students in various study areas and included in practically every introductory statistics book (e.g. Fahrmeir et al., 2016). The p-value is defined as the probability - conditional to H_0 being true - for the test statistic to be greater than the empirical estimate obtained from the replication study (Fahrmeir et al., 2016). If the p-value is smaller than an established significance level α, the result is significant and the effect replicable.

Significance of meta-analysis p-value

The p-value can also be applied in a meta-analytic fashion. The one-sided t-test is simply calculated based on the combination of original and replication study (Schweinsberg et al., 2016).

$$t = \frac{\widehat{\theta}}{\sqrt{s^2_{orig,rep} \cdot \left(\frac{1}{n_{orig}} + \frac{1}{n_{rep}}\right)}} \tag{3.28}$$

where $s^2_{orig,rep}$ is the pooled variance, computed based on the pooled empirical variance for original and replication study, s^2_{orig} and s^2_{rep}.

$$s = \frac{1}{n_{orig} + n_{rep} - 4} \cdot \left((n_{orig} - 2)s^2_{orig} + (n_{rep} - 2)s^2_{rep}\right) \tag{3.29}$$

The nominator is the average of the mean difference between condition and control group across original and replication study, weighted by the respective sample sizes.

$$\widehat{\theta} = \frac{1}{n_{orig} + n_{rep}}(\hat{\theta}_{orig} \cdot n_{orig} + \hat{\theta}_{rep} \cdot n_{rep}) \tag{3.30}$$

The replication is considered successful if its overarching p-value is smaller than the significance level α (Schweinsberg et al., 2016, Fahrmeir et al., 2016).

Criticism on frequentist criteria

One big disadvantage of the p-value based criteria - *significance of p-value* and *significance of meta-analysis p-value* - is its varying statistical power. Statistical power describes the ability of a test statistic to result in significant outcomes if the null hypothesis is actually false. It indicates how selective the test statistic t separates null and alternative hypothesis.

$$p(t) = \mathbb{P}(t \geq t_{1-\alpha}|\mu_1) = 1 - \beta \tag{3.31}$$

β denotes the Type-II error - the counter-probability to the statistical power. β implies how likely the test is to decide in favor of the null hypothesis μ_0 despite it being untrue (Fahrmeir et al., 2016). The probability for a Type-II error is unbound by hypothesis testing, contrary to the probability for the Type-I error which is regulated by the significance level α (Fahrmeir et al., 2016).
The rate of false negatives is determined by the statistical power of the test in question. The higher the statistical power, the more certain the test results in significant outcomes when H_1 is true and thus the more accurate its conclusions. The lower the statistical power the more likely Type-II errors occur and hence the more likely non-replicability is concluded despite the true effect size being non-zero (Simonsohn, 2015).

Consequently, our confidence in the correct original finding unjustifiably suffers under underpowered studies. This does not only apply to p-value based criteria but to statistical criteria in general and hence poses a great problem. Sufficiently high statistical power has hitherto been hard to obtain in some research areas, e.g. psychology and social sciences (Amrhein et al., 2018, Gelman, 2015a). Increasing statistical power to a sufficient percentage - 80% is a popular target - frequently calls for extremely large studies. In many research areas, such large studies are infeasible - in particular for replication studies which require an even bigger sample size to achieve a substantive statistical power (Anderson and Maxwell, 2016). Especially in psychology, F. Schmidt and Oh, 2016 stress how low statistical power leads to replication studies with non-significant effect estimates despite a true alternative hypothesis. Amrhein et al., 2018 goes so far as to state that "[r]eplication studies have a false-negative problem" due to low statistical power (p. 4). Mitigation can come with conducting multiple replication studies which are pooled (Maxwell et al., 2015).

Another disadvantage to p-value based criteria is that absence of evidence is not equal to evidence of absence. When implementing the selected frequentist criteria, no evidence in favor of the null hypothesis can be gathered - only against. It is therefore impossible to quantify the amount of which a certain replication study speaks for a null hypothesis (Fahrmeir et al., 2016). A non-significant study does not prove a zero effect. Regardless of the result, the null hypothesis can never be validated, only not rejected (Goodman et al., 2016, Fahrmeir et al., 2016). Consequently, p-values can only deliver evidence for a significant non-

zero effect size of a replication study (Anderson and Maxwell, 2016). However, laymen and researchers alike frequently perceive the p-value as a method measuring the probability for the null hypothesis which leads to grave misinterpretations (Goodman, 1999). This fallacy is the main reason why, in recent years, the suitability of p-values to determine hypothesis test results has been heavily criticised (Goodman, 1999, Gelman, 2016).

Additionally, assessing replicablity solely on the basis of p-value significance neglects the actual numerical difference between the effect estimates. There are two possibilities in which regarding replication and original effect estimate as comparable through the lens of p-values is greatly misleading (Simonsohn, 2015). Either the empirical estimates differ between original and replication study but coincide in their significance or the effect estimates differ in significance but have similar values.

The first scenario arises when a replication sample with a large size obtains a significant p-value. This theoretically corroborates the significant finding in the original study. However, the replication estimate exhibits a notably smaller value than the original estimate. This small replication effect estimate could have never been detected in a study as small as the original one. Hence, concluding replication success based on the two significant p-values seems wrong since they apparently do not pick up on the same effect.
In the second case, two studies with almost identical effect sizes differ in their significance due to the estimations varying in their precision, e.g. having different sample sizes. This portrays the limited meaningfulness of p-values since "a difference in significance does not always indicate that the difference is significant" (Verhagen and Wagenmakers, 2014, p. 1457).

Another criticism is the failure of the p-value to take the adequate fit of the alternative hypothesis into account. If both null and alternative hypothesis do not match the data well, the p-value still tends to reject the null and accept H_1 (Verhagen and Wagenmakers, 2014). It overestimates the evidence against the null hypothesis (Rouder et al., 2009).

The p-value also suffers under Lindley's paradox (Lindley, 1957, Rouder and Morey, 2011). P-values under the null hypothesis are uniformly distributed. For the alternative hypothesis, the distribution depends on the true parameter and the sample size. Lindley's paradox describes the scenario in which significant p-values have a higher density under H_0 than under H_1 - thus are a more likely result for a zero than a non-zero effect. Due to the mere classification on the basis of the fixed significance threshold α, the density difference is not accounted for and the effect estimate considered sufficient proof of a non-zero effect size (Rouder and Morey, 2011).

Moreover, the *significance of meta-analysis p-value* disregards any variance in treatment effect which could exists between original and replication study.

The *comparison of effect orientation* portrays a stark simplification of information available in the replication studies. It allows only limited conclusions regarding potential distortion in the original study. In addition, it does not provide any clues on the true effect size and whether it is significantly different from zero.

Bayesian criteria - Bayes factor

There are several ways to calculate the Bayes Factor based on what kind of information is known a-priori and how it is framed. However, they all rely on the same basic logic elaborated in appendix A.5 and chapter 3.1.2.

Independent Jeffreys-Zellner-Siow Bayes factor test

The standard independent Jeffreys-Zellner-Siow (JZS) Bayes factor is named for its characterizing prior

distribution - the Jeffreys-Zellner-Siow prior.
To recall, in Bayesian statistics, we require a prior distribution for the parameter of interest and a likelihood assumption describing how the data is distributed (see chapter 3.1.2). The Jeffreys-Zellner-Siow prior is a popular generic default choice for the first component, the prior distribution (Rouder and Morey, 2011).

The JZS Bayes factor tests the hypothesis $H_0 : \delta = 0$ and answers the question whether an effect, measured as Cohen's d, is present or absent in the replication study. Based on the significance of the original study, a Bayes factor favoring the alternative hypothesis of a non-zero Cohen's effect size would be considered a success (Rouder et al., 2009, Bayarri and Garcia-Donato, 2007).

Each study condition and control group - x and y - consists of observations which follow from a normal distribution,

$$x_i \sim N(\mu + \frac{1}{2} \cdot \delta\sigma, \sigma^2) \tag{3.32}$$

iid for $i = 1, ..., n_x$ and

$$y_i \sim N(\mu - \frac{1}{2} \cdot \delta\sigma, \sigma^2) \tag{3.33}$$

iid for $i = 1, ..., n_y$. μ denotes the grand mean effect size (Bayarri and Garcia-Donato, 2007, Morey and Rouder, 2011, Marsman et al., 2017, Rouder et al., 2009).

The information contained in the observed data is captured in the t-test value. This follows from the property of sufficiency that applies to the t-test (Kauermann and Hothorn, 2020). The likelihood is thus determined by (non-central) t-distributions.

- under H_0: δ is zero and so is the non-centrality parameter, the t-statistic is t-distributed with $n_{cond} + n_{crt} - 2$ degrees of freedom

$$t \overset{H_0}{\sim} t(n_{cond} + n_{crt} - 2) \tag{3.34}$$

- under H_1: the parameter has an unknown non-zero value, leading to a non-central t-distribution with non-centrality $c = \delta \sqrt{n}$, where $n = \frac{n_{cond} n_{crt}}{(n_{cond} + n_{crt})}$ and equal degrees of freedom

$$t \overset{H_1}{\sim} NCT(c, n_{cond} + n_{crt} - 2) \tag{3.35}$$

For $H_0 : \delta = 0$, the prior probability is a one-point density. The alternative hypothesis $H_1 : \delta \neq 0$ is specified more precisely by the JZS prior (Rouder et al., 2009). The JZS prior makes as few assumptions as possible a-priori and consists of prior distributions that are highly uninformative.
The prior distribution under H_1 is per default set to

$$\delta \sim N(0, \sigma_\delta^2) \tag{3.36}$$

where σ_δ^2 is also assigned an a-priori distribution

$$\sigma_\delta^2 \sim inverse\ \chi^2(1) \tag{3.37}$$

According to Liang et al., 2008, this equals

$$\delta \sim Cauchy(0, r) \tag{3.38}$$

The Cauchy distribution is characterized by its fat tails - compared to a standard normal distribution - and its lack of first or higher order moments (Morey and Rouder, 2011, Verhagen and Wagenmakers, 2014).

The default value for the scale parameter r which determines the width of the Cauchy distribution is $\frac{\sqrt{2}}{2}$ (Morey and Rouder, 2011, Marsman et al., 2017, Fahrmeir et al., 2016). The value of the Bayes factor is heavily dependent on the value of r. Unfortunately, the choice of r is arbitrary and consequently strongly disputed (Hoijtink et al., 2016).

The variability σ^2 and grand mean μ are also unknown. For σ^2, an uninformative and improper Jeffrey's prior distribution is assumed (Morey and Rouder, 2011).

$$p(\sigma^2) \propto \frac{1}{\sigma^2} \tag{3.39}$$

For μ, an uninformative prior distribution is constructed by assigning the same probability for every value in $\mathbb{R}$

$$p(\mu) \propto 1 \tag{3.40}$$

The distributions for control and condition group across null and alternative hypothesis have both parameters μ and σ^2 in common. Consequently, the exact choice of a-priori distributions for those two parameters does not have a great influence on the Bayes factor (Rouder et al., 2009).

Together, the combination of a Cauchy prior for the Cohen's d, a constant prior for the grand mean and Jeffrey's prior for the population variance is called JZS prior.

The marginal likelihoods are obtained by

$$M_0 = \int_0^\infty t_{df,\delta\sqrt{n}=0}(x)p(\sigma^2)\partial\sigma^2 \tag{3.41}$$

$$M_1 = \int_0^\infty \int_{-\infty}^\infty t_{df,\delta\sqrt{n}}(x)p(\sigma^2)p_1(\delta)\partial\sigma^2\partial\delta \tag{3.42}$$

where $p_1(\delta)$ is said Cauchy distribution. t is the non-central t-distribution, dependent on non-centrality $\delta\sqrt{n}$ and the corresponding degrees of freedom df.
Based on this, it follows

$$\Rightarrow\ BF_{01} = \frac{M_0}{M_1} \tag{3.43}$$

Overall, the Bayes factor can be expressed depending on the sample sizes for condition and control group respectively as well as the conventional one-sided t-test statistic (Rouder et al., 2009). With $v = n_{cond} + n_{crt} - 2$ and $n = \frac{n_{cond} \cdot n_{crt}}{n_{cond}+n_{crt}}$, it is obtained

$$B_{01} = \frac{\left(1 + \frac{t^2}{v}\right)^{-(v-1)/2}}{\int_0^\infty (1 + n \cdot t)^{-1/2} \left(1 + \frac{t^2}{(1+n\cdot g)v}\right)^{-(v+1)/2} (2\phi)^{-1/2} g^{-3/2} e^{-1/(2g)} \partial g} \tag{3.44}$$

g stands for Zellner's g prior. In equation 3.44, the fact that the JZS prior can be understood as a mixture of g priors is leveraged. The one dimensional integration over the g prior can be computed using a Laplace approximation (Liang et al., 2008).

The great benefit of the JZS Bayes factor is its consistency. Depending on whether the null or alternative hypothesis is true, the factor will converge towards 0 or infinity with increasing sample size. This indicates that the larger the sample is, the more decisive the test gets (Morey and Rouder, 2011). The Bayes factor also increases with increasing effect size estimate d (Rouder and Morey, 2011).

The JZS Bayes factor is considered an objective method. In Bayesian statistics subjectivity comes into play when a prior distribution is selected. The JZS prior serves as an objective prior due to its lack of information and generic nature (Verhagen and Wagenmakers, 2014, Rouder et al., 2009). Being the objective choice, it has a relatively low influence on the posterior distribution. Instead, the posterior is mainly shaped by the likelihood of the data (van Ravenzwaaij and Etz, 2021, Morey and Rouder, 2011).

Equality-of-effect-size Bayes factor test

Bayarri and Mayoral, 2002 expands the classic JZS Bayes factor to the equality-of-effect-size Bayes factor. The equality-of-effect-size Bayes factor test accounts for two studies - the original and one replication study - by assessing whether Cohen's d, δ, can be expected to be identical for the two studies

$$H_0 : \delta_{orig} = \delta_{rep} \tag{3.45}$$

$$\Rightarrow \Delta\delta = \delta_{orig} - \delta_{rep} = 0 \tag{3.46}$$

Alternatively, the null hypothesis can be formulated as

$$H_0 : \ \tau^2 = 0 \tag{3.47}$$

Setting the expected difference or heterogeneity to zero is very restrictive. Instead, Bayarri and Mayoral, 2002 assume that the variance τ^2 between original and replication effect estimate is very small. Nevertheless, this does not make a difference when calculating the Bayes factor. As mentioned in chapter 3.1.2, the point null hypothesis approximates the interval hypothesis well enough.

In contrast to other described methods, the replication study is deemed successful if the null hypothesis is not rejected (Verhagen and Wagenmakers, 2014, Simonsohn, 2015). H_0 assumes the equality of effect size between original study and replication study. If the expected effect size for both studies is identical, replicability is declared.

The Bayes factor is therefore calculated reciprocally as BF_{01}.

The equality-of-effect-size Bayes factor test relies on a hierarchical Bayes model. The hierarchical Bayes model originates from a meta-analysis approach while assuming random effects (Marsman et al., 2017). It involves three distinct levels on which distributions are assumed a-priori and samples are drawn.

Level 1: Dependence of t-test statistic of effect size

The two-sample t-test which follows a non-central t-distribution with $df_i = n_{cond,i} + n_{crt,i} - 2$ degrees of freedom and non-centrality parameter $\delta_i\sqrt{n_i}$ is utilized for

$$T_i|\delta_i \sim NCT(\delta_i\sqrt{n_i}, df_i) \tag{3.48}$$

where $n_i = \frac{n_{cond,i}n_{crt,i}}{(n_{cond,i}+n_{crt,i})}$ and $i = orig, rep$.

Level 2: Relationship between original and replication effect size
Original and replication study are related by assuming a joint probability distribution $p(\delta_{orig}, \delta_{rep})$ with mean δ and some variance τ^2.

While it would be possible to model some kind of dependence between δ_{orig} and δ_{rep}, they are assumed to be exchangeable in the approach by Bayarri and Mayoral, 2002.

To ensure higher robustness, δ_i are drawn from a prior central t-distribution. The characteristic parameters are location parameter δ, scale parameter υ and degrees of freedom df

$$\delta_i|\delta \sim t(\delta, \upsilon^2, df) \tag{3.49}$$

Level 3: Prior distributions for hyperparameters
The hyperparameters - the parameters characterizing the distribution of the parameters - are also attributed a-priori distribution. The distributions are chosen to be uninformative.

In order to simplify the calculation of the Bayes factor, the non-central and Student t-distributions for $i = orig, rep$are modeled as a composition of normal and inverse Gamma distributions (Bayarri and Mayoral, 2002).

This concludes to

$$\begin{aligned}
&\text{I } T_i|\delta_i, \sigma_i^2 \sim N(\delta_i\sqrt{n_i\sigma_i^2}, \sigma_i^2)\\
&\text{II } \delta_i|\delta, \tau^2 \sim N(\delta, \tau^2),\\
&\qquad \sigma_i^2 \sim IGa(df_i/2, df_i/2)\\
&\text{III } p(\delta) \propto 1,\\
&\qquad \tau^2 \sim IGa(a+1, ak)
\end{aligned}$$

Since $\mathcal{E}(\tau^2) = k$ and $Var(\tau^2) = \frac{k^2}{(a-1)}$, k is the expected value for the heterogeneity and a indicates the degree of certainty or trust in this value. The higher a, the smaller the variance of τ^2. Both a and k need to be either subjectively chosen or estimated on the basis of previous replications. Bayarri and Mayoral, 2002 limit themselves to the former case.

According to transformations in Verhagen and Wagenmakers, 2014, the posterior predictive probability of the t-test value T_{rep} for the replication study can be expressed conditionally on t_{orig} as

$$T_{rep}|t_{orig}, \tau^2, \sigma_{orig}^2, \sigma_{rep}^2 \sim N\left(t_{orig}\sqrt{\frac{n_{rep}\sigma_{rep}^2}{n_{orig}\sigma orig^2}}, n_{rep}\sigma_{rep}^2\left(\frac{1}{n_{rep}} + \frac{1}{n_{orig}} + 2\tau^2\right)\right) \tag{3.50}$$

The Bayes factor requires the marginal likelihood for the null and alternative hypothesis.
For the former, $\tau^2 = 0$ is inserted. For the latter, the expression is integrated over $\tau^2 \in H_1$.
Both expressions are integrated across the study variances σ_{rep}^2 and σ_{orig}^2, which can be summarized in a variance vector $\sigma^2 = (\sigma_{orig}^2, \sigma_{rep}^2)$

$$B_{01} = \frac{\int p(t_{rep}|t_{orig}, \tau^2 = 0, \sigma^2)p(\sigma^2|t_{orig})\partial\sigma^2}{\int p(t_{rep}|t_{orig}, \tau^2, \sigma^2)p(\tau^2, \sigma^2|t_{orig})\partial\sigma^2\partial\tau^2} \tag{3.51}$$

where in addition to the distribution 3.50 it is known

$$p(\tau^2, \sigma^2 | t_{orig}) = p(\tau^2) \cdot p(\sigma^2_{rep}) \cdot p(\sigma^2_{orig} | t_{orig}) \tag{3.52}$$

The unconditional probabilities follow from the prior distribution assumptions listed above. The conditional probability is determined by the sample sizes for control and condition group in the original study, $df_0 = n_{cond} + n_{crt} - 2$ (Bayarri and Mayoral, 2002)

$$\sigma^2_{orig} | t_{orig} \sim IGa(\frac{df_0 + 1}{2}, \frac{df_0}{2}) \tag{3.53}$$

The Bayes factor is implemented by a Markov Chain Monte Carlo sampler sampling from the prior distributions of τ^2 and σ^2 and averaging for the null and alternative hypothesis respectively (Verhagen and Wagenmakers, 2014).

Replication Bayes Factor

To overcome some of the weaknesses of the JZS Bayes factor, Verhagen and Wagenmakers, 2014 developed a new method, called the replication Bayes factor.

The replication Bayes factor is designed to assess whether the outcome of a replication study is compatible with the effect size estimated by the original study.

The replication Bayes factor incorporates two viewpoints: the sceptic's as H_0 and the proponent's as H_1. The sceptic believes the effect size is zero

$$H_0 : \delta = 0 \tag{3.54}$$

while the proponent's hypothesis states that the parameter is distributed according to the posterior distribution from the original study with data x_{orig}

$$H_1 : \delta \sim p_{post}(\delta | x_{orig}) \tag{3.55}$$

Verhagen and Wagenmakers, 2014 constructed the replication Bayes factor to summarize the data sets in t-test values. Under H_0, the t-test estimated from the m replication studies, $x_{rep,i}$ for $i \in \{1, ..., m\}$, follows a central t-distribution. The Bayes factor is calculated as a quotient between the likelihood of $t_{rep,i}$ under H_0 and H_1

$$B_{01} = \frac{P(t_{rep,i} | H_0)}{P(t_{rep,i} | H_1)} = \frac{t_{df_i}(x_{rep,i})}{\int t_{df_i, \delta_i \sqrt{n_i}}(x_{rep,i}) p_{post}(\delta | x_{orig}) \partial\delta} \tag{3.56}$$

The posterior distribution for the original study, $p_{post}(\delta | x_{orig})$, follows from a flat non-informative two-sided prior distribution and an update corresponding to the data. It is a mixture of non-central t-distributions with non-centrality parameter $\delta_i \sqrt{n_i}$. The choice of the initial prior distribution can be considered arbitrary since the data update has such a large dominance that different prior distributions hardly weigh in (Verhagen and Wagenmakers, 2014). Since there exists no closed form for the posterior, it is approximated by a normal distribution with high accuracy (Verhagen and Wagenmakers, 2014, Ly et al., 2019).

BF_{01} itself can be approximated arbitrarily close by sampling j times from the approximation of the posterior probability function $p_{post}(\delta | x_{orig})$, inserting the drawn values in $t_{df_i, \delta_i \sqrt{n_i}}(x_{rep,i})$ and calculating the

mean

$$B_{01} \approx \frac{1}{j} \sum_{k=1}^{j} \left(\frac{t_{df_i}(x_{rep,i})}{t_{df,\delta_i\sqrt{n_i}}(X_{rep,i})} \right)_k \tag{3.57}$$

Evidence updating Bayes factor

Ly et al., 2019 develop the replication Bayes factor further, obtaining the evidence updating Bayes factor. It leverages a general property of the Bayes factor valid for the pooled data (x_{orig}, x_{rep}).

$$BF_{01}(x_{orig}, x_{rep}) = BF_{01}(x_{orig}) \cdot BF_{01}(x_{rep}|x_{orig}) \tag{3.58}$$

A simple transformation of equation 3.58 yields

$$BF_{01}(x_{rep}|x_{orig}) = \frac{BF_{01}(x_{orig}, x_{rep})}{BF_{10}(x_{orig})} \tag{3.59}$$

This simplifies the calculations greatly because sampling from the posterior distribution of the original study is rendered superfluous. All that is necessary is computing the Bayes factor for the original and the pooled data set. The evidence updating Bayes factor can not be used if the original study data is unavailable but for its aggregated format (Ly et al., 2019). Consequently, it can only be applied to the simulated but not to the multi-lab examples.

Criticism on Bayesian criteria

The evidence the Bayes factor offers is "inherently relative: what matters is which of the two hypotheses does best, not whether a specific hypothesis does well in an absolute sense" (Marsman et al., 2017, p. 4). This relative nature of the Bayes factor can be both boon and bane. The Bayes factor compares a specific null hypothesis to a specific alternative hypothesis, characterized by a certain prior distribution. Rejecting the alternative might not necessarily mean the null hypothesis is true, but rather that it fits the data relatively better than the alternative (Wagenmakers, Verhagen, et al., 2016). In general, Bayes factors tend towards accepting the null hypothesis if the alternative hypothesis does not provide a sufficiently better fit to the data even when the null hypothesis itself only matches poorly (Verhagen and Wagenmakers, 2014). While the Bayes factor might not allow us to establish the truthfulness of an hypothesis, it enables us to distinguish between relevant and irrelevant effect sizes. This is especially applicable to the zero-effect hypothesis. While a zero effect size might be highly unlikely or even non existent in reality (e.g. Gelman, 2018, Marsman et al., 2017), the null hypothesis can be considered the better choice than the alternative hypothesis if the effect size is very small and hardly detectable due to big random error. Gelman, 2015b nicely illustrates that measuring a small effect with high measurement error equals "trying to use a bathroom scale to weigh a feather — and the feather is resting loosely in the pouch of a kangaroo that is vigorously jumping up and down." The Bayes factor prevents us from picking up potential random noise which is masking the true weight of the feather. The Bayes factor might tend to accept the null despite it contradicting reality - the a true weight of the feather is non-zero. If the sample is large - greater than 500.000 in fact - the Bayes factor is well-suited to detect even tiny effects which speak against the null hypothesis (Rouder et al., 2009).

While the consistency of the *JZS Bayes factor* can be considered beneficial, it can also lead to wrong conclusions. If the assumption of a zero effect is not entirely true, the evidence against the null and for the alternative hypothesis will increase as the sample size increases. This means that even if the true effect has a negligible size caused by trivial and uninteresting noise, the JZS Bayes factor will let us believe otherwise. Hence, the JZS Bayes factor cannot distinguish well between unimportant and relevant effect sizes (Morey

and Rouder, 2011). It is also biased against small effect sizes and tends to accept the null hypothesis despite the effect size being significantly different from zero with a small or medium true size (Nelson et al., 2018).

The challenge (and most ardent criticism) in Bayesian statistics is the choice of the prior distribution which is frequently deemed too subjective (Hoffmann, 2017). Berger and Delampady, 1987 argue that even choosing the so called JZS prior - acknowledged as the objective default (Verhagen and Wagenmakers, 2014) - imposes some information "and as such cannot claim to be objective" (p. 319). However, according to Rouder et al., 2009 frequentist hypothesis testing is wrongfully comprehended as objective. The advantage of Bayesian methods is that "the elements of subjectivity are transparent rather than hidden" (p. 235). Another counter argument towards the accusation of too much subjectivity is the limited influence of the prior distribution. If a sample has a moderately large size, the Bayes factor is rather insensitive to priors with reasonable variation (Rouder et al., 2009, Wagenmakers, Verhagen, et al., 2016). Hence, the choice of prior plays a minor role in determining the value of the Bayes factor.

While some criticise the subjectivity of the JZS prior, others criticise its objectivity. According to the opinion of some Bayesian statisticians, relying on the JZS prior means to "make the Bayesian omelet without breaking the Bayesian egg" (Savage, 1961, p. 578). The JZS Bayes factor calculates a Bayes factor and benefits from its convenient features while avoiding its difficulties - the definition of subjective priors. Selecting a subjective and informative prior distribution is challenging and thus rarely performed (Hoijtink et al., 2016, p. 4). Hardly anyone calibrates a subjective prior by adapting the default prior to the specific research subject at hand. However, if some knowledge about the true parameter value is available, incorporating such information would strongly improve performance (Rouder et al., 2009). There exist several approaches to do so - e.g. Hoijtink et al., 2016 develops two criteria which allow determining the scale parameter r such that the prior distribution is well calibrated.

The JZS Bayes factor assumes that condition and control group have an identical variance σ^2. If this assumption is not fulfilled, the Bayes factor becomes inaccurate in its conclusions and suffers under the Behrens-Fisher problem (Wetzels et al., 2009).

The JZS Bayes factor does not relate the original to the replication study. It is rather "blind to earlier experiments" (Verhagen and Wagenmakers, 2014, p. 1459). Hence, potential distorting factors that influenced the original study cannot be detected. The unrelatedness between original and replication estimate also leaves one question open: if the original study is significant but the replication study speaks for the null hypothesis, which outcome should be valued higher in order to reach a conclusion on the existence of an effect?

The *equality-of-effect-size Bayes factor* is concerned with the difference between the two groups. While it bridges the gap between original and replication estimate, the exact value of the identical effect size is not taken into consideration, making it impossible to distinguish between zero and non-zero effects (Verhagen and Wagenmakers, 2014).

The equality-of-effect-size Bayes factor is highly dependent on the sample size. For small replication studies, it can have low statistical power and consequently have troubles delivering compelling evidence for either the null or alternative hypothesis (Bayarri and Mayoral, 2002, Verhagen and Wagenmakers, 2014). Another criticism regarding the equality-of-effect-size Bayes factor concerns potential distortion. Despite a discrepancy between original and replication effect estimate, the Bayes factor tends towards not rejecting the null hypothesis. A more moderate replication effect compared to an original estimate influenced by either publication bias or selective reporting does not differ sufficiently enough to result in a significant discrepancy between original and replication estimate (Simonsohn, 2015).

For all simulated scenarios, a certain degree of heterogeneity characterizes the effect estimates across stud-

ies. Hence, the null assumption to the equality-of-effect-size Bayes factor, $H_0 : \tau^2 = 0$, is untrue in reality. The bigger the heterogeneity, the less likely the equality-of-effect-size Bayes factor is to favor the null hypothesis despite it being theoretically true.

The term replication paradox describes the scenario in which the *replication Bayes factor* delivers evidence for the alternative hypothesis - and hence for replicability of the study - due to another than the expected reason. The reason for accepting the alternative hypothesis could lie in the effect having a non-zero size and not because it is well explained by the posterior distribution based on the original study. Vice versa, the replication Bayes factor delivers evidence for the null hypothesis despite the effect being non-zero because the original study and its posterior distribution insufficiently capture the replication estimate.
The replication paradox follows as a consequence to the relative nature of the Bayes factor - the replication estimate is more likely to originate from the posterior probability relative to from a zero mean normal distribution and not in an absolute sense (Ly et al., 2019).

The *evidence updating replication Bayes factor* lacks robustness and hence cannot handle bigger heterogeneity between studies particularly well (Ly et al., 2019).

Interval-based criteria

The posterior equal-tailed credibility interval as well as the highest density posterior interval (HDI) in isolation and in a hierarchical model utilize a two-sided JZS prior and hence provide an non-directional assessment, $H_1 : \theta_0 \neq 0$. Directionality can be accounted for by visually analyzing the credibility intervals since the criteria retain information regarding the sign of the effect estimate.

Posterior equal-tailed credibility interval - in isolation
Credibility intervals are the Bayesian counterpart to frequentist confidence intervals. According to the equivalence between hypothesis testing and confidence intervals credibility intervals can be leveraged to assess the evidence for a zero or non-zero effect size (Fahrmeir et al., 2016, Kauermann and Hothorn, 2020). A replication study is considered a success if the $1 - \alpha$ credibility interval does not entail zero (Marsman et al., 2017).

The benefit of Bayesian credibility intervals - and the difference to frequentist confidence intervals - is their easy and intuitive interpretation: the true parameter value lies with a probability of $1 - \alpha$ within the interval (Kauermann and Hothorn, 2020, Marsman et al., 2017).

A $1 - \alpha$ Bayesian credibility interval is constructed based on the posterior density of the parameter of interest. For the posterior equal-tailed credibility interval, the probabilities for the parameter to lie below the lower or above the upper interval boundary are identical. The boundaries are set to the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantile of the posterior distribution (Makowski et al., 2019). Such credibility intervals are constructed for each replication study in isolation.

The default choice for a prior distribution to compute the posterior density is the uninformative JZS prior. However, similar to the JZS Bayes factor, the prior distribution has only limited influence on the posterior distribution if the collected data entails sufficient information.

Highest density posterior credibility interval - in isolation
The HDI indicates that the interval spans over the range of parameter values Δ_{HDI} which are allotted the highest posterior probability under all possible values of δ (Kauermann and Hothorn, 2020, Marsman et al., 2017)

$$\forall_{\delta \in \Delta_{HDI}} \; p_{post}(\delta) \geq \forall_{\delta \notin \Delta_{HDI}} \; p_{post}(\delta) \tag{3.60}$$

The interpretation of the HDI equals the standard credibility interval. The definition of successful replications - zero is not an element of the interval - also remains the same.

Highest density posterior credibility interval - hierarchical model

Marsman et al., 2017 also construct highest density posterior credibility intervals relying on a hierarchical model which accounts for variation in treatment effect and allows information sharing across replication studies.

In accordance with the JZS Bayes factor in chapter 3.2.4, the observations for condition and control group are assumed to originate from a normal distribution, characterized by grand mean μ, effect δ and variance σ^2 (Marsman et al., 2017, Rouder et al., 2009).

The hierarchical model is identical to the one implemented in the equality-of-effect-size Bayes factor in chapter 3.2.4. For the parameters characterizing the group distribution, (μ, σ^2), the choice of uninformative prior distributions, outlined in chapter 3.2.4, is maintained

$$p(\mu) \propto 1 \tag{3.61}$$

$$p(\sigma^2) \propto \frac{1}{\sigma^2} \tag{3.62}$$

For the group-level parameters, (δ, τ^2) uninformative priors are assumed

$$\tau \propto \frac{1}{\tau^2} \tag{3.63}$$

$$\delta \sim Cauchy(0, r) \tag{3.64}$$

The default choice for the Cauchy scale parameter r remains $\frac{\sqrt{2}}{2}$ (Marsman et al., 2017).

The hierarchical parameter estimation results in a posterior probability distribution of the effect size δ which can consequently be used to construct a $1-\alpha$ highest density credibility interval for each replication studies while acknowledging heterogeneity. Its interpretation again remains identical to the two other interval-based criteria.

Criticism on interval-based criteria

In many research fields, zero effects do not exist (Gelman, 2018). The task is rather to distinguish non-zero effects whose deviation from zero is accounted for by uninteresting and uninformative noise, from those effects whose deviation can be explained and exploited theoretically. Checking whether zero is in the credible interval or not does not acknowledge non-zero effects due to noise. The criticism regarding the choice of uninformative prior distribution brought up against Bayes factors remains for both versions of the highest density credibility interval as well as the posterior credibility interval (e.g. Marsman et al., 2017). The intervals could be constructed more precisely if the prior distribution was customized to the scenario at hand.

The hierarchical credibility interval has three benefits over the other two interval-based criteria built in isolation,

1. Each replication study contributes to the estimation of the overarching parameters, expected grand mean θ and variance in treatment effects across studies, τ^2.

2. Assuming an overarching distribution lead to regression towards the mean, i.e. the shrinkage of individual effect sizes towards the grand mean.

3. Relying on more than one replication study reduces uncertainty, i.e. diminishes the span of the credibility interval.

Small telescope criterion

The small telescope criterion has been developed by Simonsohn, 2015. Its main conclusion differs from the conclusions obtained under alternative criteria. Small telescope does not establish the existence of a significantly non-zero effect but rather assesses if the "original evidence suggesting a theoretically interesting effect exists" can be trusted (Simonsohn, 2015, p. 560). More precisely, the small telescope criterion asks if the effect in the replication study is too close to a zero effect that it would not have been detectable in the original study, i. e. would not have obtained the original significant p-value while guaranteeing a minimum of statistical power.
The name explains the criterion in an analogy from astronomy. If we zoom in using a larger telescope - the replication study - and are unable to find a significant result, this indicates that the zoomed-out result with a smaller telescope - the original study - might not be accurate.

Hence, the small telescope criterion is based on detectability. Detecability in the context of hypothesis testing equals statistical power. As a reminder, statistical power is defined as the probability for a significant effect given that the true effect size is non-zero. Detectability is therefore calculated as the probability of accepting the alternative hypothesis when it is in fact true and thus, the counter-probability to the probability of a Type-II error. According to Simonsohn, 2015, studies with a statistical power of 33% and below are deemed severely underpowered.
The corresponding effect size, $\delta_{0.33}$, is determined by

$$\mathcal{P}(t_{orig} \geq t_{0.95,n-1}|\delta_{0.33}) = 0.33 \tag{3.65}$$

A replication effect, δ_{rep}, is judged based on the power it would attribute to the original study. It is considered too small an effect if it is significantly smaller than $\delta_{0.33}$. Hence, a one-sided t-test with hypotheses is performed

$$H_0 : \delta > \delta_{0.33} \tag{3.66}$$

$$H_1 : \delta \leq \delta_{0.33} \tag{3.67}$$

Accepting H_1 does not disprove the existence of an effect. It rather implies "that sampling error alone is an unlikely explanation for why the replication resulted in such a categorically smaller effect size." (Simonsohn, 2015, p. 567). This in turn can lead to different conclusions (Simonsohn, 2015).

- The original and replication study apparently do not concern the same effect, some (hidden) mediator might have distorted the replication study.

- The original estimate suffers under publication bias or selective reporting and is a severe overestimation of the true effect size.

Criticism on small telescope

The small telescope criterion exhibits one distinct characteristic. It "evaluates the original study's design rather than its result" (Simonsohn, 2015, p. 566). While the theory behind the research hypothesis might remain correct and provable if $H_1 = \delta \leq \delta_{0.33}$ is accepted, the original study is discredited as poor evidence (Simonsohn, 2015). The small telescope criterion solely focuses on the null hypothesis and does not account for potential results under the alternative hypothesis (Wagenmakers, Verhagen, et al., 2016). Being a p-

value based criterion, it can also not provide evidence for, only against the null hypothesis (Verhagen and Wagenmakers, 2014).

Sceptical p-value

The sceptical p-value was developed by Held, 2020. In an analysis of credibility, the sceptical p-value indicates up to which confidence level the confidence interval constructed around the replication effect estimate opposes the sceptic's assumption of a zero effect.

Held, 2020 argues that the two main advantages of the sceptical p-value are that

- it accounts for both original and replication study and its estimated effect sizes and
- by disregarding the uncertainty of the original study, the sceptical p-value equals the ordinary p-value.

One requirement to the sceptical p-value is the normal distribution of the effect size. This can be guaranteed by measuring the effect as correlation ρ and applying the Fisher transformation to its empirical estimate. This different effect size estimate is denoted by $\hat{z} = tanh^{-1}(\widehat{\rho})$ (see chapter A.4). Throughout the calculations, the original effect estimate $\widehat{z}_{orig}$ and its variance σ^2_{orig} are assumed to be known. Both conditions are fulfilled thanks to the Fisher transformation and its easily derived variance. A classic one-sided t-test is performed on the original data, $t_{orig} = \frac{\widehat{z}_{orig}}{\sigma_{orig}}$.

The baseline for calculating the sceptical p-value comprises a significant original effect estimate and a sceptical prior distribution. The sceptical prior distribution is a normal distribution $N(0, \tau^2)$ which describes the sceptic's perspective: the true effect is zero with some noise around it. The first step is to find τ^2, i.e. to assess how sceptical the prior distribution would need to be in order for the posterior confidence interval - updated according to the original study - to entail zero and thus be insignificant. While the posterior is fixed, the prior variance is chosen flexibly - ending up with a reverse Bayesian approach.

Based on the original study, a conventional $1-\alpha$ prior confidence interval is constructed with lower boundary L and upper boundary U. To consider one-sided alternative hypothesis, the posterior confidence interval is fixed to have a lower limit $L_{post} = 0$ and an arbitrary upper limit U_{post}.

With this fixed target posterior confidence interval given, the credibility interval for the sceptical prior distribution is symmetrical around zero with boundaries $+/-S$, the scepticism limit, for an arbitrary choice of significance level α (Matthews, 2018).

$$S = \frac{(U-L)^2}{4\sqrt{UL}} \tag{3.68}$$

Held, 2019 proves that the variance for the sceptical prior distribution can be calculated by

$$\tau^2 = \frac{\sigma^2_{orig}}{\frac{t^2_{orig}}{z^2_{1-\alpha}} - 1} \tag{3.69}$$

According to formula 3.69, the variance of the sceptical prior is large if the original effect size is only barely significant. The more significant the original effect size and hence greater t_{orig}, the narrower the prior interval.

Replication success is subsequently defined by determining whether the replication effect size is aligned or in conflict with the sceptical prior distribution. If the former applies, the original study is deemed credible and the replication successful. The fit between the replication study and the sceptical prior is assessed by

calculating the sceptical p-value p_s. The sceptical p-value indicates up to which confidence interval $1 - p_s$ the confidence interval - constructed based on the replication estimate - is significant (Held, 2020). The sceptical p-value, $p_s = 1 - \Phi(z_{p_s})$ - Φ is the standard normal cumulative distribution - is the solution to an equation incorporating the one-sided test statistics and variances for the original study, t_{orig} and σ^2_{orig}, and for the replication study, $t_{rep,i}$ and $\sigma^2_{rep,i}$ ($i \in \{1, ..., m\}$), as well as the quantile z_{p_s}. The ratio between original and replication variance is denoted by $c = \frac{\sigma^2_{orig}}{\sigma^2_{rep,i}} = \frac{n_{rep,i}-3}{n_{orig}-3}$

$$\left(\frac{t_0^2}{z_{p_s}^2} - 1\right)\left(\frac{t_{rep,i}^2}{z_{p_s}^2} - 1\right) = c \tag{3.70}$$

The derivation of the formula relies on the prior-predictive variance for $\widehat{z}_{rep,i}$ - the sum of prior variance and random sampling error $\tau^2 + \sigma^2_{rep,i}$ - and a classic t-test statistic on the Fisher-transformed correlation (see appendix in Held, 2020).

The solution follows from formula 3.70 with several equivalent transformations

$$z_{p_s}^2 = \begin{cases} \frac{t_H^2}{2} & c = 1 \\ \frac{1}{c-1}\left(\sqrt{t_A^2(t_A^2 + (c-1)t_H^2)} - t_A^2\right) & c \neq 1 \end{cases} \tag{3.71}$$

with $t_A^2 = \frac{t^2_{orig}+t^2_{rep,i}}{2}$ and $t_H^2 = \frac{2}{t_{orig}^{-2}\, t_{rep,i}^{-2}}$.

The calculated sceptical p-value is subsequently compared to a pre-specified significance level α. The replication is deemed successful if $z_{p_s} \leq z_{1-\alpha}$.

For replication success, it is necessary that both original and replication effect size estimates are significant. This follows from

$$max\,(\,p_{orig}, p_{rep}\,) \leq p_s \tag{3.72}$$

$$p_s \leq \alpha \tag{3.73}$$

$$\Rightarrow p_{orig},\ p_{rep} \leq \alpha \tag{3.74}$$

Hence, both studies have to provide sufficient evidence on their own before contributing to the overall assessment of replicability (Held, 2020, Held et al., n.d.).

The smaller the estimated replication effect size, the bigger the ratio c due to

$$c = \frac{\sigma^2_{orig}}{\sigma^2_{rep,i}} = \frac{t^2_{rep,i}/t^2_{orig}}{\widehat{z}^2_{rep,i}/\widehat{z}^2_{orig}} \tag{3.75}$$

This in turn leads to a bigger sceptical p-value and the convenient feature that "[r]eplication studies with smaller effect estimates than the original estimates are considered less credible" than replication estimates with equal or greater size (Held, 2020, p. 432). To refer back to the sceptic's perspective, small replication effects do not convince the sceptic that the significant original effect size is indeed trustworthy and portrays the true effect (Held et al., 2020).

Criticism on sceptical p-value

The requirement for the sceptical p-value - the significance of both original and replication effect estimates - strongly limits its applicability since the insignificance of replication estimates is a rather common occurrence (e.g. Schweinsberg et al., 2016, Klein et al., 2014, Wagenmakers, Beek, et al., 2016).

Snapshot Bayesian hybrid meta-analysis method

The snapshot Bayesian hybrid meta-analysis method, in short snapshot hybrid, has been developed by Van Aert and Van Assen, 2017 to quantify the evidence for a zero, small, medium or large effect size. It calculates the posterior probability of the effect size at said four distinct snapshot levels based on the observed data and an uninformative prior while correcting for potential publication bias in the original study. Snapshot hybrid makes only a few assumptions,

- Normal distribution of the effect size
- Significance of the original effect size
- Consideration of effect size as fixed

To guarantee the validity of the first assumption, the effect size is measured as correlation ρ and Fisher transformed to z. The variance can be obtained by the known formula $\sigma^2 = \frac{1}{n-3}$.

The likelihood of the effect estimates of both original, z_{orig}, and replication study, $z_{rep,i}$ ($i \in \{1, ..., m\}$), are dependent on the true effect size z. The combined likelihood for $(z_{orig}, z_{rep,i})$ is computed as a product of the individual likelihoods - normal distributions - since the studies are independently conducted

$$L(z) = f\ (z_{orig}, z_{rep,i}|z) = f\ (z_{orig}|z) f\ (z_{rep,i}|z) \tag{3.76}$$

In order to correct for potential publication bias, the likelihood for the original effect size z_{orig} is truncated at the critical level $z_{1-\alpha}$. Since it is known that the original effect size is significant, the estimate's value must exceed a critical boundary $z_{1-\alpha}$ calculated based on classical hypothesis testing. All values below the critical boundary are assigned a zero density. The truncated likelihood is constructed by dividing the likelihood by the power of the test statistic, $\beta = \Phi(\frac{z_{1-\alpha}-z}{\sigma})$

$$f_{trunc}(z_{orig}|z) = \frac{f(z_{orig}|z)}{1-\beta} \tag{3.77}$$

A-priori, all n_p choices are attributed an equal probability of $p_i = \frac{1}{n_p}$. Hence, the posterior probability is directly proportional to the likelihood $L(z)$. Its final value is calculated by normalizing the likelihood by the sum of likelihoods for each parameter choice according to the law of total probability (Fahrmeir et al., 2016)

$$p_{post}(z_i) = \frac{L(z_i)}{\sum_{i=1}^{n_p} L(z_i)} \tag{3.78}$$

The snapshot hybrid posterior outcome can be directly interpreted as relative probability that z has a certain size z_i. Transferring the Bayes factor classification according to Jeffreys, 1961 to posterior probabilities, a posterior probability greater than 0.75 is considered substantial evidence for the respective effect size. All other results are regarded inconclusive. A simulation study in Van Aert and Van Assen, 2017 illustrates that this adopted threshold guarantees that the wrong decision rate never exceeds 0.065.

Criticism on snapshot hybrid

The criticism and counterarguments for snapshot hybrid will be discussed in more detail in chapter 5.2.

3.2.5 Methods for multiple replications

Berkson's interocular traumatic test

Before calculating any criterion, it is helpful to obtain an overview over the effect estimates of all replication studies in comparison to the original estimate. Thereby, the effect size is measured both as a raw difference between condition and control mean and as Cohen's d.

This graphical exploration is called the Berkson's interocular traumatic test. It allows a visual assessment on whether "the data are so compelling that conclusion hits the reader straight between the eyes" (Verhagen and Wagenmakers, 2014, p. 1470).

Non-central confidence interval

Non-central confidence intervals to a significance level α are constructed if the null hypothesis is false, i.e. the effect size is likely to be significantly different from zero.

Referring back to chapter 3.2.4, the non-centrality parameter is determined by

$$c = \frac{\theta - \theta_0}{\frac{\sigma}{\sqrt{n}}} = \delta \sqrt{n} \tag{3.79}$$

with $n = \frac{n_{cond} n_{crt}}{(n_{cond} + n_{crt})}$.
It is empirically approximated as

$$\widehat{c} = d \sqrt{n} = t \tag{3.80}$$

Based on the non-centrality parameter c, a confidence interval with upper limit c_u and lower limit c_l is constructed such that

$$\mathbb{P}(c_u \leq c \leq c_l) = 1 - \alpha \tag{3.81}$$

with a significance level α. The limits are computed by exploiting the Inversion Confidence Interval Principle (Kelley et al., 2007). Assuming t_{obs} denotes the observed t-test statistic for the pooled data, the limits are determined such that they attribute t_{obs} a minimum and maximum cumulative probability to be observed.

$$t(t_{obs}|c_u) = \alpha \tag{3.82}$$

$$t(t_{obs}|c_l) = 1 - \alpha \tag{3.83}$$

where t stands for the cumulative function to the non-central t-distribution with non-centrality c_u and c_l respectively (Kelley et al., 2007).

The monotonic transformation between Cohen's d and the non-centrality parameter allows us to construct the confidence interval of the former based on formula 3.81 (Smithson, 2003, Kelley et al., 2007).

$$\mathbb{P}(c_u \sqrt{\frac{1}{n}} \leq \delta \leq c_u \sqrt{\frac{1}{n}}) = 1 - \alpha$$

The non-central $1 - \alpha$ confidence interval for the weighted standard mean difference over all replication studies is easily calculated using the steps outlined above on the pooled data set (Klein et al., 2014). Replication success is defined in accordance with the symbiosis between confidence intervals and classic hypothesis

testing (Fahrmeir et al., 2016). A confidence interval entailing zero implies a lack of replicability and thus replication failure.

Criticism on non-central intervals
Due to the pooling, the non-central interval does not account for any heterogeneity between replication studies. Its interpretation is less intuitive than for its Bayesian counterpart, the credibility interval. $1 - \alpha$ does not equal the probability for the true effect to lie within the interval. It rather describes the number of the cases, the confidence interval constructed on the data sample covers the true effect size (Fahrmeir et al., 2016).

Bayesian evidence synthesis

In Bayesian evidence synthesis the aim is to answer to which extent the pooled data supports the existence or non-existence of a certain effect (Verhagen and Wagenmakers, 2014, Scheibehenne et al., 2016). The criterion can be utilized iteratively, adding more and more data to its calculation as replication studies are performed and analyzed in different timelines (Scheibehenne et al., 2016). The condition or assumption inherent to Bayesian evidence synthesis - as to every meta-analysis approach - is the exchangeability of studies. Bayesian evidence synthesis assumes that the effect sizes across replication studies stem from one distribution with a fixed mean. Variation between studies are solely attributed to sampling error.

The Bayesian evidence synthesis is fairly easy to calculate. The multiple replication studies are considered as one big replication experiment on which the default JZS Bayes factor criterion (see chapter 3.2.4) is implemented.

Criticism on evidence synthesis
Bayesian evidence synthesis imposes a rather stark simplification on the complex issue of replicability. It disregards any sort of heterogeneity or context dependency between the studies. Bayesian evidence synthesis does not acknowledge the influence of hidden mediators which can vary across replication studies (Anderson and Maxwell, 2016). Depending on the size of the variation in treatment effects, this neglect can be too generalizing.

Due to the pooling, Bayesian evidence synthesis is also not suited to determine any relationship between replication and original study which is subjected to potential distortions.

Fixed-effect meta-analysis JZS Bayes factor

Another Bayes factor applicable to multiple replication studies simultaneously is the fixed-effect meta-analysis Bayes factor. While Bayesian evidence synthesis assumes a constant variance for all replications, the fixed-effect meta-analysis Bayes factor allows for a variable variance for each replication study (Rouder and Morey, 2011).

As the name suggests, the fixed-effect meta-analytical Bayes factor criterion incorporates fixed effects (Wagenmakers, Beek, et al., 2016, Verhagen and Wagenmakers, 2014). It is constructed as a generalization of the JZS Bayes factor in order to apply to m independent replications and the original study. It calculates the overall likelihood as the product of the individual likelihoods. Again, the relevant information is captured by the t-value since the t-test is sufficient (Rouder and Morey, 2011, Kauermann and Hothorn, 2020)

$$f(t_1, ..., t_m|\delta) = \prod_{i=1}^{m} f(t_i|\delta) \tag{3.84}$$

The prior distributions for the parameters remain the same - a JZS prior $p(\delta)$. Hence, the meta-analytic

Bayes factor is obtained by

$$BF_{01} = \frac{\prod_{i=1}^{m} f(t_i|\delta = 0)}{\int \prod_{i=1}^{m} f(t_i|\delta)p(\delta)\partial\delta} \tag{3.85}$$

The interpretation for both Bayes factor versions remains identical to previous implementations of Bayes factors.

Criticism on meta-analysis

Sceptics oppose the popularity of meta-analysis in combination with direct replication studies due to the apples-and-oranges argument. Given that direct replication studies differ from the original study, the studies are considered unsuited to be uniformly analyzed - it is like comparing apples and oranges (Hunter, 2001). According to Verhagen and Wagenmakers, 2014, meta-analysis disregards the sequential logical order of original and replication study. The main criticism itemized in many papers (e.g. Held, 2020) is this assumption of exchangeability between original and replication studies. The information across all studies is equally leveraged to gather evidence for or against the null hypothesis without any hierarchical ordering (Verhagen and Wagenmakers, 2014). This does not mirror the naturally asymmetric question asked when performing a replication study. We are concerned to find out if the original study withstands the replication attempts and thus can be considered valid evidence for an effect (Held, 2020).

Another disadvantage of meta-analysis is its inability to handle publication bias. It suffers the common "Garbage in - Garbage out" dilemma (Ferguson and Heene, 2012). Publication bias leads to only significant studies with high effect sizes being published. Calculating some sort of mean or average effect size from these biased studies consequently results in an overestimation and overconfidence in the true effect size. Publication bias is thus regarded as "a major threat to the validity of meta-analyses" (Van Aert and Van Assen, 2017, p. 2). It could be argued that in order to perform stable and convincing meta-analysis, a series of trustworthy replication studies - not subjected to any distortions - is required. Thus, replication and meta-analysis go hand-in-hand and are two sides of one coin (Eden, 2002). Nevertheless, simply regarding meta-analysis as the solution for the replication crisis creates a wrong sense of simplicity and security (Ferguson and Heene, 2012).

In fixed-effect meta-analysis, it is assumed that each replication study is a direct replication and its observations are sampled from one and the same underlying fixed-effect distribution. Hence, fixed-effect meta-analysis does not account for any heterogeneity between the different studies (Bonett, 2009).

An alternative meta-analysis approach would model random effects. However, while random-effect meta-analysis considers variation in treatment effect across different replication studies, it assumes that the effect sizes per study are randomly sampled from an overarching superpopulation of possible effect sizes. This superpopulation of effect sizes follows a normal distribution. According to Bonett, 2009, "the critical random sampling assumption of the [random effect, AN] methods will almost never be satisfied in practice" (p. 226). However, random sampling is required so that the inference based on a certain number of replication studies can be generalized to the overall effect.

Focusing on the fixed-effect meta-analysis Bayes factor in particular, the choice of objective JZS prior can be doubted due to the same criticism noted against the JZS Bayes factor.

3.3 Data sets - Multi-lab analysis

As a consequence of the replication crisis, an increasing number of researchers have performed large replication studies reassessing seemingly established scientific findings (Held et al., 2020). These multi-lab replication studies involve many labs across the world. They each conduct the identical experiment to the one described in the original paper and analyse the data with the same methods (e.g. Klein et al., 2014, Open Science Collaboration, 2015, Schweinsberg et al., 2016, Wagenmakers, Beek, et al., 2016).

According to Simons, 2014, "[d]irect replication by other laboratories is the best (and possibly the only) believable evidence for the reliability of an effect." (p. 76)

The results from multi-lab replications have corroborated the claims by Ioannidis, 2005 that "most research findings are false." (p. 696) Many firmly established and highly accredited studies were proven to lack replicability and to have been heavily influenced by questionable practices and/ or systematic distortions (e.g. Open Science Collaboration, 2015, Klein et al., 2014).

The methods selected to determine replication success will be implemented on both the multi-lab examples and a simulation study. We have utilized three real data sets covering relevant scenarios.

- Facial feedback hypothesis (Strack et al., 1988): replication studies and their assessment by Wagenmakers, Beek, et al., 2016 have shown a lack of replicability.
- Imagined contact hypothesis (Turner et al., 2007): the analysis of the replication studies by Klein et al., 2014 do not conclusively refute or corroborate replicability claims.
- Sunk costs hypothesis (Oppenheimer et al., 2009): the replication studies by Klein et al., 2014 substantiate the original study and thus speak for replicability.

The simulation study also entails three scenarios, as described in chapter 3.4.

Wagenmakers, Beek, et al., 2016 conducted a multi-lab analysis on the facial feedback hypothesis across 17 different labs. Overall, 1958 individuals were observed.

Klein et al., 2014 performed a multi-lab analysis, replicating 13 classic and contemporary psychological effects in 36 samples and settings. The endeavour included replications of the imagined contact and sunk cost hypothesis, for which 6330 and 6336 individuals respectively were observed.

Different methods were selected by the respective authors to conclude replication success or failure. The chosen replication methods are,

- Wagenmakers, Beek, et al., 2016
    1. Random effect meta analysis
    2. One-sided JZS Bayes factor
    3. Replication Bayes factor
- Klein et al., 2014
    1. 95% or 99% non-central confidence intervals for weighted mean
    2. 95% or 99% central confidence intervals for unweighted mean
    3. Null hypothesis testing on individual studies

4. Null hypothesis testing on aggregated data across all replication studies

3.3.1 Facial feedback hypothesis

The facial feedback hypothesis can be found in most introductory psychology books and counts as the prime example on how facial expressions can influence affective responses - independent of the actual presence of an emotion (Strack et al., 1988, Wagenmakers, Beek, et al., 2016). According to the hypothesis, participants who mimic a smile by holding a pen with their teeth show a more intense response to humor than participants who mimic a frown by holding a pen with their lips. The perception of humor is measured on a Likert scale from 0 (not funny at all) to 9 (very funny) by indicating how funny the participants found each of 4 cartoons. The average rating per participant is calculated as mean of the individual scores.

Measure	value
Number of observations	64
Mean condition (smile)	5.14
Mean control (frown)	4.32
t-statistic	1.85
Cohen's d *	0.46

Table 3.1: Results for facial feedback hypothesis in the original paper

(*) Cohen‘s d was not indicated in the original paper. It was determined by leveraging the relationship between t-statistic and Cohen's d, $t = \sqrt{n}\delta$.

Wagenmakers, Beek, et al., 2016 conducted 17 independent direct replication studies in which they found that the facial feedback hypothesis is not replicable based on the selected replication criteria.

One notable difference in the experimental set-up between the original and the replication studies is the webcam monitoring conducted during the replications. Some researchers have argued that this change might lead to different reactions of the participants and hence distort the findings towards a non-significant effect size (e.g. Gollwitzer and Schwabe, 2020).

For more information on the replication studies, the reader might refer to the original paper by Strack et al., 1988 and the replication paper by Wagenmakers, Beek, et al., 2016.

3.3.2 Imagined contact hypothesis

Several studies (e.g. Turner et al., 2007) have suggested that the mental simulation of a positive interaction with a member of an ethnic minority suffices to improve attitudes towards them. Husnu and Crisp, 2010 extends this hypothesis by showing that "imagined contact" enhances intentions to actively engage with the respective minority in the future. The study was performed on 33 British Non-Muslim undergraduate students who are randomly allotted to either a no-contact control scene or a contact scene group. The scenarios are laid out by simple instructions:

- non-contact control scene: "I would like you to take a minute to imagine you are walking in the outdoors. Try to imagine aspects of the scene about you (e.g., is it a beach, a forest, are there trees, hills, what's on the horizon)."

- contact scene: "I would like you to take a minute to imagine yourself meeting a British Muslim stranger for the first time. During the conversation imagine you find out some interesting and unexpected things about the stranger."

To assess the degree of intentions, the study asks four questions regarding the likelihood of future interaction with the minority. The participants indicate their answer on a 9-point Likert scale, from 1 (not at all) to 9 (very much). The average willingness is estimated as mean of the four questions. The result shows a significantly greater intention to meet British Non-Muslims in the future amongst the contact scene group compared to the control group.

Measure	value
Number of observations	33
Mean condition (contact)	5.93
SD condition (contact)	1.67
Mean control (no contact)	4.69
SD control (no contact)	1.26
Cohen's d	0.86
t-statistic	2.39

Table 3.2: Results for imagined contact hypothesis in original paper

The replication studies are considered successful according to some chosen replication metrics but failed according to others. For more information on the experimental set-up and results, the reader might refer to the original paper by Turner et al., 2007 and the replication paper by Klein et al., 2014.

For this thesis, additional 8 observations in total across multiple replication studies are excluded due to missing values in the original data.

3.3.3 Sunk costs hypothesis

Oppenheimer et al., 2009 explored how sunk costs, meaning costs that have already been paid and cannot be refunded, influence behaviour and decision making. The participants are faced by the following scenario: "Imagine today is a big game for your favorite football team which you plan on attending. However, it is bitter cold and freezing."
"Would you still go to the match for which..."

- "you got your tickets for free from a friend?"
- "you bought your tickets already without the possibility to return them?"

The participants are randomly allotted one of the scenarios and are instructed to indicate their willingness to attend the game based on a 9-point Likert scale, ranging from 1 (not motivated at all) to 9 (very motivated). Participants who imagined having paid for their tickets were more likely to express a higher willingness to attend the game than those imagining to have gotten the tickets for free.

Measure	value
Number of observations	213
Mean condition (paid ticket)	7.46
Mean control (free ticket)	6.93
F-statistic	2.74
Cohen's d *	0.23

Table 3.3: Results for sunk cost hypothesis in original paper

(*) indicates that Cohen's d was calculated and not included in the original paper.

Klein et al., 2014 repeated this experiment without changes in the multi-lab setting. The replication studies clearly replicate the hypothesis according to all replication metrics implemented by the authors.

For detailed information on how the replication studies for both experiments were performed, the reader might refer to the original paper by Oppenheimer et al., 2009 and the replication paper by Klein et al., 2014.

For this thesis, additional 14 observations in total across multiple replication studies are excluded due to missing values.

3.4 Simulation study

3.4.1 Purpose

The implementation of a simulation study is a central tool in assessing the performance of statistical methods. Simulation studies can be viewed as experiments in a controlled setting in which the true parameters for the data generating process are known. They are conducted by pseudo-random sampling from a specified distribution (Morris et al., 2019).

In this thesis, a simulation study is conducted in order to explore the performance of the selected criteria determining replication success in three different scenarios. These three scenarios are influenced by different distorting factors to varying degrees.

We limit ourselves to the two main distortion phenomena, publication bias and researcher degrees of freedom in combination with selective reporting, due to their frequent occurrence in reality and on their relevance for the method assessment. Additionally, we assume variation in treatment effects (see chapter 3.2.2).

3.4.2 Methodological approach

In order to construct a simulation study mirroring the reality as accurately as possible, the data generation is conducted with parameters set according to the multi-lab data sets as elaborated in section 3.3.

Three scenarios with the following characteristics are simulated:

1. *null effect with publication bias*: The original study obtains a false positive significant effect. In reality, the parameter is a null effect.
2. *a positive effect with selective reporting*: The original study indicates a significant effect which is greater than the effect size calculated in the replication studies. The original effect size is an overestimation of the true effect size.
3. *a positive effect without distortion*: The original effect size is an adequate estimation of the true parameter and lies within the same range of effect sizes estimated by the replication studies.

Scenario 1 is expected not to replicate given that the original effect size is simply a false positive.
Scenario 2 is expected to be wrongly classified as replicable by some criteria and rightly classified as not replicable by others that take the difference in estimate values into account.
In order to correctly classify scenario 2 as not replicable, a criterion must distinguish between parameter heterogeneity and variance due to different underlying distribution parameters across multiple studies.
Scenario 3 is expected to replicate since the original effect size estimates the true parameter without the influence of any distorting factors.

The heterogeneity between studies is simulated by randomly sampling the effect size parameter θ_i for each replication from a normal distribution $N(\theta_0, \tau^2)$. This distribution - or more its parameters - must be known a-priori. The mean θ_0 of this normal distribution is the mean of the true effect size. The heterogeneity τ^2 describes how much the individual true effect sizes θ_i deviate from this mean θ_0. It is easy to imagine study settings in which the true effect varies in size due to e.g. socioeconomic factors prevalent in the group of participants (e.g. Griffin et al., 2002).

Each study consists of a condition group of size n_{cond} and a control group of size n_{crt}. We simulate the control and condition samples to be independent of each other and have an identical variance, σ^2.

The control group sample is generated by adding a random error $\epsilon \sim N(0, \tau^2)$ to the average value β_0. For the condition group, the variable treatment effect θ_i is added additionally. Since we are exclusively interested in the difference between control and condition group, β_0 can be arbitrarily chosen. It will be neglected in most of the following formulas.

The steps in the simulation of a data set are,

1. Sample θ_i from the prespecified normal distribution
2. Condition group: sample observations $X_i \sim N(\beta_0 + \theta_i, \sigma^2)$, iid for $i \in \{1, ..., n_{cond}\}$.
3. Control group: sample observations $X_i \sim N(\beta_0, \sigma^2)$, iid for $i \in \{1, ..., n_{crt}\}$.

For each scenario, an original study and a number of replication studies of a certain size for both, condition and control group, are simulated. The simulation of the control group remains unchanged across all scenarios and will not be mentioned explicitly in the scenario explanations further on. The exact proceedings on how to simulate the condition group based on θ_i and s^2 depend on the scenario type. It will be explained in the following sections.

Scenario 1: null effect with publication bias

Publication bias, as discussed in section 2.1.2, means that given a certain research hypothesis, corresponding studies claiming a significant effect are more likely to be published than corresponding studies supplying evidence for a null effect. Thus, publication bias leads to a distorted impression, assuming the existence of a significant effect where either none is existent or it is insignificantly close to zero (e.g. Gelman, 2015a, Zwaan et al., 2018).

For scenario 1, θ_0 is set to 0, to simulate a null effect.
θ_i for all simulated condition studies, $i = 1, ..., m$, are randomly drawn from $N(0, \tau^2)$.

The condition study observations are subsequently simulated by drawing from a normal distribution with mean θ_i and variance σ^2. To simulate publication bias, we require the original study to have a significant effect. Hence, the original study sample is drawn from $N(\theta_1, \sigma^2)$ until it leads to a significant empirical effect estimate according to the one-sided t-test. The observations for all replication studies $j = 2, ..., m+1$ are randomly sampled from the respective normal distribution $N(\theta_i, \sigma^2)$.

Scenario 2: positive effect with selective reporting

In scenario 2, the effect θ_0 is positive and significantly different from zero. The concrete value can be arbitrarily chosen, as long as the confidence interval with the default variance value τ^2 does not entail zero.

Scenario 2 describes the distortion caused by selective reporting. Selective reporting entails publication bias, HARKing and fishing for the most surprising effect. It results in an overly optimistic effect size

estimate in whose favor all other effect size estimates are discarded (Chan et al., 2004).

Selective reporting is simulated by multiplying the defined true parameter θ_0 with the percentage by which the original effect size estimate typically overestimates the true effect size, denoted by p_{over}. The data for the original study is hence drawn from

$$X_i|\theta_0 \sim N(\theta_0 \cdot (1 + p_{over}, \sigma^2) \tag{3.86}$$

We define typical as the amount of overestimation detected in Klein et al., 2014. Based on this overestimated effect size a sample is drawn such that its t-test is significant.

The means $\theta_i, i = 2, ..., m + 1$ for the replication studies are sampled from $N(\theta_0, \tau^2)$ with θ_0 being the positive effect. The study observations themselves are then sampled with size n from $N(\theta_i, \sigma^2)$.

Scenario 3: positive effect without distorting factors

The simulation of scenario 3 is straightforward since there are no distorting factors influencing the sampling. Again, we model θ_0 as an arbitrary but positive effect size.

For $j = 1, ..., m$, the values θ_i are sampled from $N(\theta_0, \tau^2)$. The observations for each study are drawn from $N(\theta_i, \sigma^2)$. No further adaptions are required.

3.4.3 Generating the simulation study

To guarantee replicability of the simulated data sets, a seed was set using the *set.seed()* function with the seed 5.
We generate 100 replication studies for each of the three scenarios described in 3.4.2. The number of observations for condition and control group are determined as the average group size across replication studies and multi-lab hypotheses. For the condition group, the size amounts to 83, for the control group to 81. In accordance with the multi-lab examples and their use of Likert scales ranging from 0 to 9, β_0 is set to the average of such scale, $\beta_0 = 4.5$.

To simulate the original study and its replications, the intra-study variation is empirically estimated as the average of the empirical variation for the three multi-lab examples. The average for each hypothesis follows from $\overline{s^2} = \frac{1}{n} \sum_{i=1}^{n} s_j^2$.

For scenario 1, we obtain $\overline{s_1^2} = 2.39$, for scenario 2 $\overline{s_2^2} = 3.35$ and for scenario 3 $\overline{s_3^2} = 3.94$.

For the normal distribution $N(\theta_0, \tau^2)$ from which the replication mean θ_i is sampled, the overall mean θ_0 is defined by the type of simulation scenario. The variance τ^2 can be arbitrarily chosen and is set to 0.01. The empirical estimate for θ_0 again originates from the real multi-lab hypotheses.

For scenario 2, we model an overestimation of the true parameter, θ_0, in the original study due to selective reporting. The amount of overestimation is determined as the average percentage by which the original estimates overestimate the replication effect mean for selected hypotheses in Klein et al., 2014. Figure 1 in Klein et al., 2014 lets us suspect selective reporting for four multi-lab hypotheses.

- Imagined contact hypothesis
- Gain vs loss framing hypothesis
- Implicit math attitudes relations with self-reported attitudes hypothesis
- Sex differences in implicit math attitudes hypothesis

The flag priming hypothesis is excluded due to its non-significant replication effect size.

Based on said hypotheses, we obtain an average 190% overestimation of the true effect by the original study. Hence, θ_1 for the second scenario is set to 0.448 which is obtained through multiplying the percentage of overestimation times the mean effect from the imagined contact hypothesis.

For scenario 1, we sample repeatedly from a normal distribution with a zero mean.

For scenario 3, no distortion is simulated and therefore, we simply draw samples from the normal distribution with mean 0.576 according to the sunk cost hypothesis.

3.5 Application to multi-lab examples and simulation data

3.5.1 Software and packages

The analysis of the multi-lab examples and the simulated data was performed in R (R Core Team, 2020). The packages utilized include,

- Data Table by Dowle and Srinivasan, 2019
- MBESS by Kelley, 2020
- metafor by Viechtbauer, 2010
- BayesFactor by Morey and Rouder, 2018
- pwr by Champely, 2020
- ReplicationSuccess by Held et al., n.d.
- rstan by Stan Development Team, 2020
- puniform by van Aert, 2021b
- TeachingDemos by Snow, 2020
- effectsize by Ben-Shachar et al., 2020
- DescTools by Signorell, 2021
- compute.es by Del Re, 2013

The code was build in a modular fashion. Functions are defined in separate files. Data is pre-processed separately and saves as R.Data files.

A total of nine scripts prepare the data, implement the criteria and test the outcome.

- check of assumptions in 1_assumptions_check.R
- data preparation in 2_data_prep.R
- calculation of original parameters in 3_original_parameters.R
- calculation of empirical parameter estimates and manual selection of parameter values for simulation study in 4_parameter.R
- generation of simulation study in 5_simulation.R

- implementation of criteria for one replication study including on pooled data in 6_oneReplication.R
- implementation of criteria for multiple replication studies in 7_multipleReplications.R
- comparison of Bayes factors with one-sided and two-sided prior assumption in 8_BayesPrior.R
- testing for correctness of criteria application in 9_tests.R

The necessary functions are defined in different files according to their purpose.

- functions preparing data by calculating t-values, Cohen's d, variance and mean for condition and control group in functionsPrep.R
- functions implementing the criteria for one replication study in functionsOneReplication.R
- functions implementing the criteria for multiple replication studies in functionsMultiReplication.R
- functions available from Wagenmakers, Beek, et al., 2016 in Repfunctionspack.R

3.5.2 Pre-processing of the multi-lab examples

The authors to the multi-lab examples have made their data sets and code freely available.

1. Klein et al., 2014: https://osf.io/wx7ck/ and https://github.com/ManyLabsOpenScience/ManyLabs1
2. Wagenmakers, Beek, et al., 2016: https://osf.io/h2f98/

The functions on how to read and process the csv files including exclusion criteria are adopted from the code by Klein et al., 2014 and Wagenmakers, Beek, et al., 2016.

Each data set is attributed four RData files which consist of relevant test statistics and empirical estimates required to apply the replication criteria.

- mean data set: mean effect size, variance and sample size for condition and control group
- Cohen's d data set: effect size as Cohen's d
- t value data set: t-test statistic, sample size for condition and control group
- unaggregated data list: individual observations

The t-test statistic is calculated by the base function *t.test()*. Cohen's d is computed by the function *tes()* from the package compute.es.

3.5.3 Method implementation in R

Criteria for one replication

The first two criteria require the calculation of p-values. In this thesis, we leverage a one-sided two sample t-test with significance level $\alpha = 0.05$ (see chapter 3.1.2). This is performed by the base function *t.test()*.

The meta-analysis p-value is manually coded as a the t-test statistic depending on the combined squared sum of differences between condition and control group mean across original and replication study and the pooled variance. Given the variance for the original study are unknown, we reconstruct it by transforming the formula to calculate Cohen's d. We obtain the population variance which we assume equals the variance in the condition and control group. For the replication study, the variation can be easily estimated given the unaggregated data. The total variation is computed as a pooled variation for the condition and control

groups across original and replication study. Based on these key statistics, the meta-analysis p-value is computed and the test decision made to a significance level $\alpha = 0.05$.

The function *ttestBF()* from the BayesFactor package implements the JZS Bayes factor and is also leveraged to calculate the evidence updating Bayes factor.

Due to our interest in an oriented hypothesis $H_1 : \delta > 0$, the JZS prior is adapted by being specified one-sided (Marsman et al., 2017, Wagenmakers, Verhagen, et al., 2016). Under H_1 we assume a positive-only prior distribution. This indicates the distributions to be truncated at 0 and renormalized (Wetzels et al., 2009, Wagenmakers, Beek, et al., 2016). Morey and Wagenmakers, 2014 have proven that the one-sided Bayes factor can also be computed based on the traditional two-sided Bayes factor and the ratio between the marginal posterior and the marginal prior probability. Practically, the JZS Bayes factor and the evidence updating Bayes factor are performed one-sidedly by leveraging the opportunity to specify the null interval in the R function *ttestBF*. The effect such prior assumptions have on the value of the Bayes factor is investigated in chapter 4.4.

The replication Bayes factor was coded by Wagenmakers, Beek, et al., 2016. In the context of this thesis, the original replication Bayes factor is applied to a two-sample t-test. This leads to a change in how the non-centrality parameter is calculated, in particular the factor under the square root, as described in chapter 3.1.2.

The code for the equality-of-effect-size Bayes factor was also written by Wagenmakers, Beek, et al., 2016. In accordance with Bayarri and Mayoral, 2002 and Verhagen and Wagenmakers, 2014, we set the hyper-paramters for the prior distributions to $a = 2$ and $k = 2$.

According to the proposal by Jeffreys, 1961, Bayes factors are classified into four groups.

- $BF_{10} < \frac{1}{3}$: the Bayes factor indicates strong evidence for H_0, it is labeled “very H0”
- $BF_{10} \in \left[\frac{1}{3}, 1\right]$: the Bayes factor indicates weak evidence for H_0, it is labeled “H0”
- $BF_{10} \in [1, 3]$: the Bayes factor indicates weak evidence for H_1, it is labeled “H1”
- $BF_{10} > 3$: the Bayes factor indicates strong evidence for H_1, it is labeled “very H1”

This grouping including its labeling is maintained for every Bayes factor criterion applied to one replication study. The Bayes factor criteria computed for multiple replication studies simultaneously are cut off at an arbitrary value of 1 billion to ensure readability of the results. Values above the threshold are indicated as ’> 1B’.

Regarding the intervall-based methods, we set $\alpha = 0.1$, leading to a 90% credibility interval. The function *ttestBF()* can alternatively output a sample from the posterior distribution for θ_i which we utilize to calculate the posterior credibility interval.

The STAN-code for the HDI - in isolation and in a hierarchical model - originates from Marsman et al., 2017 and is available under https://osf.io/bqwzd/. While Marsman et al., 2017 consider all effect sizes across different experiments to originate from one overarching distribution, we assume separate group-level distributions for each hypothesis and simulation scenario.

The relevant Cohen’s d value $\delta_{0.33}$ which is essential to the small telescope criterion is found through the function *pwr.t.test()* from the package pwr.

The sceptical p-value approach is implemented in the package ReplicationSuccess in the function *pSceptical()*. Its significance level is set to 0.05 in accordance with the established choice in null hypothesis

significance testing. Since we are considering a one-sided alternative hypothesis, the original hypothesis test must also have been one-sided. If we only assume a one-sided test post hoc after performing a two-sided test on the original data, the original significance level α needs to be divided by two to obtain the transformed significance value for the sceptical p-value, $\tilde{\alpha} = 0.5 \cdot \alpha$. This applies to the sunk cost and imagined contact hypothesis. Hence, we compare the sceptical p-value to a significance level of $\tilde{\alpha} = 0.025$.

The snapshot hybrid method is also easily applied given the designated function *snapshot()* from the puniform package. According to the original paper by Van Aert and Van Assen, 2017, we calculate the posterior distribution for four distinct choices of the correlation r.

1. zero effect $r_z = 0$
2. small effect $r_s = 0.1$
3. medium effect $r_m = 0.3$
4. large effect $r_l = 0.5$

We attribute each choice an a-priori probability of $p_i = 0.25$.

Since multiple replications studies are available for each multi-lab hypotheses, the methods for one replication are applied on each replication study individually. The rate of successful replications is subsequently determined as the percentage of successful replication studies - defined by the respective method - under all available replication studies.

Considering the outcome in total, replicability is implied if the success percentage is greater than 50%.

Criteria for multiple replications

The non-central confidence interval was implemented in code by Klein et al., 2014. It utilizes the function *ci.smd()* from the package MBESS and requires the standardized mean difference as an input which is delivered by the function *escalc()* from the metafor package. $\alpha = 0.01$ is chosen as a significance level in accordance with the computation by Klein et al., 2014.

The Bayesian evidence synthesis approach relies on the already mentioned function *ttestBF()* which is applied to the pooled data set.

The fixed-effect meta-analysis Bayes factor has been calculated via two functions,

1. Leveraging the function meta.ttestBF() from the package BayesFactor
2. Applying the function metaBF() written by Wagenmakers, Beek, et al., 2016

While the code by Wagenmakers, Beek, et al., 2016 assumes a two-sided prior, the R function allows a specification of a one-sided JZS prior. Additionally, both functions use different approximation methods. Subsequently, estimates are similar but not identical Bayes factors.

Chapter 4

Results

In this section, we will discuss the results of applying the different replication criteria on the simulation data and the multi-lab examples,

- Facial feedback hypothesis - short: *facial*
- Sunk cost hypothesis - short: *cost*
- Imagined contact hypothesis - short: *contact*

The result analysis is divided, as is the method explanation in chapter 3.2, into results for methods for one replication (chapter 4.1) and for methods for multiple replications (chapter 4.3), implemented for the replication studies across all data sets. In addition, we analyze the performance of the criteria for one replication on the pooled research data (chapter 4.2). The assumptions adopted by the various replication criteria are verified in chapter 3.2.2.

The relevant data sets and code are available on GitLab under https://gitlab.lrz.de/stephaniearmbruster/replicationSuccess.git.

4.1 Criteria for one replication

The results obtained after applying the criteria for one replication to the multi-lab examples are shown in tables 4.1, 4.2 and 4.3. For the simulation study, the results are illustrated in tables 4.4, 4.5 and 4.6. The percentages in the second column of each overview table indicate the rate of replication successes among all replication studies. Success is hereby defined individually for each criterion according to its underlying theory, as explained in chapter 3.2. For the criteria with a non-binary outcome, additional information is adjunct in subsequent columns.

An unaggregated result overview for each multi-lab replication study can be found in appendix C. The plots depicting the exact intervals constructed by the interval-based methods - posterior credibility interval, HDI in isolation and in a hierarchical model - are included in appendix D. In these intervals, distorting factors can be distinguished by visually analyzing the intervals and setting the original effect size into comparison. If the original effect size lies at the upper boundary, we can consider it an overestimation. Depending on the true effect size, it is either due to publication bias (with a true zero effect) or selective reporting (with a true non-zero but small effect).

For each multi-lab example and simulation scenario, we include a normal posterior density plot for grand mean θ and heterogeneity τ^2, modeled in the hierarchical model for the HDI (see appendix D).

The HDI - in isolation and in a hierarchical model - are constructed for each hypothesis by sampling from the posterior density of the corresponding replication study with STAN.

For HDI in isolation, warning messages occurred indicating that MCMC chains did not mix well. For HDI in a hierarchical model, some MCMC chains diverged. While some initial error diagnosis was unsuccessfully attempted, further debugging lay out-of-scope for this thesis. Consequently, the validity of the conclusions can be questioned and the following interpretation has to be taken with a pinch of salt.

4.1.1 Facial feedback hypothesis

According to the original paper by Strack et al., 1988, the facial feedback hypothesis is supported by a significant effect estimate, measured as Cohen's d, 0.46.

The replication studies performed by Wagenmakers, Beek, et al., 2016, however, have proven a lack of replicability - and deliver evidence for a zero effect. This conclusion is corroborated when applying our selected replication criteria to the data.

While over half of the replication effect estimates have a positive sign like the original estimate, the rate of successes only slightly surpasses the rate of failures, hinting at an equal distribution between positive and negative effect estimates and with that a zero effect size.

While none of the replication studies for the facial feedback hypothesis regarded in isolation concluded a significant result, the pooled data set estimates are significant for 23.5% of studies. This discrepancy illustrates the dependency of the p-value computation on sample size. Nevertheless, the majority of replication studies conclude an unsuccessful replication.

The JZS Bayes factor leads to the same conclusion of non-replicability - the data supports the claim of a zero effect size over a non-zero in 76.5% of studies with substantial evidence.

The equality-of-effect-size Bayes factor considers the relationship between replication and original study by testing for equality between the respective effect estimates. For the facial feedback hypothesis, the criterion only concludes weak evidence for equality. This implies some sort of distortion by which the original study was influenced. In accordance with the assumption of a true zero effect - given the results of previous criteria - we can suppose the original study to likely have been subjected to publication bias (or other distorting factors).

According to the replication Bayes factor, the majority of replication studies again obtain evidence for a zero effect. The replication effect sizes seem to contradict the posterior distribution based on the original data. This apparent discrepancy between original and replication study points towards non-replicability.

The results obtained for the small telescope criterion slightly disagree with the conclusions drawn so far. The majority of replication studies conclude replicability since their estimates suffice in magnitude to have been detected by the original study. However, the effects observed in 23.5% of the studies are too small to obtain statistical power above the threshold of 33%, indicating some degree of evidence towards non-replicability.

The disadvantage of the sceptical p-value criterion is its requirement of significance for both original and replication effect size if considered in isolation. For the facial feedback hypothesis, this does not apply to any replication study - all are insignificant and hence, the sceptical p-value cannot be implemented.

The majority of studies lead to an inconclusive result for the snapshot hybrid criterion. For over a third of replication studies, however, the highest posterior likelihood attributed to the zero effect lies above the threshold of 75%. This hints at a true zero effect and thus non-replicability.

None of the constructed intervals - neither for the posterior credibility interval, nor for HDI in isolation, nor in a hierarchical model - exclude zero. Through examining the intervals visually and additionally plotting the original estimate (see appendix D), we can assume that the original estimate suffers under publication bias.

Criterion	Replication success in %				
Effect orientation	52.9				
Significance	0				
Meta significance	23.5				
JZS BF	very H0 (76.5)	H0 (23.5)	H1 (0)	very H1 (0)	
Equality BF	H0 (58.8)	very H0 (23.5)	H1 (17.6)	very H1 (0)	
Replication BF	very H0 (76.5)	H0 (23.5)	H1 (0)	very H1 (0)	
Posterior CI	0				
HDI isolation	0				
HDI hierarchical	0				
Small telescope	76.5				
Snapshot hybrid	inconclusive (64.7)	p.0 (35.3)	p.sm (0)	p.me (0)	p.la (0)

Table 4.1: Results for facial feedback hypothesis

4.1.2 Imagined contact hypothesis

In the original paper by Husnu and Crisp, 2010, the effect to the imagined contact hypothesis is computed as a significant positive effect, 0.86, measured as Cohen's d.

The replication studies conducted by Klein et al., 2014 deliver evidence for the replicability of the original study. However, some criteria applied by Klein et al., 2014 yield contradictory results. According to our analysis, the original paper to the imagined contact hypothesis suffers under selective reporting. The original effect estimate is an overestimation of a rather small true positive effect.

While for most replication studies the JZS Bayes factor supports the zero effect more than the non-zero, the rate does not lie above 50%. Hence, we can consider the results inconclusive and could interpret them as an indication for a small non-zero true effect which is challenging to detect due to variation. We also have to keep in mind that the JZS Bayes factor is hesitant in rejecting the null hypothesis.

According to the equality-of-effect-size Bayes factor the highest percentage of replication studies conclude weak evidence for equality, followed by those that obtain weak evidence against the equality. This points towards a subtle but existent divergence between the original study and the replications.

The replication Bayes factor indicates that the replication effect estimates seem to contradict the posterior distribution based on the original data. This apparent discrepancy between original and replication study again contributes to the claim of non-replicability and the presumption of selective reporting.

The snapshot hybrid criterion lets us suppose a small magnitude for the true effect size. 75% of the replication studies are inconclusive since the posterior probability tends to be spread across the zero and small effect size. We can suspect the true effect size to lie between zero and the small effect size ($p = 0.1$).

This is corroborated by the effect size intervals constructed as 90% posterior credibility interval and HDI in isolation. Only a small percentage of replication studies obtain an interval that excludes zero. Apparently,

the uncertainty inherent in each study inhibits a decisive conclusion.

When taking all performed replications into account, none of intervals constructed as HDI in a hierarchical model include zero - shifting the interpretation from non-replicable to replicable. This exemplifies very well how sharing information across studies decreases uncertainty and hence allows the statistic to pick up on close-to-zero effects.

The small telescope criterion proves that the majority of replication effect estimates have a sufficiently high magnitude to have been detected with a statistical power above the threshold of 33% in the original study. This contributes to our assumption that the effect size is small but different from zero.

For the imagined contact hypothesis, the original papers perform two-sided tests and consequently the relevant significance level for the sceptical p-value lies at $\tilde{\alpha} = 0.025$. 5 replication studies exhibit a significant effect size estimate and are thus suitable for the sceptical p-value criterion. However, none of the calculated sceptical p-values lie below the threshold 0.025. Consequently, we can interpret the replication effect estimates to be insufficient in persuading the sceptic that the sceptical prior is too sceptical and hence, that the effect size is non-zero. All studies are deemed non-replicable.

Criterion	Replication success in %				
Effect orientation	75				
Significance	13.9				
Meta significance	25				
JZS BF	very H0 (47.2)	H0 (27.8)	H1 (13.9)	very H1 (11.1)	
Equality BF	H0 (47.2)	H1 (36.1)	very H1 (13.9)	very H0 (2.8)	
Replication BF	very H0 (63.9)	H0 (22.2)	very H1 (8.3)	H1 (5.6)	
Posterior CI	13.9				
HDI isolation	13.9				
HDI hierarchical	100				
Small telescope	75				
Sceptical p-value	0				
Snapshot hybrid	inconclusive (75)	p.0 (19.4)	p.sm (5.6)	p.me (0)	p.la (0)

Table 4.2: Results for imagined contact hypothesis

4.1.3 Sunk cost hypothesis

The original study by Oppenheimer et al., 2009 obtains a Cohen's d effect estimate, 0.23, for the sunk cost hypothesis.

Klein et al., 2014 corroborate this finding in their replication studies. The replication analysis conducted in this thesis also supports the claim that the original study can in fact be replicated and that the original study obtains a representative positive effect estimate - without any distortion.

This conclusion is supported by the majority of investigated criteria. The equality-of-effect-size Bayes factor decisively indicates the equality between replication and original effect estimate. The p-value based methods - orientation, significance and meta significance - as well as the JZS Bayes factor speak for a non-zero effect. For the latter, some replication studies portray weak evidence for a zero effect size. This might indicate that the effect size is small in magnitude, rendering it difficult to detect amidst variance and heterogeneity.

The results obtained when applying the replication Bayes factor suggest a similar interpretation. While 50% of the replication studies indicate that the data is more likely to originate from the posterior distribution

defined by the original study than from a zero effect, 27.8% of replications deliver weak evidence for a zero effect.

Considering the snapshot hybrid criterion, the results for the sunk cost hypothesis seem equally ambiguous. 77.8% of replication studies are inconclusive since the posterior probability is spread across the small and zero effect size. This results from the pitfall of the snapshot hybrid method - considering a discrete number of distinct parameter values. While the true effect size of the sunk cost hypothesis is likely not equal zero, it is small and thus lies between the zero and the small parameter choice.

The interval based methods only considering one study at a time tend towards excluding zero. However, the uncertainty of the replication studies renders the detection of a small effect size difficult. When considering the relationship and thus the shared information across the studies, the HDI excludes zero for all studies and implies a true effect significantly different from zero.

Criterion	Replication success in %				
Effect orientation	91.7				
Significance	58.3				
Meta significance	83.3				
JZS BF	very H1 (41.7)	H0 (25)	H1 (22.2)	very H0 (11.1)	
Equality BF	very H0 (94.4)	H0 (5.6)	very H1 (0)	H1 (0)	
Replication BF	very H1 (50)	H0 (27.8)	H1 (22.2)	very H0 (0)	
Posterior CI	52.8				
HDI isolation	50				
HDI hierarchical	100				
Small telescope	100				
Sceptical p-value	0				
Snapshot hybrid	inconclusive (77.8)	p.0 (11.1)	p.sm (11.1)	p.me (0)	p.la (0)

Table 4.3: Results for sunk cost hypothesis

4.1.4 Simulation study - scenario 1

Scenario 1 is simulated from a zero mean distribution with heterogeneity and variance. The original study is subjected to publication bias.

Regarding the comparison between the sign of the original and replication effect estimate, the studies can be considered equally distributed between same and opposite direction. This is a natural consequence to sampling from a zero effect distribution and can be interpreted as evidence for non-replicability.

The overwhelming majority of replication studies correctly do not reject the null hypothesis. The significant effect estimates can be considered Type-I errors - false positives. While there is a higher rate of significant meta-analysis p-values, the overall conclusion still correctly indicates non-replicability, i.e. hints at a zero effect size.

The JZS Bayes factors supports this claim by accepting the null hypothesis. The snapshot hybrid method detects a correct zero effect in the majority of replication studies, while the rest of the studies are inconclusive.

The intervals also uniformly speak for a zero effect since they all include zero - apart from a small percentage of posterior credibility intervals and HDI in isolation. Publication bias can be identified thanks to the discrepancy between the replication estimates and the original finding. The criteria indicate a zero effect size while the original study claims a positive effect which is significantly different from zero.

The equality-of-effect-size Bayes factor surprisingly delivers evidence for the equality of effect estimates between original and replication study. This indicates that the criteria is rather ill fit to handle publication bias in combination with an actual zero effect and heterogeneity between studies.

The same judgement can be made when observing the results for the replication and the evidence updating Bayes factor. Both criteria only deliver weak evidence for a zero effect size. Due to ambiguity in the data, the Bayes factors cannot correctly detect that the replication studies do not align with the original false positive effect estimate.

For the small telescope criterion, 84% replications studies do not reject the null hypothesis despite the true underlying effect being zero and hence in theory not detectable. The variation in treatment effects seems to lead to the wrong conclusion of a positive effect size.

Criterion	Replication success in %				
Effect orientation	55				
Significance	4				
Meta significance	31				
JZS BF	very H0 (75)	H0 (20)	H1 (3)	very H1 (2)	
Equality BF	very H0 (64)	H0 (26)	H1 (9)	very H1 (1)	
Replication BF	H0 (87)	H1 (10)	very H1 (3)	very H0 (0)	
Evidence updating BF	H0 (44)	very H0 (36)	H1 (16)	very H1 (4)	
Posterior CI	4				
HDI isolation	4				
HDI hierarchical	0				
Small telescope	84				
Sceptical p-value	0				
Snapshot hybrid	p.0 (69)	inconclusive (31)	p.sm (0)	p.me (0)	p.la (0)

Table 4.4: Results for simulation scenario 1

4.1.5 Simulation study - scenario 2

Scenario 2 resembles the imagined contact hypothesis. It is generated to have a small true effect size which is overestimated by the original study due to selective reporting.

The positive nature of the effect size is indicated by the high rate of significant meta-analysis p-values. The null hypothesis assuming a zero effect size is rightfully rejected. P-value based criteria, however, do not allow for any further analysis regarding relationship between studies. The p-value does not account for any selective reporting but simply establishes the existence of a non-zero effect.

The small telescope criterion as well as the interval-based approaches also deliver evidence for a non-zero effect size. For all HDI in a hierarchical model, the intervals exclude zero, again drawing the right conclusion that the effect size is not zero. The intervals constructed in isolation are ill-fit to detect small effect sizes if variance is high. Potential distortions can be identified by examining the intervals and plotting the original effect estimate for a visual comparison.

Snapshot hybrid obtains inconclusive results for 90% of replication studies. The probability is distributed between the zero and small effect size, indicating that the true size lies in between.

The Bayes factors - JZS, replication and evidence updating - conclude at least weak evidence for a zero effect in 49%, 74% and 35% of replication studies respectively. The true effect has an insufficiently low magnitude to be detected by the JZS Bayes factors. For the replication and evidence updating Bayes factor,

the results indicate selective reporting. The posterior distribution constructed based on the original study does not align with the replication estimates.

On the contrary, selective reporting is not accurately detected by the equality-of-effect-size Bayes factor. The majority of replication studies obtain Bayes factor values indicating equality between the original and replication estimate.

Criterion	Replication success in %				
Effect orientation	80				
Significance	22				
Meta significance	93				
JZS BF	very H0 (49)	H0 (27)	H1 (16)	very H1 (8)	
Equality BF	very H0 (54)	H0 (36)	H1 (9)	very H1 (1)	
Replication BF	H0 (74)	H1 (17)	very H1 (9)	very H0 (0)	
Evidence updating BF	very H0 (35)	H0 (28)	very H1 (20)	H1 (17)	
Posterior CI	21				
HDI isolation	21				
HDI hierarchical	100				
Small telescope	98				
Sceptical p-value	0				
Snapshot hybrid	inconclusive (90)	p.0 (10)	p.sm (0)	p.me (0)	p.la (0)

Table 4.5: Results for simulation scenario 2

4.1.6 Simulation study - scenario 3

Scenario 3 is modeled to incorporate a positive effect size which is correctly estimated by the original study without any distortions.

The positive replication effect size which is in line with the original estimate is detected by the p-value based criteria, the replication and the evidence updating Bayes factors as well as the small telescope criterion and the interval-based methods alike. Both, the replication and the evidence updating Bayes factor provide evidence for the representative nature of the original estimate. The equality-of-effect-size Bayes factor supports this interpretation by concluding equality between the original and replication estimates amongst the vast majority of studies.

Snapshot hybrid again lets us assume that the magnitude of the true effect size lies between the zero and small effect size since the majority of studies are inconclusive due to an equal probability distribution between the zero and small effect size.

In accordance with the results for the sunk cost hypothesis, JZS Bayes factor is surprisingly inaccurate. Only in 39% of the studies it correctly delivers strong evidence for a non-zero effect while it obtains weak evidence for a zero effect size in 24% of replications. A reason for this low performance might be the small magnitude of the true effect in combination with heterogeneity and variance inherent to each study.

4.2 Criteria for one replication on pooled data

4.2.1 Estimation of heterogeneity

The heterogeneity of effect estimates across the original and all replication studies is calculated by the EB estimator (see chapter 3.2.2). Its computation requires the variance for the original effect size which is extracted from the t-test value.

Criterion	Replication success in %				
Effect orientation	97				
Significance	56				
Meta significance	82				
JZS BF	very H1 (39)	H0 (24)	H1 (23)	very H0 (14)	
Equality BF	very H0 (93)	H0 (7)	very H1 (0)	H1 (0)	
Replication BF	very H1 (46)	H1 (31)	H0 (23)	very H0 (0)	
Evidence updating BF	very H1 (54)	H1 (31)	H0 (15)	very H0 (0)	
Posterior CI	53				
HDI isolation	53				
HDI hierarchical	100				
Small telescope	100				
Sceptical p-value	0				
Snapshot hybrid	inconclusive (85)	p.0 (15)	p.sm (0)	p.me (0)	p.la (0)

Table 4.6: Results for simulation scenario 3

We obtain the following empirical variance estimate for our three hypotheses.

Hypothesis	$\widehat{\sigma}^2$
facial	0
contact	0.011
cost	0

Table 4.7: EB estimates of effect size heterogeneity

The EB estimator for heterogeneity results in (close-to-) zero estimates for every hypothesis - indicating that there exists no notable variation in treatment effect.
Cochran's Q test corroborates this conclusion by not rejecting the null hypothesis, $\tau^2 = 0$, for any of the considered multi-lab examples. Since all replication studies are designed as direct replications, there is little leeway given for the influence of hidden mediators, resulting in low heterogeneity between studies. As a consequence, it is valid to consider $\tau^2 = 0$ and the true effect size θ_0 to directly underlie all observations regardless of the exact study. Hence, it is justified to concatenate the replication studies to one combined data set. We can leverage the resulting large sample size for high statistical power (Fahrmeir et al., 2016).

However, pooling is only a valid option when we can be certain that no relevant variation in treatment effect or other statistically important differences between the replication studies exist. Consequently, the simulation data which is generated to exhibit heterogeneity is not analyzed through pooling. For the facial feedback, sunk cost and imagined contact hypothesis, the criteria for one replication study are applied to the pooled data set.

4.2.2 Implementation of replication success criteria

The results for each criterion are summarized in the table below.

Overall, considering all criteria, we can draw the conclusion that a zero effect size underlies the facial feedback hypothesis with high probability. The equality-of-effect-size Bayes factor only delivers weak evidence for the equality of effect size estimates between the original study and the pooled replication data. The sceptical p-value cannot be applied since the p-value to the pooled data is not significant and hence the condition for the method is not fulfilled.

The evidence towards a small positive effect for the imagined contact hypothesis is less conclusive compared

to the sunk cost hypothesis. Nevertheless, the majority of criteria speak for a non-zero effect. However, the original effect estimate seems to have been subject to distortion in the form of selective reporting. This observation follows from the equality-of-effect-size Bayes factor which delivers weak evidence for inequality between the estimates. The result from the small telescope criterion also speaks for selective reporting since the pooled replication estimate is not detectable.

The results for the sunk cost hypothesis speak for a small positive effect size which is significantly different from zero. This is indicated by the significance of both p-value and meta-analysis p-value. The JZS and replication Bayes factor both decisively speak for the alternative hypothesis. The small telescope criterion shows that the replication effect estimate could potentially be detected by the original study and hence confirms the original significant effect estimate. The equality-of-effect-size Bayes factor concludes that the original and pooled replication study align in their effect estimate and consequently that the original study was not heavily influence by any distorting factors.

Criterion	Facial Feedback	Imagined contact	Sunk cost
Effect estimate orientation	FALSE	TRUE	TRUE
Significance test: p-value	0.516	0.000	0.000
Meta-analysis p-value	0.452	0.229	0.000
JZS Bayes factor	5.000000e-02	1.883815e+04	8.440864e+23
Equality-of-effect-size Bayes factor	0.731	1.215	0.091
Replication Bayes factor	4.800000e-02	7.387137e+03	3.113805e+24
Small telescope: significant (1)	1	1	0
Sceptical p-value: p-value		0.480	0.475
Snapshot hybrid	p.0 (100)	p.sm (100)	p.sm (100)

Table 4.8: Results overview for replication success criteria applied to pooled multi-lab examples

The below graph 4.1 plots the 90% credibility interval for the posterior distribution in combination with the original effect estimate for all three multi-lab examples. For the facial feedback hypothesis, zero is included in the interval, implying a zero effect and hence non-replicability. For the two other multi-lab examples, the posterior equal-tailed credibility interval excludes zero, indicating the existance of an effect.

Overall, we can draw following conclusions,

- *facial*: The effect size is not significantly different from zero and the original estimate according to Strack et al., 1988 is an extreme overestimation of the true effect.
- *contact*: While a small non-zero effect seems to exist, the original effect estimate by Husnu and Crisp, 2010 is an overestimation due to selective reporting.
- *costs*: We can observe a positive effect size, significantly different from zero and aligned with the original estimate by Oppenheimer et al., 2009.

4.3 Criteria for multiple replications

The following section analyzes the outcome of applying the replication success criteria introduced in chapter 3.2.5. They are implemented on all available replication studies simultaneously. Therefore, they allow for an overarching decision regarding replicability while acknowledging all conducted replication studies.

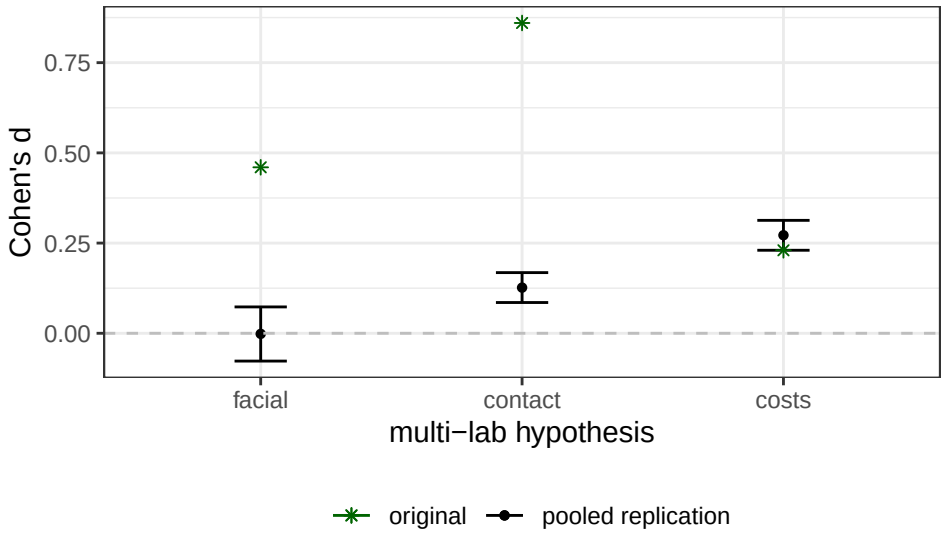

Figure 4.1: 90% credibility interval for the one-sided posterior distribution

4.3.1 Facial feedback hypothesis

In figure 4.2 Berkson's interocular traumatic test for the facial feedback hypothesis lets us assume that the true effect size is either zero or close-to-zero. The original significant effect estimate is an overestimation, suffering under publication bias, and is consequently not replicable.

Bayesian evidence synthesis	Bayes factor acc. to BayesFactor	Bayes factor acc. to Wagenmakers
0.05	0.08	0.05

Table 4.9: Results for facial feedback hypothesis (arbitrary threshold value 1 billion, introduced for readability)

All three criteria speak for the null hypothesis. The observed data across all replication studies is more likely to have been generated from a zero effect than a positive effect.

4.3.2 Imagined contact hypothesis

While Berkson's interocular traumatic test in figure 4.3 implies a small non-zero effect, the original estimate seems to be subjected to selective reporting since it strongly overestimates the true effect size.

Bayesian evidence synthesis	Bayes factor acc. to BayesFactor	Bayes factor acc. to Wagenmakers
18838.15	216890.05	152096.38

Table 4.10: Results for imagined contact hypothesis (arbitrary threshold value 1 billion, introduced for readability)

Again, all three criteria support the alternative hypothesis - a non-zero effect size. However, the evidence is less conclusive than for the sunk cost hypothesis.

4.3.3 Sunk cost hypothesis

For the sunk cost hypothesis, Berkson's interocular traumatic test in figure 4.4 indicate a non-zero positive effect estimate for which the original estimate is an undistorted representation.

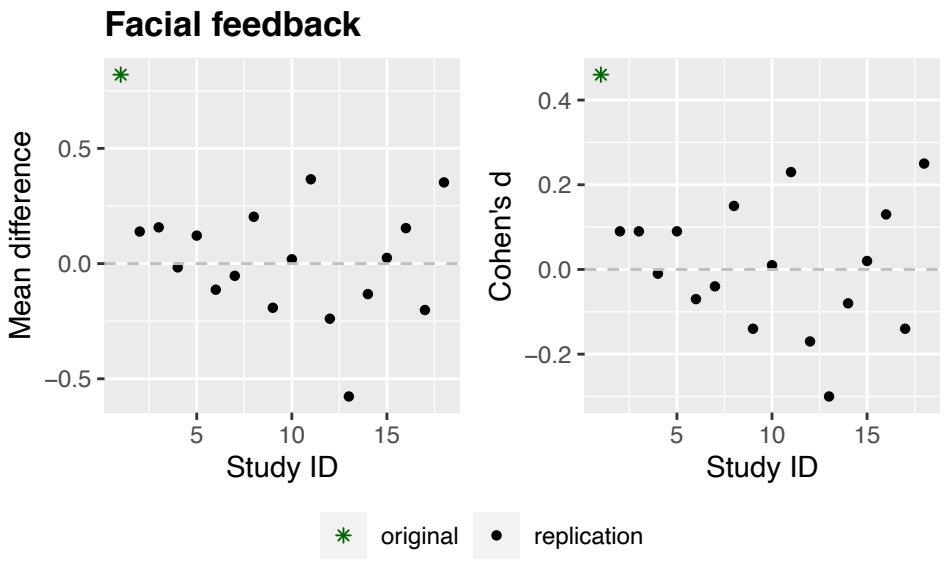

Figure 4.2: Overview: mean difference between condition and control group and Cohen's d - facial feedback hypothesis

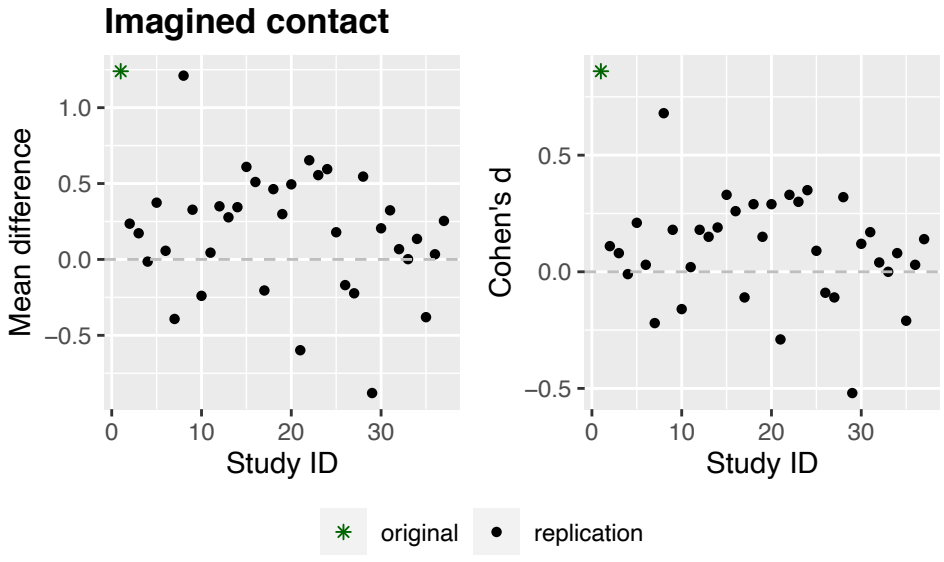

Figure 4.3: Overview: mean difference between condition and control group and Cohen's d - imagined contact hypothesis

Bayesian evidence synthesis	Bayes factor acc. to BayesFactor	Bayes factor acc. to Wagenmakers
$>$ 1B	$>$ 1B	$>$ 1B

Table 4.11: Results for sunk cost hypothesis (arbitrary threshold value 1 billion, introduced for readability)

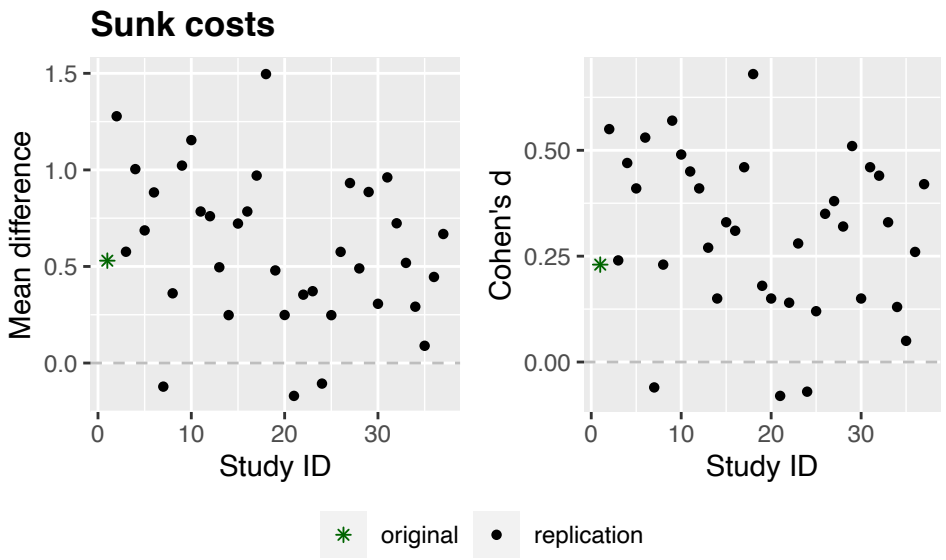

Figure 4.4: Overview: mean difference between condition and control group and Cohen's d - sunk cost hypothesis

All three criteria deliver decisive evidence for a non-zero effect size with very high Bayes factor values.

4.3.4 Simulation study - scenario 1

Berkson's interocular traumatic test for simulation scenario 1 (in figure 4.5) - as well as for scenario 2 (see figures 4.6) - is less conclusive. This exemplifies the limited detective and explanatory power of Berkson's test and simple graphical visualization in general.

However, the overview of raw difference between condition and control mean as well as the effect size measured as Cohen's d illustrate an equal deviation around zero, indicating that the true underlying effect size is zero. The original effect estimate is located at the boundary of observed values but there exist even more extreme values.

Bayesian evidence synthesis	Bayes factor acc. to BayesFactor	Bayes factor acc. to Wagenmakers
0.05	0.07	0.04

Table 4.12: Results for simulation scenario 1 (arbitrary threshold value 1 billion, introduced for readability)

For scenario 1, the criteria which we apply to all replication studies simultaneously conclude a zero effect.

4.3.5 Simulation study - scenario 2

As aforementioned, the Berkson's interocular traumatic test for scenario 2 is rather inconclusive. The distribution of the effect size - measured as both mean difference and Cohen's d - is skewed towards positive values, indicating a non-zero positive effect size. However, we cannot accurately identify distortion due to selective reporting since the effect estimates show substantial deviation.

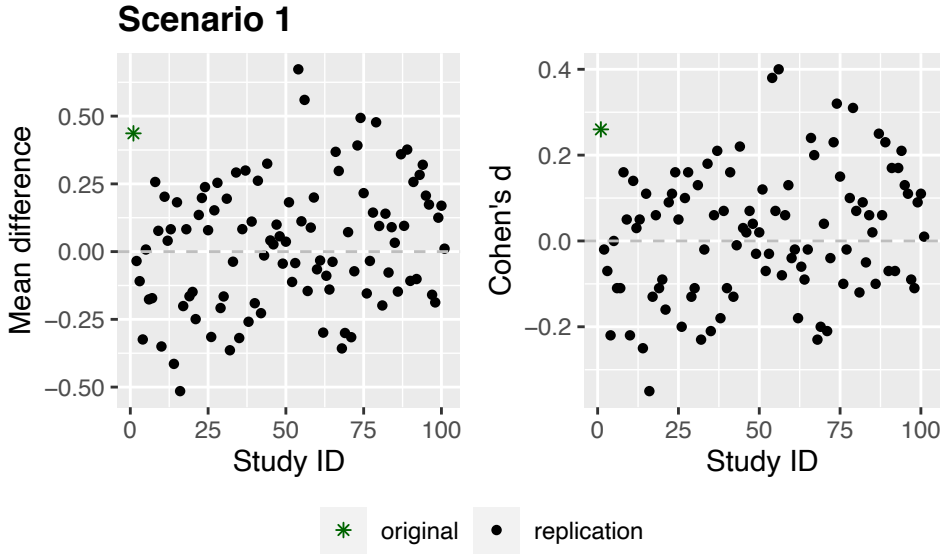

Figure 4.5: Overview: mean difference between condition and control group and Cohen's d - scenario 1

Bayesian evidence synthesis	Bayes factor acc. to BayesFactor	Bayes factor acc. to Wagenmakers
$>$ 1B	$>$ 1B	$>$ 1B

Table 4.13: Results for simulation scenario 2 (arbitrary threshold value 1 billion, introduced for readability)

All three criteria provide evidence for the alternative hypothesis. This allows us to rightfully conclude a non-zero effect size underlying scenario 2 - without delivering any information on potential distortions.

4.3.6 Simulation study - scenario 3

Berkson's interocular traumatic test for scenario 3 indicates that the original estimate is representative for the estimates obtained in the replication studies. This lets us infer that the original study is likely to not have been subjected to any distorting factors. Its conclusion of a significant, positive effect can be considered justified.

Bayesian evidence synthesis	Bayes factor acc. to BayesFactor	Bayes factor acc. to Wagenmakers
$>$ 1B	$>$ 1B	$>$ 1B

Table 4.14: Results for simulation scenario 3 (arbitrary threshold value 1 billion, introduced for readability)

All three criteria deliver large Bayes factor values, rightfully exhibiting strong evidence for the alternative hypothesis.

4.3.7 Graphical summary: non-central confidence interval

Figure 4.8 shows replicability is not given for the facial feedback hypothesis but indeed present in the sunk cost and imagined contact hypothesis.

However, the lower limit to the confidence interval of the imagined contact hypothesis is in close proximity to zero. The confidence interval thus suggests an effect only slightly different from zero (Klein et al., 2014).

For scenario 1, the non-central confidence interval includes zero and hence rightfully concludes a zero effect. The original Cohen's d estimate - plotted as dark green star - shows that the original estimate is

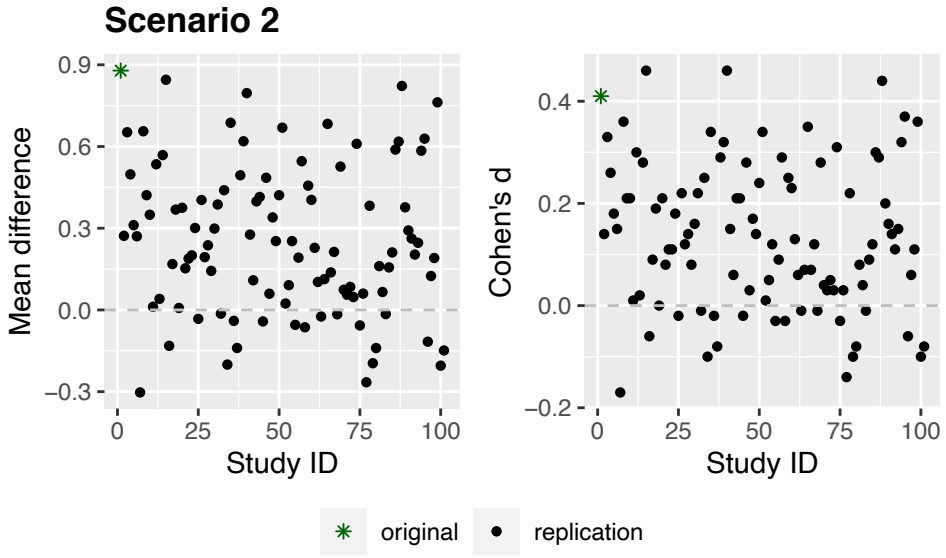

Figure 4.6: Overview: mean difference between condition and control group and Cohen's d - scenario 2

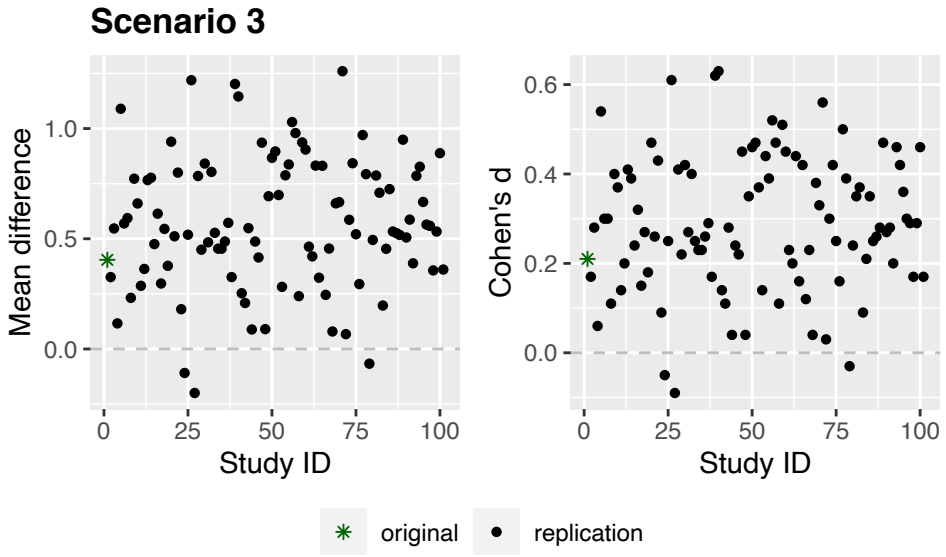

Figure 4.7: Overview: mean difference between condition and control group and Cohen's d - scenario 3

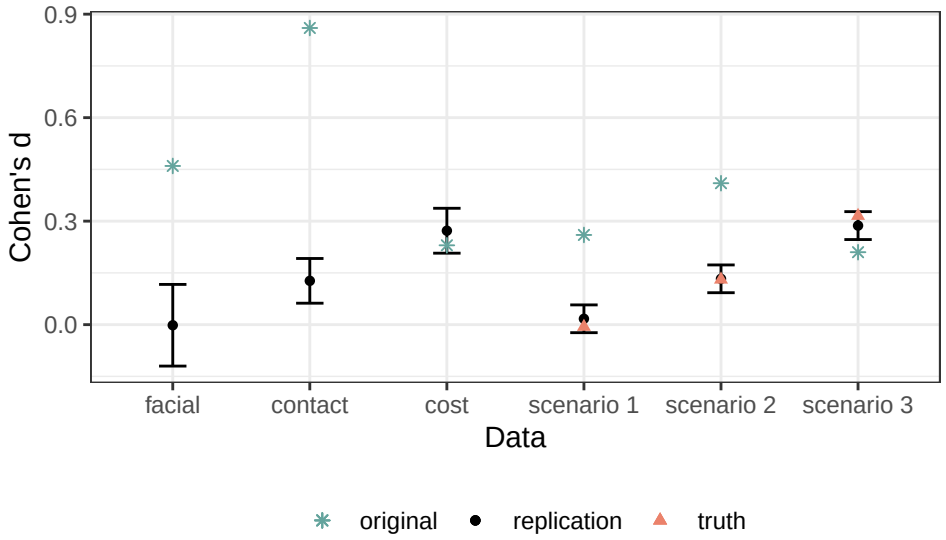

Figure 4.8: Non-central confidence intervals for Cohen's d - multi-lab examples and simulation data

an overestimation of the true effect. The same can be observed for scenario 2, for which the discrepancy between confidence interval and original estimate is even more distinct. Due to variation in treatment effect, the Cohen's d estimate for the original study differs notably from the non-central interval for scenario 3 despite it being simulated without distortions. For scenario 2 and 3, the non-central interval does not entail zero and hence allows the conclusion of a non-zero effect.

In general, the non-central interval seems a rather practical criterion since it allows a double assessment - whether the replication studies deliver evidence for a zero effect as well as what value the true effect size is likely to have. The accuracy of the latter can be illustrated by considering the true parameter sizes available for the simulation data. Plotting the median of the true Cohen's d effect size for all scenarios (pink triangle) showcases the precise overlap between true value and constructed non-central effect size interval.

4.4 The influence of a one-sided prior distribution

4.4.1 One-sided priors for Bayes factors

While the equality-of-effect-size Bayes factor and the replication Bayes factor rely on a two-sided prior, the JZS Bayes factor, the evidence updating Bayes factor,evidence synthesis and the fixed effect meta-analysis Bayes factor incorporate an oriented prior probability distribution which only attributes non-zero densities to positive effect size values.

In order to assess the influence of such a prior assumption, we have evaluated each of the one-sided criteria twice - with and without the one-sided limitation.

The tables 4.15 to 4.20 illustrate the difference between the limited and unlimited implementations of each criterion - labelled '(one)' for the limited and '(two)' for the unlimited version.

The results for the criteria designed to be applied to one replication study at a time - JZS Bayes factor and evidence updating Bayes factor - are indicated as success rates in %. The results for the criteria assessing all available replication studies simultaneously - evidence synthesis and fixed effect meta-analysis Bayes factor - are given directly as Bayes factor value.
To ensure readability, Bayes factor values are cut off at an arbitrary threshold of 1 billion.

As aforementioned, the fixed effect meta-analysis JZS Bayes factor according to our self written code is not identical to the Bayes factor calculated according to Wagenmakers, Beek, et al., 2016 due to different approximations.

Overall, no major changes in the outcome for any criterion can be observed. The interpretation regarding the effect size for the multi-lab examples as well as the simulation study seem to be independent of the prior choice.

Criterion	Replication success			
JZS BF (one)	very H0 (76.5)	H0 (23.5)	H1 (0)	very H1 (0)
JZS BF (two)	very H0 (82.4)	H0 (17.6)	H1 (0)	very H1 (0)
Evidence synthesis (one)	0.05			
Evidence synthesis (two)	0.052			
Meta-analysis BF (one)	0.076			
Meta-analysis BF (two)	0.056			
Meta-analysis BF (Wagenmakers)	0.046			

Table 4.15: Bayes factor comparison for facial feedback hypothesis (arbitrary threshold value 1 billion, introduced for readability)

Criterion	Replication success			
JZS BF (one)	very H0 (47.2)	H0 (27.8)	H1 (13.9)	very H1 (11.1)
JZS BF (two)	very H0 (55.6)	H0 (30.6)	very H1 (8.3)	H1 (5.6)
Evidence synthesis (one)	18838.154			
Evidence synthesis (two)	9419.079			
Meta-analysis BF (one)	216890.054			
Meta-analysis BF (two)	108445.029			
Meta-analysis BF (Wagenmakers)	152096.38			

Table 4.16: Bayes factor comparison for imagined contact hypothesis (arbitrary threshold value 1 billion, introduced for readability)

Criterion	Replication success			
JZS BF (one)	very H1 (41.7)	H0 (25)	H1 (22.2)	very H0 (11.1)
JZS BF (two)	very H0 (27.8)	H1 (25)	very H1 (25)	H0 (22.2)
Evidence synthesis (one)	$>$ 1B			
Evidence synthesis (two)	$>$ 1B			
Meta-analysis BF (one)	$>$ 1B			
Meta-analysis BF (two)	$>$ 1B			
Meta-analysis BF (Wagenmakers)	$>$ 1B			

Table 4.17: Bayes factor comparison for sunk cost hypothesis (arbitrary threshold value 1 billion, introduced for readability)

4.4.2 One-sided priors for posterior equal-tailed credibility interval

The posterior equal-tailed credibility intervals can also be constructed based on either a one-sided or a two-sided JZS prior. The one-sided posterior credibility interval tests a directional alternative hypothesis, $H_1 : \theta_0 \geq 0$.

The resulting mean credibility intervals for the multi-lab examples and the simulation study are plotted in figure 4.9. The one-sided version (with a more shallow width) is on the left, the two-sided version (with a wider width) is on the left for each hypothesis.

Criterion	Replication success			
JZS BF (one)	very H0 (75)	H0 (20)	H1 (3)	very H1 (2)
JZS BF (two)	very H0 (78)	H0 (18)	H1 (3)	very H1 (1)
Evidence updating BF (one)	very H0 (40)	H0 (39)	H1 (17)	very H1 (4)
Evidence updating BF (two)	H0 (44)	very H0 (36)	H1 (16)	very H1 (4)
Evidence synthesis (one)	0.055			
Evidence synthesis (two)	0.032			
Meta-analysis BF (one)	0.069			
Meta-analysis BF (two)	0.039			
Meta-analysis BF (Wagenmakers)	0.035			

Table 4.18: Bayes factor comparison for scenario 1 (arbitrary threshold value 1 billion, introduced for readability)

Criterion	Replication success			
JZS BF (one)	very H0 (49)	H0 (27)	H1 (16)	very H1 (8)
JZS BF (two)	very H0 (65)	H0 (22)	H1 (10)	very H1 (3)
Evidence updating BF (one)	very H0 (35)	H0 (28)	very H1 (20)	H1 (17)
Evidence updating BF (two)	very H0 (35)	H0 (28)	very H1 (20)	H1 (17)
Evidence synthesis (one)	$>$ 1B			
Evidence synthesis (two)	$>$ 1B			
Meta-analysis BF (one)	$>$ 1B			
Meta-analysis BF (two)	$>$ 1B			
Meta-analysis BF (Wagenmakers)	$>$ 1B			

Table 4.19: Bayes factor comparison for scenario 2 (arbitrary threshold value 1 billion, introduced for readability)

Criterion	Replication success			
JZS BF (one)	very H1 (39)	H0 (24)	H1 (23)	very H0 (14)
JZS BF (two)	H0 (34)	very H1 (28)	very H0 (25)	H1 (13)
Evidence updating BF (one)	very H1 (56)	H1 (30)	H0 (12)	very H0 (2)
Evidence updating BF (two)	very H1 (54)	H1 (31)	H0 (15)	very H0 (0)
Evidence synthesis (one)	$>$ 1B			
Evidence synthesis (two)	$>$ 1B			
Meta-analysis BF (one)	$>$ 1B			
Meta-analysis BF (two)	$>$ 1B			
Meta-analysis BF (Wagenmakers)	$>$ 1B			

Table 4.20: Bayes factor comparison for scenario 3 (arbitrary threshold value 1 billion, introduced for readability)

The zero effect underlying the facial feedback hypothesis and scenario 1 is detected by the two-sided interval, not by the one-sided one. The two-sided interval is unable to observe the small non-zero effect in the imagined contact hypothesis and scenario 2. However, the mean of the median of the effect estimate for scenario 2 computed by the two-sided version is closer to the real mean effect size than the one-sided version.

For scenario 3, the one-sided mean estimate has lower bias to the true effect.

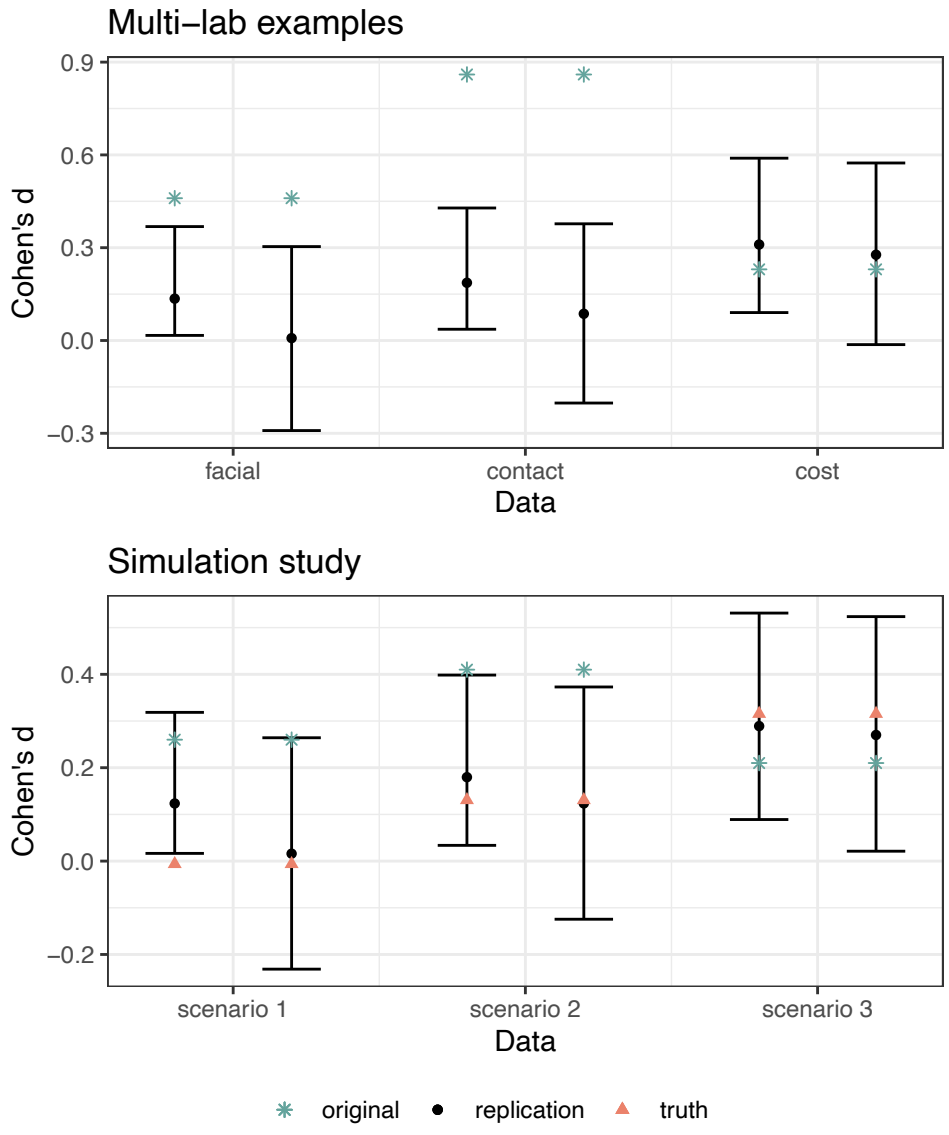

Figure 4.9: Posterior equal-tailed credibility interval with one- and two-sided JZS prior

4.5 Assessment of assumptions

After implementing the criteria on the multi-lab examples and the simulation data, the question arises whether the initial assumptions adopted by the criteria (see chapter 3.2.2) are actually fulfilled.

For the simulation study scenarios both assumptions - normality and variance homogeneity - are guaranteed by the simulation process itself.

In the following, all three multi-lab hypotheses are analyzed regarding normality and variance homogeneity between condition and control group.

4.5.1 Assumption of normality

Whether the mean effect difference is normally distributed or not is determined graphically and numerically leveraging

1. QQ-Plot
2. Shapiro-Wilks test: $H_0 : \overline{d} \sim N(\mu, \sigma^2)$

The Shapiro-Wilks test results in non-significant p-values for each hypothesis. Consequently, the null hypothesis that the mean effect difference follows a normal distribution cannot be rejected.

Hypothesis	**p-value**
facial	0.59
contact	0.44
cost	0.93

Table 4.21: Shapiro-Wilks test results

The points in the QQ-Plots lie more or less close to the bisection line and within the confidence interval for every hypothesis, see plots 4.10.

The boundary quantiles of the imagined contact hypothesis divert from the bisection line and its confidence interval.

Both methods combined allow accepting the normality assumption of the effect size for all three hypotheses.

Several Bayes factor based criteria (e.g. JZS Bayes factor, replication Bayes factor) require the individual observations in condition and control group across all replication studies to follow a normal distribution. This is assessed through visualizing the observations, separated for condition and control group, across all studies in a density plot. For the facial feedback and imagined contact hypothesis, the density plot proves a bell shaped distribution. For the sunk costs hypothesis, the normality assumption does not apply. The distribution seems to be suffering under a ceiling effect (Garin, 2014), showing that people are highly willing to attend a game by their favorite soccer team despite bad weather and independent of the applied treatment. This could indicate that selecting soccer - an emotional sport for many (Jones et al., 2012) - for the experimental set up is an unfortunate choice.

In conclusion, while the raw mean differences between condition and control group for the sunk cost hypothesis follow a normal distribution, neither the normality assumption for the individual observations nor the variance homogeneity are fulfilled.

It has to be to pointed out that while the criteria in question model the individual data points to be sampled from a normal distribution, they only utilize the t-test in their explicit calculation (Rouder et al., 2009). It

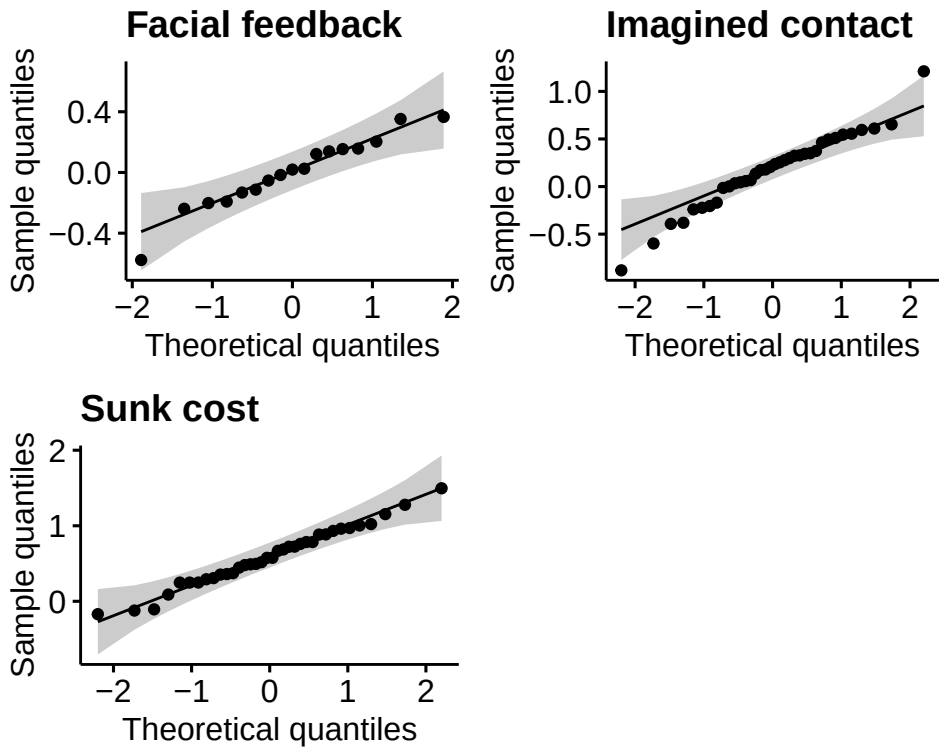

Figure 4.10: QQ-Plot for multi-lab hypotheses

can be argued that the corresponding t-distribution is valid despite non-normality due to the central limit theorem (Fahrmeir et al., 2016, Ghasemi and Zahediasl, 2012).

4.5.2 Assumption of equal variance

The two-sample t-test as well as the criteria computing Bayes factors assume identical variance for condition and control group. This assumption is guaranteed to be met for the simulation data.

For the multi-lab examples, the validity of the homogeneity assumption is assessed by plotting the difference in variance between condition and control group, Δ, in figure 4.12 and by performing a FK test of homogeneity of variances.

For the facial feedback hypothesis, homogeneity in variance is assumed by Wagenmakers, Beek, et al., 2016 and consequently maintained for our purposes. The difference in variance between condition and control group for the imagined contact hypothesis lies within the same range as for the facial feedback hypothesis. Hence, the conclusion of homogeneity seems justified. This is also corroborated for both hypotheses by the FK test. The overwhelming majority of replication studies - 0.941% for the facial feedback hypothesis, 0.972% for the imagined contact hypothesis - result in an insignificant p-value to the significance level of 0.05.

For the sunk cost hypothesis, Δ tends to have negative values, indicating a heteroscedastic variance with higher values in the control than in the condition group. However, positive differences do exist. The FK test obtains an equally inconclusive result. 0.472% of replication studies have an insignificant p-value, 0.528% a significant one.

In accordance with the other two multi-lab hypotheses and the inconclusive result, we assume variance homoscedasticity for the sunk cost hypothesis.

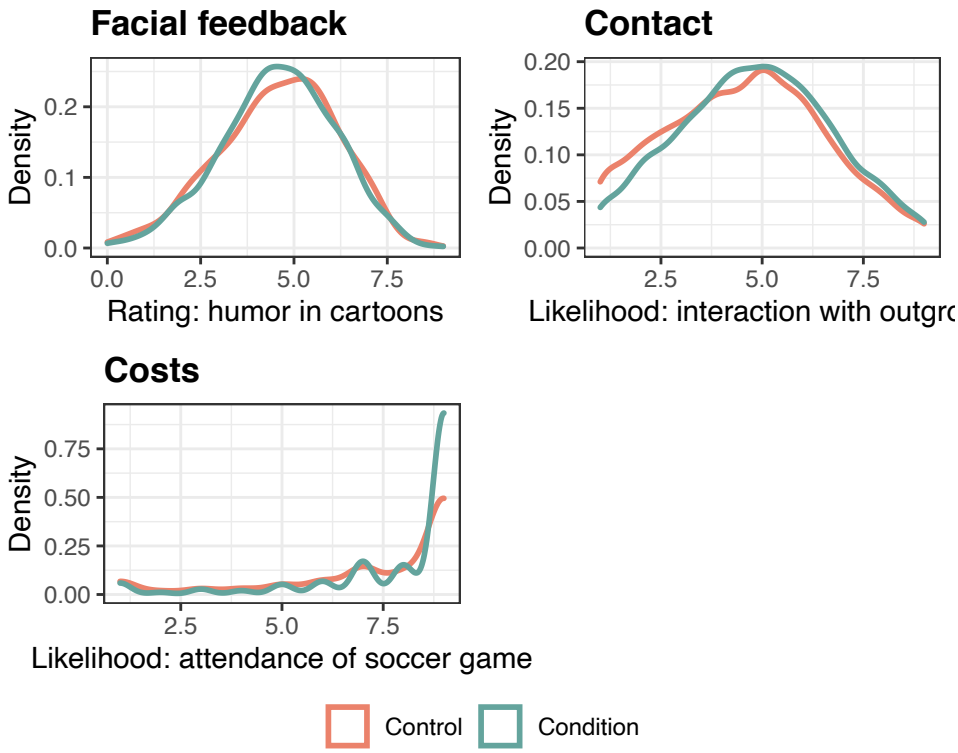

Figure 4.11: Density plot for multi-lab hypotheses

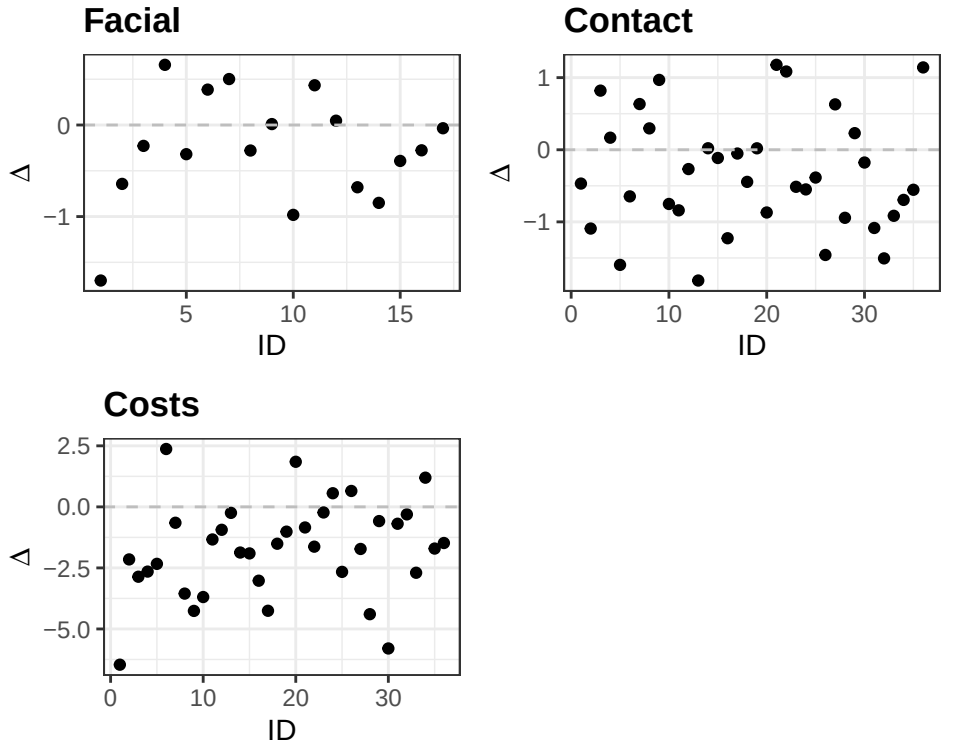

Figure 4.12: Difference Δ in variance between condition and control group for multi-lab hypotheses

Chapter 5

Discussion

The aim of this thesis was to investigate the performance of selected criteria to define replication success in light of potential distortions and different true effect sizes. To do so, we implement the criteria on three multi-lab examples as well as a simulation study including three scenarios. In order to explore the effects of distortions on the criteria, one example-scenario pair is affected by publication bias and another by selective reporting. The third serves as a control instance and is uninfluenced by any distorting factors.

In the following chapter, we continue to discuss the performance of the selected replication criteria after having embedded the result into the criticism given in the wider literature in chapter 4.

To recall, we are particularly interested in answering two questions:

1. Is there a zero effect and the original estimate an artefact of publication bias?
2. Is the original effect an overestimate due to selective reporting?

Hence, the assessment will in particular focus on each criterion's performance when either publication bias or selective reporting is present.

5.1 Discussion of criteria for one and multiple replications

5.1.1 Criticism on frequentist criteria

If the true parameter is (close to) zero and multiple replication studies are available, the criterion of *comparing the effect orientation* can detect publication bias by indicated an approximately equal distribution of studies between same and different direction. However, it cannot account for selective reporting since neither the concrete effect size, nor any significant discrepancy from the original effect estimate is taken into account. Simply comparing the signs of the estimates summarizes the information from the studies rather rudimentary - it disregards too much of reality.

The p-value criteria - be it the *simple p-value* or the *meta-analysis p-value* - do not allow us to account for any distorting factors. Whether a study leads to a significant estimate is highly dependent on the sample size - this contributes to the difference between p-value and meta-analysis p-value. The simple p-value also overestimates the evidence for the null hypothesis. Despite the fact that scenario 2 and 3 - as well as the imagined contact and sunk cost hypothesis - have a true non-zero effect size, at least a half of the replication studies obtain a non-significant p-value.

There is increasing criticism from the scientific community against the use of p-values to define replication success - and to perform hypothesis testing in general - due to the many weaknesses elaborated in chapter 3.2.4. The disadvantages of p-value based methods are well showcased in our analysis on the multi-lab examples as well as the simulation study. Relying on the p-values for decision-making leaves multiple open questions. When the replication turns out non-significant, does that mean the original study was a false-positive? Or is rather the replication study a false-negative (Laraway et al., 2019)? Deciding solely depending on the replication p-value omits all evidence collected in the original study (Van Aert and Van Assen, 2017). The extent to which p-values can be interpreted as an objective and informative representation is questionable, in particular since they depend on sample sizes and the statistical power of tests. Both are parameters which vary across replications (Verhagen and Wagenmakers, 2014).

Overall, the p-value and null hypothesis significance testing are more limited and problematic in their applicability than their popularity and ubiquitous usage let us believe. Hence, there is increasing opposition and lobbying against their widespread application and faulty interpretation (e.g. Goodman, 1999).

5.1.2 Criticism on Bayesian criteria

The *JZS Bayes factor* only exhibits a mediocre performance in determining replication success. As observable from the results obtained for the pair scenario 2 and imagined contact hypothesis, it is ill fit to detect a non-zero effect size if it is characterized by a small magnitude and within-study variance. Applying the JZS Bayes factor on the pooled data sets proves the dependence on sample size. The bigger the sample, the more likely the JZS Bayes factor is to pick up on small effects.

While the *equality-of-effect-size Bayes factor* can score with creating an explicit relationship between replication and original effect, it is ill-equipped to detect distortions and handle variation in treatment effects. This can be observed in the results for scenario 1 and the facial feedback hypothesis as well as scenario 2 and the imagined contact hypothesis. All parameters in said examples and scenarios are distorted by publication bias and selective reporting. However, the equality-of-effect-size Bayes factor wrongly finds at least weak evidence for equality between the replication and original estimate.

As aforementioned, the *replication Bayes factor* and its modification, the *evidence updating Bayes factor*, suffer from the so called replication paradox. This follows from the relative nature of its evidence (Ly et al., 2019) and is well illustrated in the results for scenario 2 and the imagined contact hypothesis - both subject to selective reporting. Despite a true non-zero effect size, the majority of Bayes factors exhibits at least weak evidence for the null hypothesis. Despite the deviation from zero for the replication estimate, the posterior distribution in H_1 explains the replication results worse than a distribution assuming a zero effect size. Consequently, the both Bayes factor versions are unsuited to determine replication success when the original study was subject to selective reporting. They perform better in detecting publication bias, as evident in the results for the facial feedback hypothesis. However, for scenario 1, the performance is less convincing. The replication Bayes factor and the evidence updating Bayes factor seem negatively affected by heterogeneity.

When considering all replication studies simultaneously, the fixed-effect meta-analysis Bayes factor and Bayesian evidence synthesis come to equal conclusions. Both criteria do not allow for any statistical exploration of the relationship between the original study and its replications but only to establish the existence - or non-existence - of a true non-zero effect.

5.1.3 Criticism on interval-based methods

Testing whether the credibility interval constructed based on the posterior density of the effect entails zero constitutes a combination of frequentist null hypothesis testing and Bayesian statistics. The 90% credibility interval describes - as the name suggests - the co-domain in which the effect size lies with a probability of 90%. We chose a significance level of 90% due to the one-sided nature of our analysis. We are interested in estimating a lower boundary which will allow us to assess the proximity between the effect size and zero. In theory, we could construct a one-sided credibility interval with lower limit $\alpha = 0.05$ quantile and upper limit infinity. However, the lower limit for the one-sided interval equals the lower limit of a two-sided interval with the significance level, $\alpha = 0.1$. As a benefit to the two-sided interval, the upper limit at the $1 - \frac{\alpha}{2}$ quantile can be leveraged to detect overestimation in the original estimate (Fahrmeir et al., 2016).

Identifying random noise from a true non-zero effect is challenging. In this thesis, we mitigate the problem by rounding the lower and upper boundary to two decimal places.

We assume two-sided priors for all interval-based methods due to the danger of artificially skewing the credibility interval. The comparison between a two- and one-sided JZS prior when constructing the posterior equal-tailed credibility interval in chapter 4.4.2 has indicated that a directional assumption in form of a one-sided prior leads to an overestimation of the effect size. For both, scenario 1 and the facial feedback hypothesis, it wrongfully concludes a significant non-zero effect. The effect estimates from multiple replication studies scatter around zero. The one-sided prior causes the positive effect estimates - even though they are positive due to random noise - to be interpreted as evidence for the alternative hypothesis.

For the truly non-zero effects in the remaing examples and scenarios, the one-sided prior draws the right conclusions and adds certainty by delivering intervals with a more shallow width.

Consequently, one-sided priors should only be leveraged if we are convinced that the effect is positive. If we are interested in establishing the existence of a non-zero effect size, a two-sided prior can be considered more appropriate.

Both the *posterior credibility interval* and the *HDI in isolation* perform inaccurately when a true positive effect is characterized by a small magnitude and accompanied by larger variance - as seen for the imagined contact and sunk cost hypothesis. The effect estimate incorporates too much uncertainty and hence the intervals have a large width. For scenario 2 and scenario 3, the performance is better. This can most probably be traced back to the increased sample size and the consequent smaller variance for the effect estimator. In general, remedy to increase certainty is found by leveraging information across the replication studies. The *HDI based on a hierarchical model* exhibit the shallowest width and middle point nearest to the true value. It uniformly concludes the correct effect estimate for all hypotheses and scenarios. Consequently, it should be preferred to the alternative two interval-based criteria.

Distortions could be conveniently detected through either visually comparing the original and the replication estimate or through testing whether the original estimate lies within the credibility interval.

One main discussion point resulting from the Bayesian nature of the interval-based criteria is the definition of prior distribution. For the posterior credibility interval, we have opted for a one-sided prior, identical to the one-sided JZS prior. The validity and usefulness of such assumption can be questioned and requires further attention.

5.1.4 Criticism on small telescope

The biggest disadvantage of the *small telescope* criterion is its reliance on classic frequentist hypothesis testing via p-values. Hence, small telescope can statistically only prove non-replicability - since it cannot accept the null hypothesis which entails the replicability claim.

The implementation of the small telescope method on the simulation scenario 1 and the facial feedback hypothesis indicates that the criterion is also unsuited to detect a true zero effect in the presence of within variance and heterogeneity. Even a small effect caused by random noise can obtain sufficient statistical power for the original study. The threshold of 33% is set extremely low, especially in comparison to the established target statistical power of 80%.

In addition, our analysis has shown that small telescope cannot identify selective reporting.

Overall, it seems as if the sceptical p-value is not well suited to detect non-zero effect sizes with the magnitude and variance we are observing.

However, the performance of the small telescope criterion can only be restrainedly evaluated by the simulation study. To obtain sufficient statistical power, the replication study requires a 2.5 times bigger sample size than the original study - a condition that is not fulfilled in any of the simulated scenarios and only few of the multi-lab replication studies.

5.1.5 Criticism on sceptical p-value

None of the replication studies - neither in the multi-lab examples nor in the simulated data - obtained a significant result when applying the *sceptical p-value* approach. The sceptical p-value seems unsuited for the low correlation coefficients and the high variability inherent in our data.

Its applicability is strongly limited by its condition that both original and replication effect estimate are required to be significant. Especially the latter is very infrequently significant (e.g. Open Science Collaboration, 2015).

5.2 The power of the Bayesian factor and snapshot hybrid

Overall, *Bayesian methods* allow for more a meaningful and flexible interpretation of their results. By returning a Bayes factor or a posterior probability, Bayesian criteria provide evidence for and against the null hypothesis while the ordinary frequentist methods prove absence of evidence - and not evidence of absence (Fahrmeir et al., 2016).

The Bayes factor can also better depict ambiguity in data when the obtained results speak neither for H_0 nor for H_1 convincingly (Verhagen and Wagenmakers, 2014). P-values, on the other hand, tend to overstate the evidence for the alternative hypothesis when data is ambiguous, suffering under Lindley's paradox.

However, some Bayesian methods fall short in detecting selective reporting and publication bias (replication Bayes factor and evidence updating Bayes factor) and in accounting for variation in treatment effects (equality-of-effect-size Bayes factor).

The comparisons between Bayes factors relying on either a one-sided or two-sided prior distribution in chapter 4.4.1 have illustrated that the two-sided version tends to interpret evidence more in favor of the null hypothesis than the one-sided one. This follows from the relative nature of the Bayes factor. Since the examined effects for the imagined contact and sunk cost hypothesis as well as scenario 2 and 3 have

a positive size, the alternative hypothesis assuming a-priori distribution which equally attributes non-zero probabilities to positive and negative effect sizes matches the data as inaccurately as the null hypothesis. It is hence a choice of the lesser evil. The one-sided prior distribution correctly modeling the positive sign of the effect matches said data better and is therefore accepted with a higher rate.

The comparison between one- and two-sided priors illustrates well why customizing the prior distribution according to the prior knowledge available of the effect is an essential step when implementing Bayesian methods. It optimizes performance and allows for a more detailed explanation of the underlying research theory.

We consider the *snapshot Bayesian hybrid meta-analysis* method by Van Aert and Van Assen, 2017 a very promising criterion. In a nutshell, it calculates the posterior probability for an effect of a certain size given the replication study data and the original data. If the posterior probability for a certain effect size lies above the threshold of 75%, we consider it substantial evidence for the effect size to be the true effect. The default effect sizes tested in snapshot hybrid are a zero effect (0), a small effect (0.1), a medium effect (0.3) and a large effect (r = 0.5).

Prior distribution
The default priori distribution is non-informative, i.e. each choice is allotted equal probability beforehand. Van Aert and Van Assen, 2017 prove that informative subjective prior probabilities can be easily incorporated into the snapshot hybrid method without having to repeat the entire calculation process. Prior probabilities can either remain discrete or can also be indicated in the form of a continuous distribution. Overall, snapshot hybrid has great flexibility in integrating prior assumptions regarding the potential effect size (Van Aert and Van Assen, 2017).

Discrete effect size values
Another criticism is the limited number of distinct effect size choices. This is especially disadvantageous when the true effect size lies between two given choices. In such situations, snapshot hybrid will deliver approximately equal evidence for both effect sizes - rendering a conclusive decision impossible. This phenomenon can be observed when applying snapshot hybrid to the sunk cost and imagined contact data as well as scenario 2 and 3 .

However, the principle of snapshot hybrid is independent of the number and magnitude of effect size options. In theory, they can be arbitrarily chosen by the researcher. We can regard this selection as the criterion's subjective component in accordance with the subjective prior choice in Bayesian methods. Based on expert knowledge, the researcher can hypothesize probable effect sizes and calculate their posterior probabilities based on replication and original data (Van Aert and Van Assen, 2017). The effect sizes tested in snapshot hybrid might also be set to values which impose some sort of relevance in the specific field - comparable to the dual-criterion design (Rosenkranz, 2021).

Identical sample sizes
One of the criticized points of snapshot hybrid is its inherent assumption of same sample sizes across replication and original paper. The authors of snapshot hybrid have developed a web application van Aert, 2021a in which unequal sample sizes for original and replication study can be incorporated. This feature is sadly not available in the R function *snapshot()* available in the puniform package (van Aert, 2021b).

Assumption of homogeneity and increase of sample size through pooling
Snapshot hybrid cannot account for any heterogeneity of the true effect size between original and replication study but rather considers all replication studies to be identically and independently drawn from one overarching distribution (Van Aert and Van Assen, 2017). As our analyses have indicated, there is hardly

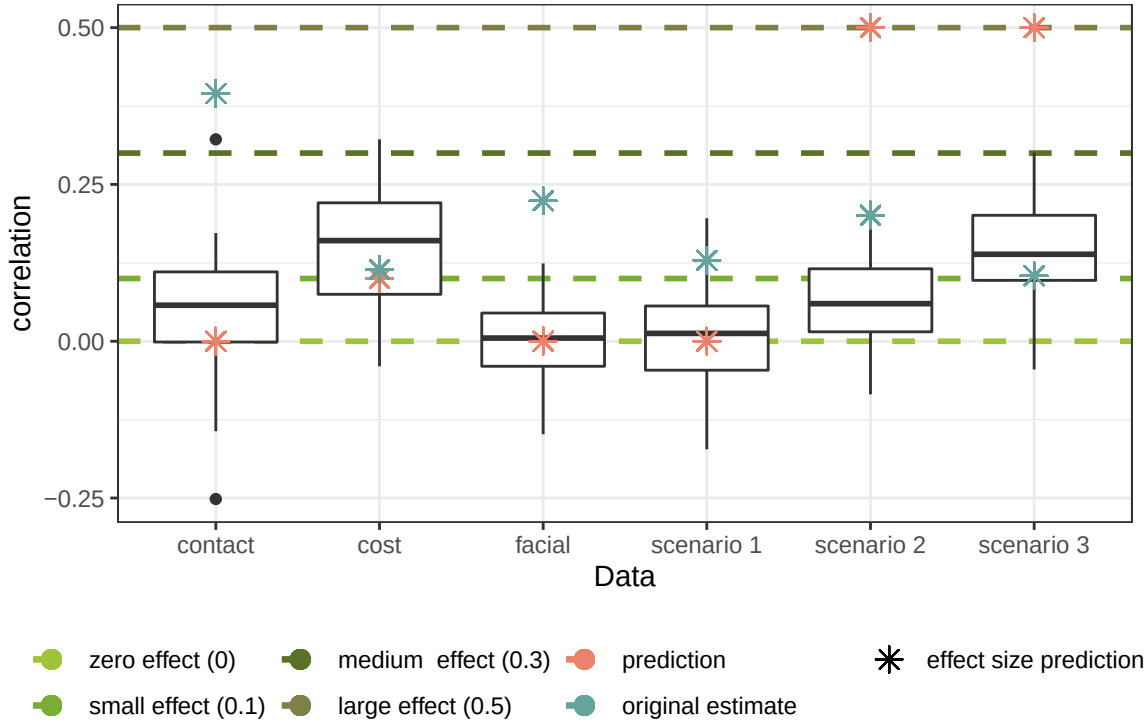

Figure 5.1: Correlation coefficients for all datasets and default effect sizes for snapshot hybrid

any measurable and certainly no significant heterogeneity in the multi-lab examples. Thus, the neglect of heterogeneity in snapshot hybrid can be justified. As a consequence, data across replication studies can be considered to originate from one overarching distribution, allowing us to pool the data.

Snapshot hybrid requires big sample sizes for sufficient statistical power - e.g. a sample size of 300 to 1000 to achieve 80% statistical power with a zero or small true effect size. Hence, the probability of inconclusive results is quite high for the combination of small sample size and small true effect size, customary in certain research areas (Van Aert and Van Assen, 2017). Pooling data serves as a remedy. Multiple smaller replication studies are combined to one large data set providing high statistical power when applying snapshot hybrid (Van Aert and Van Assen, 2017).

Ambiguity of data

Snapshot hybrid only considers posterior distributions over the threshold of 75% substantial evidence for a certain effect size. Other results are regarded inconclusive. Hence, snapshot hybrid accounts for ambiguity of data comparable to Bayes factors. It does not create a wrong sense of security as do other criteria, e.g. p-value significance (Van Aert and Van Assen, 2017).

Relationship between original and replication effect size estimate

Snapshot hybrid takes the original data set into account when calculating the posterior distribution. It additionally allows a comparison between the most likely effect estimate and the original estimate. Publication bias would result in a zero effect size as a prediction from snapshot hybrid while the original estimate has some positive significant effect. We can detect selective reporting if the predicted effect size based on snapshot hybrid has a lower value than the original estimate. This analysis can be guaranteed by choosing corresponding snapshot levels beforehand (Van Aert and Van Assen, 2017).

5.3 The strength of confidence and credibility intervals

Several authors emphasize that a single or a small number of replication studies do not provide enough information to determine replicability (e.g. Amrhein et al., 2018). Consequently, it is of great relevance to provide statistical criteria which determine replication success based on several replication studies simultaneously. However, the research for this thesis has primarily discovered methods for one replication study and only few designated for multiple. Even the multi-lab analysis experiments conducted by e.g. Wagenmakers, Beek, et al., 2016 and Schweinsberg et al., 2016 predominantly apply criteria for one replication study on all available replication studies individually.

If a limited number of replication studies are conducted, visual analysis methods are quick to apply, simple to comprehend and often surprisingly expressive.

Berkson's interocular traumatic test allows an easy assessment of publication bias and selective reporting. For both types of distortion, the original effect estimate lies at the outer edges and might even be considered an outlier.

Nevertheless, visual exploration can only be the first step in conducting a more thorough inference of the effect estimates across replication studies.

The construction of the non-central credibility interval combines the strengths of visualization with statistical inference. It enables detecting distortions due to publication bias and selective reporting as well as determining the most realistic empirical effect estimate - all well founded in a non-central t-distribution for the raw mean difference θ. However, its construction relies on pooling the data. Hence, it disregards all heterogeneity between replication studies. While this does not have any effects on the multi-lab examples since their heterogeneity is insignificant, it might render the criterion less suited to the simulation data and data with high variation in treatment effects in general.

An alternative method which incorporates heterogeneity is the HDI in a hierarchical model. Averaging the lower and upper limit as well as the median across all available replication studies leaves us with one overall interval. The HDI has the advantage of being a Bayesian credibility interval and with that in particular an easy and intuitive interpretation.

5.4 Success or failure - criticism of a binary classification

In the context of this thesis, we consider replication studies as either successful or unsuccessful. Several researchers (e.g. Gelman, 2018, Amrhein et al., 2018) have pleaded to substitute such binary classification with continuous measurements of how much the replication corroborates the original findings. Gelman, 2018 argues that instead of a binary classification researchers should "express the difference between old and new studies in terms of the expected variation in the treatment effect between conditions" (no page available). Anderson and Maxwell, 2016 even go so far as to claim that "[p]art of the reason for the crisis of confidence may be [...] viewing replication as a black and white dichotomy" (p.10). It is certainly true that a criterion which only performs binary classification - e.g. comparison of effect estimate orientation, p-values - oversimplifies the complex question of establishing replicability. While this might discredit certain methods it does not invalidate binary classification itself. We argue that the problem does not lie within the classification into success or failure to replicate per se but rather within the implementation of certain criteria that classify replication studies in a binary matter.

While we do acknowledge that scientific findings are never black-or-white, we are convinced that scientific findings require a deterministic element. A study is either replicable or not replicable within a certain

context - replicability is an all-or-nothing feature. Consequently, we hold on to a binary classification as the final conclusion when implementing the replication criteria.

Nevertheless, binary classification and continuous measurement of evidence for or against replicability are not mutually exclusive. The continuous measurement is utilized to perform classification and serves as valuable additional information acknowledging the uncertainty inherent to statistical inference. Essential for a theoretical interpretation of the research hypothesis in a wider sense is the binary classification.

Therefore, the focus should lie on applying metrics which identify replicability despite distortions and can be leveraged to classify a study as success or failure with high accuracy.

5.5 Limitations and further investigation

This thesis aims to provide a first step towards a systematical review of criteria which can be utilized to define replication success. While it is by no means an exhaustive investigation, the analysis of three multi-lab examples and three corresponding simulated scenarios allow an initial assessment of which criteria perform better given an arbitrary effect size and accommodate potential distortions - publication bias and selective reporting.

In future work, the selected criteria should be applied to more multi-lab examples and simulation scenarios. This would not only further consolidate the initial evaluation but also determine the performance in light of other distorting factors - such as p-hacking or HARKing. Throughout this thesis, we focused on publication bias and selective reporting in the interpretation and assessment of criteria performance. However, it is either impossible or very challenging to differentiate between distortion factors. Hence, the discrepancies between the original study and its replications encountered in our analysis might potentially originate from other influencing factors than interpreted. In order to further investigate distortions and their varying impact on the performance of replication criteria in more detail, more elaborate simulation studies are required (Hoffmann et al., 2020).

Thereby, we should also check for heterogeneity between replication studies to establish the prevalence of context dependency in multi-lab data and the influence it has on the different criteria.

How the individual method behaves given different values for the effect size, for the variance within a study and for the sample size can be examined by performing a more thorough simulation study in which these parameters are systematically varied across a range of reasonable options and combinations.

Throughout the thesis, we have touched upon the issue of sufficient sample sizes for the replication studies to obtain high statistical power for the respective criterion. Future work should further examine and compare the required size for substantial statistical power between criteria and hence investigate the practicability of the respective methods.

While the replication crisis is an interdisciplinary phenomenon, we assume the biggest, most effective and most sustainable impact is achieved when provided solutions are catered to the individual requirements, circumstances and idiosyncrasies of specific research fields. Hence, future reviews might incorporate area specific assessments and practical instructions on how to apply the replication criteria in day-to-day research.

Chapter 6

Conclusion

6.1 Suggestions and alternatives for the statistical evaluation of replication success

6.1.1 Replication report

Throughout this thesis, we have established the paramount importance of conducting replication studies. Different statistical criteria on how to define replication success have been applied to multi-lab examples and simulation data, analysing their suitability to handle potential distortion (selective reporting, publication bias).
The explicit purpose behind a replication study can be manyfold. "[R]eplication should be viewed as a construct that can be amenable to varying purposes and flexible in answering the questions that are most beneficial to moving [a] field forward in the domain of interest" (Anderson and Maxwell, 2016, p.10). Consequently, the various statistical approaches explore replicability in a different light. They contain different underlying conditions and assumptions, suffer from different weaknesses and shortcomings and answer different questions (Anderson and Maxwell, 2016, Verhagen and Wagenmakers, 2014). One method hardly suffices to grasp the complexity inherent to replication studies and can only be compared to other criteria to a certain extent.

Therefore, Marsman et al., 2017 argue "that future analyses of replication studies will be more inclusive by employing a range of different, complementary techniques." (p. 16)
Neither the criteria - if potential adaptions for multiple testing are considered - nor the purposes of testing are exclusive. Several criteria can and should be implemented simultaneously (Anderson and Maxwell, 2016). Including multiple replication metrics in "a comprehensive and coherent replication report" (Verhagen and Wagenmakers, 2014, p. 1469) will allow us to paint a more accurate and nuanced picture of replicability.

As discussed in chapters 5.1 and 5.3, the snapshot Bayesian hybrid meta-analysis method as well as the non-central confidence interval and the hierarchical credibility interval are considered promising and well equipped to conclude replicability despite publication bias and selective reporting. Hence, we would recommend incorporating them in the replication report.

6.1.2 Alternative approaches

The scope of methods described and analysed in this thesis is by no means exhaustive. Throughout the thesis, we limit ourselves to a one-sided t-test. Evidently, most methods can be expanded to incorporate other test statistics.

Of course, an abundance of alternative criteria to assess replication success exists in the frequentist as well as in the Bayesian framework. The interested reader might refer to Anderson and Maxwell, 2016 who identify six potential goals a replication study can pursue and give an extensive overview of methods that can be implemented to achieve each goal.

Beyond the existence of non-zero effects

Simonsohn, 2015 explains that "effects obtained in replication studies are currently evaluated almost exclusively on the basis of whether or not they are significantly different from zero." (p. 560) This also applied to the methods selected here. The majority of methods implemented in this thesis focus on assessing whether the replication study also indicates a true effect size that is significantly different from zero. However, this is only the first step. Our interest in effects expands from the simple question of existence to its actual size. Only the equality-of-effect-size Bayes factor by Bayarri and Mayoral, 2002, the replication Bayes factor by Verhagen and Wagenmakers, 2014 and its extension by Ly et al., 2019 as well as the snapshot hybrid method by Van Aert and Van Assen, 2017 allow more concrete conclusions on effect size than a simple classification into zero and not zero.

The replication crisis has showcased that empirical effect size estimates are subject to multiple distorting factors, leading to a systematic overestimation (e.g. Zwaan et al., 2018) While establishing the existence - or non-existence for that matter - of an effect is important, it is also of great interest to explore how replication studies can augment the accuracy in estimating the effect sizes. This is challenging - Van Aert and Van Assen, 2017 explain that "it is hard to obtain conclusive results about the magnitude of the true effect size in situations with sample sizes that are illustrative for current research practice." (p.14) Some of the methods proposed in this thesis make a good start and show great potential for further development.

6.1.3 Essential role of preregistration and disclosure

What has been neglected so far are the prerequisites necessary to perform a replication study - proper experimental logging and disclosure (Plesser, 2018, Nelson et al., 2018). Methodological approaches in data gathering and analysis have to be recorded to the extent of enabling other researchers to retrace the steps and reproduce original experiments (Zwaan et al., 2018). Additionally, full disclosure of the detailed results including all performed measures and exclusions is required (Nelson et al., 2018).
Vazire, 2018 points out that increased transparency in reporting experimental set-up, methods of data analysis and all results will play a essential role in guaranteeing replicability. Only by making studies more accessible through meticulous reporting can researchers comprehend and trust findings. Brandt et al., 2014 has provided a detailed replication recipe which instructs researchers on how to log their studies sufficiently.

One major remedy for HARKing and P-hacking is preregistration. Marsman et al., 2017 include preregistration as part of the "trinity of replication" (p. 2) next to collaboration with the original authors and high-powered studies.
According to Center for Open Science, 2021, over 280 scientific papers have adopted preregistration in the form of registered reports. A registered report is handed to the publishing paper prior to conducting the study, explaining the scientific hypothesis and respective study design as well as the planned data analysis and predicted results (Nosek and Errington, 2017, Vazire, 2018). This registered report is then peer reviewed, focusing not on the outcome of the study but its methodological correctness, and accepted for

publication. The publication of the study is therefore guaranteed independent of the final outcome and can only be revoked in case of concerns regarding quality and inexplicable procedural deviations from the preregistered approach.

6.2 Current situation and outlook

6.2.1 Concerns regarding replicability

The replication crisis and its inert accusations of poor science - unintentionally made but perceived as such (Zwaan et al., 2018) - have created a heated debate and considerable controversy in the scientific community (e.g. Zwaan et al., 2018, discussion between Ioannidis, 2005, Ioannidis, 2007 and Goodman and Greenland, 2007, discussion between Jager and Leek, 2014a, Jager and Leek, 2014b and Ioannidis, 2014).

Zwaan et al., 2018 enumerate some of the concerns regarding replicability as such, its assessment through various statistical criteria and its practicability in the scientific day-to-day.

Variability in context
As mentioned before, many argue that one main reason for lacking replicability in research is change in context - be it due to time passed since the original study, geographical differences or other tacit factors.

In the context of one of the research hypothesis - the facial feedback hypothesis - utilized in this thesis, the original author Strack, 2016 strives to undercut the evidence delivered by Wagenmakers, Beek, et al., 2016 through replication studies. By questioning if the cartoons which were used in the study and "were iconic for the zeitgeist of the 1980s instantiated similar psychological conditions 30 years later" (p.929), Strack, 2016 suggests the reason for non-replicability is merely historical change and not the non-existence of the effect described in Strack et al., 1988.

The danger behind this assumption of context variability is that it "renders the original theory unfalsifiable" (Zwaan et al., 2018, p. 6). Any failed direct replication study can be explained by deeming missing features post-hoc as essential to reproducing the effect. This attitude indicates wrongful hierarchical thinking, positioning the original study as more truthful or trustworthy than any replication study - without any justification but simple sequential ordering. In the logic of the critics, rejection results in a new, more specific hypothesis which embeds the effect in an increasingly complex context dependency. Every one of these refinements in hypothesis leads to more mediators added to the list of factors that must be included in a replication study (Klein et al., 2014). Following the theory of sophisticated falsificationism by Lakatos, 1970, aforementioned briefly, the study becomes degenerative. It is caught up in a "fruitless cycle of constantly invoking auxiliary hypotheses that fail to garner support" (Zwaan et al., 2018, p. 6) and are ideally dismissed as wrong (Nosek and Errington, 2017, Musgrave and Pigden, 2021). However, the custom of denying non-significant studies has factually lead to "virtually unkillable" theories (Ferguson and Heene, 2012, p. 559).
The importance of replicability as a method of falsification has to be acknowledged without finding a permanent excuse for failed replication in context dependency.

Limited value of direct replications
Another critical point held against direct replications in particular is its seemingly limited value creation. Eden, 2002 claims that the more original and replication study differ from each other, the higher the added value and potential insight created by the replication study. Direct replications either succeed or fail in their goal to replicate the original significant effect. Neither do they apply a new method, nor explore the hypothesis in a new setting, as conceptual replications do. Zwaan et al., 2018 argues that this perception

of direct replications not providing any new scientific discoveries is only valid as long as the significant original effect was in fact true. Given the current replication crisis and the conclusive evidence for missing replicability in many studies (e.g. Ioannidis, 2005, Open Science Collaboration, 2015), there is sufficient reason to doubt so. Conceptual replication has to, therefore, be replenished with direct replication to allow for a wholesome verification of a hypothesis (Zwaan et al., 2018).

Replication as a wrong focal point

Several voices in the scientific community (e.g. F. Schmidt and Oh, 2016) can be heard stating that the replication crisis should be less concerned with replication and more focused on other distorting factors - namely publication bias and questionable research practices (Zwaan et al., 2018). Overcoming these issues would solve the current crisis, so the argument. We disagree with this position. Firstly, distortion factors are embedded in the wider discussion surrounding replicability and while they might heavily contribute, they do not reflect the entire complexity underlying the replication crisis. Secondly, simply changing nomenclature has neither measurable impact nor does it support a holistic approach towards a solution to the replication crisis. Nevertheless, we admit that replication is "just one element of the toolbox of methodological reform" necessary for change (Zwaan et al., 2018, p. 9).

Confidence in wrong methodology

Direct replication studies per definition repeat the original study in its methodological approach. Of course, a central condition for their validity is the statistically justified and correct implementation of the approach. As illustrated in Rotello et al., 2015, an analysis performed in the original study can be unsound and lead to flawed results. Replicating such studies and repeating the misleading methodologies could consequently increase confidence in the wrong study results. According to Rotello et al., 2015, checking statistical assumptions before applying statistical methods to data is equally important as replicating studies. We would argue that ensuring a mathematically sound and correct implementation of test statistics is an irrefutable and even more fundamental component of good scientific practices than replicability. Therefore, whether all relevant assumptions are met has to be considered separately from the test for replicability.

6.2.2 Consequences of the replication crisis - in public and science

The effects of the replication crisis on the scientific community have been manyfold - some call it a "renaissance" (Nelson et al., 2018), some a "revolution" (Spellman, 2015). Others deny its existence or downplay its relevance (e.g. Stroebe and Strack, 2014). Again others see it as a degradation of science to a 'paint-by-numbers' process - simply reproducing studies by instructions instead of discovering new effects independently.

As the previous section 6.2.1 has hopefully indicated to the reader, the concerns regarding replicability, its exact definition and its relevance are abundant.

Fact is that the replication crisis justifiably questions the quality of research as a whole. Therefore, it has implications for science as a whole. It has endangered the societal stance of science and has diminished public trust and confidence in scientific findings and scientific expertise in general (German Research Foundation, 2017). The lack of replicability also undermines the responsibility of science to serve as a foundation for informed political decision making (Wingen et al., 2020). Consequently, conducting replication studies in the public eye is essential to enhancing the believability of science and to regaining trust in the scientific process (Laraway et al., 2019).

What will the tangible effects of the new sense of alert for replicability on researchers be? Heated debates have erupted and divided the scientific community (Chawla, 2016). It revolves around questions like,

- How many times must or should a study be replicated?
- When can a study be accepted as proven and how can research make new discoveries while concentrating on replicability?

Baumeister, 2016 and Vazire, 2018 agree that while science will become more sustainable in the sense that findings will be more substantial and replicable, the ingenuity and creativity of researchers might suffer. Replicability could be perceived as a hindrance, inducing researchers to be more risk-adverse and less bold in defining their hypothesis. Dorothy Bishop, a developmental neuropsychologist at the University of Oxford, even goes as far as to tweet, "the opposite of the reproducibility crisis! Stasis" (Chawla, 2016).

However, publishing false positive or overestimated effects is not much of scientific progress either.
Throughout this thesis, we have established the importance of replication studies - from a philosophical, societal and statistical viewpoint. While we have mainly focused on the latter and the more practical aspects of replication - how to statistically determine replication success - the replication crisis has to be considered as a wider, more complex phenomenon. It is a symptom of many shortcomings - not only by researchers and publishers. The public perception of science also pays a great contribution to it.

6.2.3 The battle against the myth of science

The public is suffering from the myth of science, i. e. the impression that scientific findings are certain and true and once established they are accepted by the scientific community for eternity. Researchers are suffering from the "myth that a scientist can do a perfect study" (Hunter, 2001, p. 151). For the public and the research community alike, science is understood to ask questions and finds answers without a shadow of a doubt (Gelman, 2015a).

This idealistic understanding of science and scientific studies neglect scientific reality. Science has always been shaped by new discoveries but also by self-improvement and self-correction. It is an evolving system of exploring, gaining new knowledge and testing established facts - all in the eternal quest for truth (Simmons et al., 2011). However, this truth cannot be determined with complete and unconditional assurance through statistical analysis - since statistics does not equal reality. Unfortunately though, statistics is the only tool available to capture at least some of reality (Białek, 2018). Statistical inference is a "thought experiment" (Amrhein et al., 2018, p. 2) assuming a simplified version of reality to construct models and draw predictive conclusions. Neglecting some of reality's complexity is essential to fit statistical models with a limited amount of data. Hence, as Vazire, 2018 summarizes, "error is inherent to the scientific progress" (p. 415) and uncertainty in the form of random sampling error or systematic distortion error will always remain.

However, this discrepancy between reality and statistics is all too often forgotten in science communication. Positioning scientific findings as absolutely certain leads to inevitable and profound misunderstanding (Gelman and Carlin, 2017). Therefore, the replication crisis might be perceived not as a crisis of science and scientific conduct but rather as a crisis of scientific communication, a denial of scientific uncertainty and an overestimation of statistical inference (Amrhein et al., 2018).

Consequently, researchers need to develop and practice a better and more veridical communication for their findings such that they are accessible to the layman (Laraway et al., 2019). This will empower the wider public to recalibrate their definition of science and accept uncertainty as one of its unavoidable components - and not as a sign of "junk science" (Białek, 2018, p.2, and Gelman, 2016).
Simultaneously, researchers have to contemplate their fundamental values shared across fields - the value of self-correction in light of unavoidable scientific error (Amrhein et al., 2018).

Only with an holistic approach including a shift in statistical practices as well as in mindsets will the

replication crisis be effectively and sustainably overcome. We are currently standing at the brink of a fundamental methodological and cultural transformation - to rejoice in an age of 'replication renaissance'. To quote Nelson et al., 2018, "it is clear that the Middle Ages are behind us, and the Enlightenment is just around the corner." (p. 529)

Appendix A

Mathematical and statistical foundations

A.1 Fundamental empirical estimators

The empirical mean is computed by

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{A.1}$$

The within study variance is computed by

$$s^2 = \frac{1}{n-1} \sum_{i=n}^{n} (x_i - \overline{x})^2 \tag{A.2}$$

If the sample includes two groups, we frequently require the pooled variance

$$s_p^2 = \frac{1}{n_1 + n_2 - 2} \left(s_1^2 \cdot (n_1 - 1) + s_2^2 \cdot (n_2 - 1) \right) \tag{A.3}$$

which is calculated based on the variance estimate and sample size for both groups respectively.

The standard deviation is the square root of the variance

$$s_p = \sqrt{s_p^2} \tag{A.4}$$

The covariance between two random variables X and Y is computed by

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}) \tag{A.5}$$

A.2 Cochran's Q test

Cochran's Q test is a popular statistic to test for heterogeneity between effect sizes across studies. It establishes whether a significant non-zero variance in treatment effect, τ^2, exists

$$H_0 : \tau^2 = 0 \tag{A.6}$$

It is calculated based on the empirical effect size and standard error estimates, $\widehat{\theta}_i$ and s_i, for all m replication studies and one original study

$$Q = \sum_{i=1}^{m+1} w_i \cdot (\widehat{\theta}_i - \overline{\theta}_w)^2 \tag{A.7}$$

where $\overline{\theta}_w$ denotes a weighted mean of the mean difference estimator, $\widehat{\theta}_i$. The weights are determined by the reciprocal of the corresponding variance $w_i = \frac{1}{\sigma_i^2}$

$$\overline{\theta}_w = \frac{\sum_{i=1}^{m+1} w_i \widehat{\theta}_i}{\sum_{i=1}^{m+1} w_i} \tag{A.8}$$

Since σ_i^2 is unknown, we estimate $\hat{w}_i = \frac{1}{s_i^2}$ (Hoaglin, 2016).

Under the null, Q is assumed to follow a χ^2 distribution with m degrees of freedom (Borenstein et al., 2021).

However, Hoaglin, 2016 argues that the Q statistic was not designed by its inventor, William G. Cochran, to test for heterogeneity and is hence unsuited to be used as such.

Nevertheless, Cochran's Q statistic enjoys great popularity and will be leveraged in this thesis to establish heterogeneity between the original and replication effect size estimates.

A.3 Correlation

Correlation is an alternative method to measure effect size (Borenstein et al., 2021). Its benefits include being bound by an upper and a lower limit, its wide popularity and hence accessible interpretation (Open Science Collaboration, 2015).

Correlation indicates how accurately one of two continuous variables X and Y can be explained as a linear transformation of the other, i.e.

$$Y = a \cdot X + b \tag{A.9}$$

The two variables run together: if one increases, the other one either increases ($a > 0$) or decreases ($a < 0$) proportionally.

Correlation is empirically measured by the Bravais-Pearson correlation coefficient. The covariance between the two variables is divided by their respective standard deviations

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} \tag{A.10}$$

The approximation depends on the empirical estimate of the covariance, s_{XY}, between X and Y and the

empirical variance estimate for X and Y (Fahrmeir et al., 2016)

$$\hat{\rho}_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} \tag{A.11}$$

The correlation coefficient lies within the range $[-1, 1]$. It takes the values of the boundary if the relationship between X and Y is perfectly linear. The less a linear relationship matches the data, the closer to zero the correlation coefficient is (Fahrmeir et al., 2016).
The variance to the Bravais-Pearson correlation coefficient ρ depends on the estimate itself and the sample size (Borenstein et al., 2021)

$$\sigma_\rho^2 = \frac{(1 - \hat{\rho}_{XY}^2)^2}{n - 1} \tag{A.12}$$

According to Borenstein et al., 2021, the empirical Cohen's d estimate, d, which describes the effect of one variable on the other, can be leveraged to determine the correlation coefficient

$$\widehat{\rho}_{XY} = \frac{d}{\sqrt{d^2 + a}} \tag{A.13}$$

If the sample sizes of the condition and control group are known, we insert

$$a = \frac{(n_{cond} + n_{ctr})^2}{n_{cond} \cdot n_{ctr}} \tag{A.14}$$

If the exact distribution is unknown, we assume equal group sizes and hence $a = 4$ (Borenstein et al., 2021).

While establishing correlation requires two continuous variables, measuring a certain variable X in a dichotomous experiment - consisting of a condition and control group - only provides one. Consequently, we assume that a continuous variable Y underlies the allocation to either control or condition group. This allows us to follow the established interpretation: if X increases, the hidden Y either increases or decreases proportionally, visibly leading to a corresponding allocation to either condition and control group (Borenstein et al., 2021). Thanks to above relationship between Cohen's d and the correlation coefficient ρ_{XY}, the calculation of $\hat{\rho}_{XY}$ is fairly simple compared to its interpretation.

Next to already mentioned benefits, effect sizes are frequently captured as correlation coefficient to ensure its normal distribution. This is provided by applying the Fisher transformation to ρ_{XY}.

A.4 Fisher transformation

It is common in meta-analysis to apply the so called Fisher transformation to obtain Fisher's z

$$z = tanh^{-1}(\widehat{\rho}_{XY}) \tag{A.15}$$

$$= \frac{1}{2}\, ln\left(\frac{1 + \widehat{\rho}_{XY}}{1 - \widehat{\rho}_{XY}}\right) \tag{A.16}$$

from the correlation estimate $\widehat{\rho}_{XY}$. Fisher's z always follows a normal distribution - independent of the distribution of the individual observations (Held et al., 2020, Borenstein et al., 2021). It stabilizes the variance of the correlation coefficient. As we have seen, the variance of $\widehat{\rho}_{XY}$ depends on sample size and

its own value. The variance to Fisher's z is solely determined by the sample size n (Borenstein et al., 2021)

$$\sigma_z = \frac{1}{n-3} \tag{A.17}$$

A.5 Bayes Theorem

Bayes Theorem was discovered by Thomas Bayes in the 18^{th} century. It uses reverse probability to estimate the probability of an antecedent event A based on the occurrence of the a subsequent event B (Bolstad and Curran, 2016).

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{A.18}$$

where $P : \phi \to \mathbb{R}$ is some probability function and $A, B \in \phi$ some event.

The fundamental idea behind Bayesian statistics is to consider the parameter of interest, γ, a random variable and not a fixed value. γ therefore follows a certain distribution. The aim is to estimate this distribution - or more precisely the parameters characterising it - as accurately as possible. The a-priori distribution $P(A) = p_{priori}(\gamma)$ is assigned before the data X is collected and represents a particular subjective belief in γ, i.e. which values we assume are most likely. By assigning such a prior distribution, we acknowledge the uncertainty of its exact value (Marsman et al., 2017).
The belief in the a-priori distribution is updated by the observed data through multiplication with the likelihood of γ. The likelihood $\ell(\gamma)$ is synonymous to the density function f of X - however, we regard x fixed and γ variable (Kauermann and Hothorn, 2020, Moreno, 2005).

We leverage the Bayes Theorem (Bolstad and Curran, 2016).

$$p_{post}(\gamma) = P(\gamma|x) = \frac{P(x|\gamma) \cdot P(\gamma)}{P(x)} = \frac{f(x|\gamma) \cdot p_{priori}(\gamma)}{f(x)} = \frac{p_{priori}(\gamma) \cdot \ell(\gamma)}{f(x)} \tag{A.19}$$

The posterior distribution is thus proportional to the product of a-priori distribution and likelihood (Bolstad and Curran, 2016) and we can simplify to,

$$p_{post}(\gamma) \propto p_{priori}(\gamma) \cdot \ell(\gamma) \tag{A.20}$$

Marsman et al., 2017 fittingly summarize the construction of a posterior distribution as follows: "Hence, the Bayesian updating process is guided by predictive success: parameters that predict [the observed data, AN] well receive a boost in plausibility, whereas parameters that predict poorly suffer a decline" (p. 3).

Appendix B

Descriptive analysis of multi-lab examples

The table below gives an overview over all replication studies and their key statistics for each hypothesis.

B.1 Facial feedback

The facial feedback hypothesis was analyzed in 17 replication studies. The mean difference between mean ratings on how funny the cartoons were perceived for condition and control group across all replication studies, weighted according to the study size, is 0.014. The median is 0.019. The minimum difference in rating is -0.577, the maximum difference in rating is 0.366 for the replication study with ID 12 and 10 respectively. The variation in ratings between condition and control group within each study ranges from a minimum value of 1.498 to a maximum value of 3.793. The median is 2.111.

The effect size measured in Cohen's d spans across -0.3 on the lower and 0.25 on the upper end.

The table B.1 gives an overview over the group sizes, mean differences, group variances (var) and Cohen's d effect size estimate for the condition and control group in each replication study.

	condition			control					
ID	mean	var	size	mean	var	size	mean difference	pooled var	Cohen's d
1	4.20	1.70	67	4.06	3.40	72	0.14	2.58	0.09
2	5.05	2.44	55	4.89	3.08	70	0.16	2.80	0.09
3	4.69	1.80	58	4.70	2.03	57	-0.02	1.92	-0.01
4	4.61	2.31	50	4.49	1.65	51	0.12	1.98	0.09
5	4.91	2.36	59	5.02	2.68	58	-0.11	2.52	-0.07
6	5.01	2.38	47	5.06	2.00	47	-0.05	2.19	-0.04
7	4.91	2.21	49	4.71	1.71	50	0.20	1.95	0.15
8	4.93	1.75	50	5.12	2.03	50	-0.19	1.89	-0.14
9	4.14	2.95	51	4.12	2.94	50	0.02	2.94	0.01
10	4.54	2.03	59	4.18	3.01	67	0.37	2.55	0.23
11	4.63	2.18	53	4.87	1.75	57	-0.24	1.96	-0.17
12	3.77	3.81	52	4.34	3.77	35	-0.58	3.79	-0.30
13	3.78	2.72	61	3.91	3.40	59	-0.13	3.05	-0.08
14	4.36	1.69	57	4.34	2.54	55	0.02	2.11	0.02
15	4.94	1.30	65	4.79	1.69	65	0.15	1.50	0.13
16	4.75	1.94	53	4.95	2.21	57	-0.20	2.08	-0.14
17	4.93	1.95	50	4.58	1.99	58	0.35	1.97	0.25

Table B.1: Replication studies for facial feedback hypothesis

B.2 Imagined contact hypothesis

36 replication studies research the imagined contact hypothesis. Condition and control group vary in the mean rating how likely the participant is to interact with a minority group. The difference has a lower boundary of -0.88 and an upper boundary 1.211 with a weighted mean of 0.258 and median 0.22. The random error within each study has a minimum value of 1.843 and a maximum value of 4.406. The median within-variance is 3.365. The effect size measured as Cohen's d ranges from -0.52 to 0.68. The median value lies at 0.115.

The table B.2 summarizes the key empirical estimates for all replication studies.

	condition			control					
ID	mean	var	size	mean	var	size	mean difference	pooled var	Cohen's d
1	5.01	4.01	50	4.78	4.48	34	0.24	4.20	0.11
2	4.94	3.82	56	4.77	4.92	64	0.17	4.41	0.08
3	4.68	2.67	49	4.69	1.85	35	-0.01	2.33	-0.01
4	5.36	3.29	46	4.98	3.12	49	0.37	3.20	0.21
5	4.02	3.08	47	3.96	4.67	49	0.06	3.89	0.03
6	4.50	2.83	51	4.90	3.48	51	-0.39	3.16	-0.22
7	5.44	3.45	47	4.23	2.82	43	1.21	3.15	0.68
8	4.64	3.38	97	4.31	3.08	77	0.33	3.25	0.18
9	5.31	2.64	58	5.55	1.68	55	-0.24	2.17	-0.16
10	4.99	3.71	56	4.95	4.46	56	0.04	4.09	0.02
11	5.27	3.40	140	4.92	4.24	137	0.35	3.82	0.18
12	5.47	3.18	72	5.19	3.45	74	0.28	3.32	0.15
13	5.11	2.33	46	4.76	4.15	52	0.34	3.30	0.19
14	5.45	3.45	44	4.84	3.43	41	0.61	3.44	0.33
15	4.42	3.73	482	3.91	3.84	518	0.51	3.79	0.26
16	3.71	3.01	55	3.92	4.24	52	-0.20	3.61	-0.11
17	6.16	2.46	61	5.69	2.51	61	0.46	2.48	0.29
18	5.32	3.50	662	5.02	3.95	666	0.30	3.73	0.15
19	4.32	2.83	51	3.82	2.81	44	0.49	2.82	0.29
20	4.16	3.91	53	4.76	4.78	48	-0.60	4.32	-0.29
21	4.57	4.55	47	3.92	3.37	38	0.65	4.02	0.33
22	4.85	3.93	81	4.30	2.84	81	0.56	3.39	0.30
23	4.51	2.71	43	3.92	3.23	36	0.59	2.94	0.35
24	4.72	3.39	73	4.54	3.94	95	0.18	3.70	0.09
25	4.24	3.73	94	4.41	4.12	93	-0.17	3.92	-0.09
26	4.39	3.04	37	4.62	4.50	50	-0.22	3.88	-0.11
27	4.07	3.26	127	3.52	2.64	98	0.55	2.99	0.32
28	3.72	2.41	43	4.60	3.35	37	-0.88	2.84	-0.52
29	4.60	3.10	70	4.39	2.87	57	0.21	3.00	0.12
30	5.14	3.57	69	4.81	3.75	74	0.32	3.67	0.17
31	4.86	2.77	34	4.79	3.85	47	0.07	3.40	0.04
32	5.14	3.48	45	5.14	4.98	62	0.00	4.35	0.00
33	4.56	2.24	47	4.43	3.16	49	0.14	2.71	0.08
34	4.11	3.03	57	4.49	3.73	46	-0.38	3.34	-0.21
35	5.45	1.58	47	5.41	2.13	43	0.03	1.84	0.03
36	4.84	3.81	35	4.58	2.67	52	0.25	3.13	0.14

Table B.2: Replication studies for imagined contact hypothesis

B.3 Sunk cost hypothesis

The sunk cost hypothesis was explored in 36 replication studies. Its mean difference between condition and control group ranges from -0.17 to 1.496 with a weighted mean of 0.604 and median 0.576. The within-variance has a minimum value of 1.8 and a maximum value of 6.792. The median variance is 3.53.

The effect size measured as Cohen's d has a lower limit of -0.08, a median of 0.325 and an upper limit of 0.68.

The table B.3 gives an overview over key characteristics for all replication studies for the sunk cost hypoth-

esis.

	condition			control					
ID	mean	var	size	mean	var	size	mean difference	pooled var	Cohen's d
1	8.28	1.75	36	7.00	8.21	48	1.28	5.45	0.55
2	7.83	4.89	65	7.25	7.05	55	0.58	5.88	0.24
3	7.80	2.93	35	6.80	5.79	49	1.00	4.60	0.47
4	8.22	1.60	50	7.53	4.25	45	0.69	2.86	0.41
5	8.34	1.70	50	7.46	4.03	46	0.88	2.82	0.53
6	7.74	5.00	58	7.86	2.63	44	-0.12	3.98	-0.06
7	8.34	2.00	35	7.98	2.65	55	0.36	2.40	0.23
8	8.51	1.35	81	7.48	4.90	93	1.02	3.25	0.57
9	7.71	3.19	52	6.56	7.45	61	1.15	5.49	0.49
10	8.36	1.20	55	7.58	4.89	57	0.78	3.08	0.45
11	8.21	2.81	155	7.45	4.15	121	0.76	3.40	0.41
12	8.15	2.87	74	7.65	3.81	72	0.50	3.33	0.27
13	8.12	2.63	51	7.87	2.87	46	0.25	2.74	0.15
14	8.10	3.73	40	7.38	5.60	45	0.72	4.73	0.33
15	7.30	5.66	522	6.52	7.57	477	0.78	6.57	0.31
16	7.79	2.82	47	6.82	5.85	60	0.97	4.52	0.46
17	8.38	2.53	56	6.88	6.79	66	1.50	4.83	0.68
18	7.37	6.06	683	6.90	7.57	640	0.48	6.79	0.18
19	8.31	2.18	48	8.06	3.19	47	0.25	2.68	0.15
20	7.58	5.29	53	7.76	3.44	49	-0.17	4.40	-0.08
21	7.73	5.85	41	7.38	6.69	45	0.35	6.29	0.14
22	8.53	1.06	88	8.16	2.69	74	0.37	1.80	0.28
23	8.15	2.39	40	8.26	2.62	39	-0.11	2.50	-0.07
24	7.89	4.32	90	7.64	3.77	78	0.25	4.07	0.12
25	8.47	1.39	94	7.89	4.05	93	0.58	2.72	0.35
26	7.53	6.38	47	6.60	5.73	40	0.93	6.09	0.38
27	8.46	1.62	131	7.97	3.34	94	0.49	2.34	0.32
28	8.51	0.70	37	7.63	5.10	43	0.89	3.07	0.51
29	8.05	3.98	61	7.74	4.56	66	0.31	4.28	0.15
30	8.25	1.76	81	7.29	7.56	63	0.96	4.30	0.46
31	8.15	2.44	46	7.43	3.13	35	0.72	2.74	0.44
32	8.41	2.36	54	7.89	2.67	54	0.52	2.51	0.33
33	7.60	3.35	48	7.31	6.05	48	0.29	4.70	0.13
34	7.73	4.34	44	7.64	3.15	58	0.09	3.66	0.05
35	8.10	2.22	48	7.66	3.93	41	0.45	3.01	0.26
36	8.47	1.86	47	7.80	3.34	40	0.67	2.54	0.42

Table B.3: Replication studies for sunk cost hypothesis

Appendix C

Overview of criteria results

For each multi-lab hypothesis, several replication criteria are applied and calculated. The majority of these criteria can only account for one replication study. Replication success when considering all replication studies is assessed through the relative frequency of individual studies delivering evidence for or against replication success.

In the tables below, an overview of the results for each individual replication study is given. The rows indicate the different replication studies, the columns enumerate the criteria according to which replication success is determined.

The entries consist of

- p-values for columns 2, 3, 8 and 9
- Bayes factor for columns 4, 5, 6 and 7
- Predicted effect size and posterior probability for column 10

In column 8, the p-value to the t-test performed in the small telescope criteria is indicated. If it is below the significance value $\alpha = 0.05$, the replication estimate is deemed to small to be detected with sufficient statistical power and hence the study considered non-replicable.

For column 9, the sceptical p-value criteria, an entry of 1 indicates the lack of significance of the replication effect size. Hence, the sceptical p-value criterion could not be implemented since its precondition is not fulfilled.

Column 10 entails the effect level, for which the snapshot hybrid method calculates the highest posterior density. To recall, possible levels are p.0 (zero effect), p.sm (correlation 0.1), p.m (correlation 0.3) and p.l (correlation 0.5).

Study ID	orientation	p-value	Meta-analysis p-value	JZS BF	Equality-of-effect-size BF	Replication BF	Small telescope	Sceptical p-value	Snapshot
1	TRUE	0.31	0.08	0.28	2.70	0.27	0.22	1.00	inconclusive
2	TRUE	0.30	0.07	0.30	2.81	0.27	0.25	1.00	inconclusive
3	FALSE	0.53	0.12	0.19	1.86	0.24	0.11	1.00	inconclusive
4	TRUE	0.33	0.06	0.30	2.77	0.26	0.26	1.00	inconclusive
5	FALSE	0.65	0.20	0.15	1.38	0.26	0.06	1.00	p.0 with 0.776
6	FALSE	0.57	0.12	0.19	1.76	0.24	0.11	1.00	inconclusive
7	TRUE	0.24	0.04	0.40	3.32	0.31	0.36	1.00	inconclusive
8	FALSE	0.76	0.21	0.13	1.04	0.30	0.04	1.00	p.0 with 0.807
9	TRUE	0.48	0.12	0.22	2.08	0.24	0.15	1.00	inconclusive
10	TRUE	0.10	0.02	0.71	4.16	0.54	0.53	1.00	inconclusive
11	FALSE	0.81	0.27	0.12	0.82	0.36	0.02	1.00	p.0 with 0.849
12	FALSE	0.91	0.48	0.11	0.47	0.60	0.01	1.00	p.0 with 0.867
13	FALSE	0.66	0.23	0.15	1.34	0.26	0.06	1.00	p.0 with 0.785
14	TRUE	0.46	0.10	0.21	2.09	0.24	0.15	1.00	inconclusive
15	TRUE	0.24	0.04	0.36	3.12	0.31	0.30	1.00	inconclusive
16	FALSE	0.77	0.24	0.13	0.98	0.31	0.03	1.00	p.0 with 0.826
17	TRUE	0.10	0.02	0.77	4.33	0.56	0.57	1.00	inconclusive

Table C.1: Overview over criteria to define replication success: facial feedback hypothesis

Study ID	orientation	p-value	Meta-analysis p-value	JZS BF	Equality-of-effect-size BF	Replication BF	Small telescope	Sceptical p-value	Snapshot
1	TRUE	0.30	0.09	0.36	1.16	0.14	0.20	1.00	inconclusive
2	TRUE	0.33	0.15	0.28	0.90	0.14	0.11	1.00	inconclusive
3	FALSE	0.52	0.14	0.22	0.69	0.12	0.07	1.00	inconclusive
4	TRUE	0.16	0.04	0.57	1.52	0.20	0.33	1.00	inconclusive
5	TRUE	0.44	0.17	0.24	0.77	0.12	0.09	1.00	inconclusive
6	FALSE	0.87	0.49	0.11	0.22	0.23	0.01	1.00	p.0 with 0.844
7	TRUE	0.00	0.00	37.49	3.91	17.92	0.96	0.39	inconclusive
8	TRUE	0.12	0.08	0.56	1.25	0.25	0.21	1.00	inconclusive
9	FALSE	0.81	0.37	0.12	0.28	0.18	0.01	1.00	p.0 with 0.823
10	TRUE	0.45	0.20	0.22	0.71	0.12	0.07	1.00	inconclusive
11	TRUE	0.07	0.10	0.70	1.16	0.37	0.15	1.00	inconclusive
12	TRUE	0.18	0.09	0.42	1.15	0.19	0.18	1.00	inconclusive
13	TRUE	0.17	0.05	0.51	1.44	0.19	0.29	1.00	inconclusive
14	TRUE	0.07	0.01	1.13	2.17	0.38	0.55	1.00	inconclusive
15	TRUE	0.00	0.06	622.53	1.52	604.57	0.26	0.46	p.sm with 0.998
16	FALSE	0.71	0.35	0.14	0.39	0.14	0.02	1.00	p.0 with 0.762
17	TRUE	0.05	0.02	1.18	1.94	0.45	0.48	1.00	inconclusive
18	TRUE	0.00	0.17	6.19	0.94	6.41	0.00	0.47	p.sm with 0.972
19	TRUE	0.08	0.02	0.97	1.98	0.34	0.48	1.00	inconclusive
20	FALSE	0.92	0.65	0.09	0.15	0.34	0.00	1.00	p.0 with 0.88
21	TRUE	0.07	0.02	1.10	2.17	0.37	0.54	1.00	inconclusive
22	TRUE	0.03	0.03	1.78	1.95	0.76	0.50	0.44	inconclusive
23	TRUE	0.07	0.01	1.19	2.29	0.40	0.58	1.00	inconclusive
24	TRUE	0.27	0.16	0.28	0.87	0.15	0.09	1.00	inconclusive
25	FALSE	0.72	0.45	0.11	0.35	0.14	0.00	1.00	p.0 with 0.859
26	FALSE	0.70	0.32	0.16	0.42	0.14	0.03	1.00	inconclusive
27	TRUE	0.01	0.02	3.81	1.95	1.87	0.54	0.44	inconclusive
28	FALSE	0.99	0.78	0.08	0.04	1.74	0.00	1.00	p.0 with 0.92
29	TRUE	0.25	0.10	0.34	1.04	0.15	0.15	1.00	inconclusive
30	TRUE	0.16	0.08	0.48	1.23	0.20	0.21	1.00	inconclusive
31	TRUE	0.43	0.13	0.27	0.86	0.12	0.12	1.00	inconclusive
32	TRUE	0.50	0.23	0.21	0.66	0.12	0.06	1.00	inconclusive
33	TRUE	0.34	0.10	0.30	0.98	0.13	0.14	1.00	inconclusive
34	FALSE	0.85	0.49	0.11	0.23	0.21	0.01	1.00	p.0 with 0.835
35	TRUE	0.45	0.10	0.24	0.78	0.12	0.10	1.00	inconclusive
36	TRUE	0.26	0.07	0.40	1.27	0.15	0.24	1.00	inconclusive

Table C.2: Overview over criteria to define replication success: imagined contact hypothesis

Study ID	orientation	p-value	Meta-analysis p-value	JZS BF	Equality-of-effect-size BF	Replication BF	Small telescope	Sceptical p-value	Snapshot
1	TRUE	0.00	0.01	6.31	3.04	11.58	0.98	0.44	inconclusive
2	TRUE	0.10	0.02	0.74	7.14	1.33	0.74	1.00	inconclusive
3	TRUE	0.01	0.01	3.06	4.19	5.19	0.95	0.45	inconclusive
4	TRUE	0.03	0.01	2.31	5.02	3.97	0.91	0.45	inconclusive
5	TRUE	0.01	0.01	7.62	3.27	14.68	0.97	0.45	inconclusive
6	FALSE	0.62	0.12	0.17	3.46	0.60	0.18	1.00	p.0 with 0.878
7	TRUE	0.13	0.04	0.63	6.36	1.03	0.71	1.00	inconclusive
8	TRUE	0.00	0.00	176.38	2.06	477.77	1.00	0.45	inconclusive
9	TRUE	0.00	0.00	8.03	3.70	16.06	0.98	0.45	inconclusive
10	TRUE	0.01	0.01	4.72	4.48	9.01	0.96	0.45	inconclusive
11	TRUE	0.00	0.00	59.65	5.35	167.39	0.99	0.45	p.sm with 0.831
12	TRUE	0.05	0.01	1.15	7.47	2.19	0.82	1.00	inconclusive
13	TRUE	0.23	0.04	0.41	6.41	0.75	0.56	1.00	inconclusive
14	TRUE	0.06	0.02	1.16	5.91	1.84	0.84	1.00	inconclusive
15	TRUE	0.00	0.00	12550.33	9.25	59769.13	1.00	0.46	p.sm with 0.998
16	TRUE	0.01	0.01	4.56	4.36	8.53	0.96	0.45	inconclusive
17	TRUE	0.00	0.00	175.56	1.06	460.13	1.00	0.44	inconclusive
18	TRUE	0.00	0.01	30.47	10.49	150.52	0.88	0.47	p.sm with 0.987
19	TRUE	0.23	0.05	0.42	6.42	0.76	0.57	1.00	inconclusive
20	FALSE	0.66	0.13	0.16	3.10	0.62	0.16	1.00	p.0 with 0.89
21	TRUE	0.26	0.06	0.40	6.08	0.71	0.54	1.00	inconclusive
22	TRUE	0.05	0.01	1.34	7.56	2.67	0.83	0.45	inconclusive
23	FALSE	0.62	0.10	0.19	3.44	0.60	0.21	1.00	p.0 with 0.861
24	TRUE	0.21	0.04	0.35	6.97	0.79	0.51	1.00	inconclusive
25	TRUE	0.01	0.00	4.38	6.82	9.64	0.94	0.45	inconclusive
26	TRUE	0.04	0.02	1.63	5.46	2.65	0.89	0.45	inconclusive
27	TRUE	0.01	0.00	4.01	7.45	9.29	0.92	0.45	p.sm with 0.752
28	TRUE	0.01	0.01	4.01	3.63	6.95	0.96	0.44	inconclusive
29	TRUE	0.20	0.04	0.41	6.89	0.81	0.57	1.00	inconclusive
30	TRUE	0.01	0.00	11.12	4.20	24.07	0.97	0.45	inconclusive
31	TRUE	0.03	0.02	2.29	4.60	3.75	0.92	0.45	inconclusive
32	TRUE	0.05	0.02	1.39	6.36	2.42	0.86	0.45	inconclusive
33	TRUE	0.26	0.05	0.38	6.20	0.71	0.53	1.00	inconclusive
34	TRUE	0.41	0.07	0.25	5.17	0.59	0.36	1.00	p.0 with 0.808
35	TRUE	0.12	0.03	0.73	6.41	1.19	0.74	1.00	inconclusive
36	TRUE	0.03	0.02	2.24	4.88	3.75	0.91	0.45	inconclusive

Table C.3: Overview over criteria to define replication success: sunk cost hypothesis

Appendix D

Credibility intervals

We include all intervals - posterior equal-tailed credibility interval, HDI in isolation and HDI in a hierarchical model - that have been constructed for each hypothesis and scenario. Additionally, we plot the distribution of the two overarching parameters, the grand mean θ and the heterogeneity τ^2, which characterize the normal distribution from which the individual replication means are drawn.

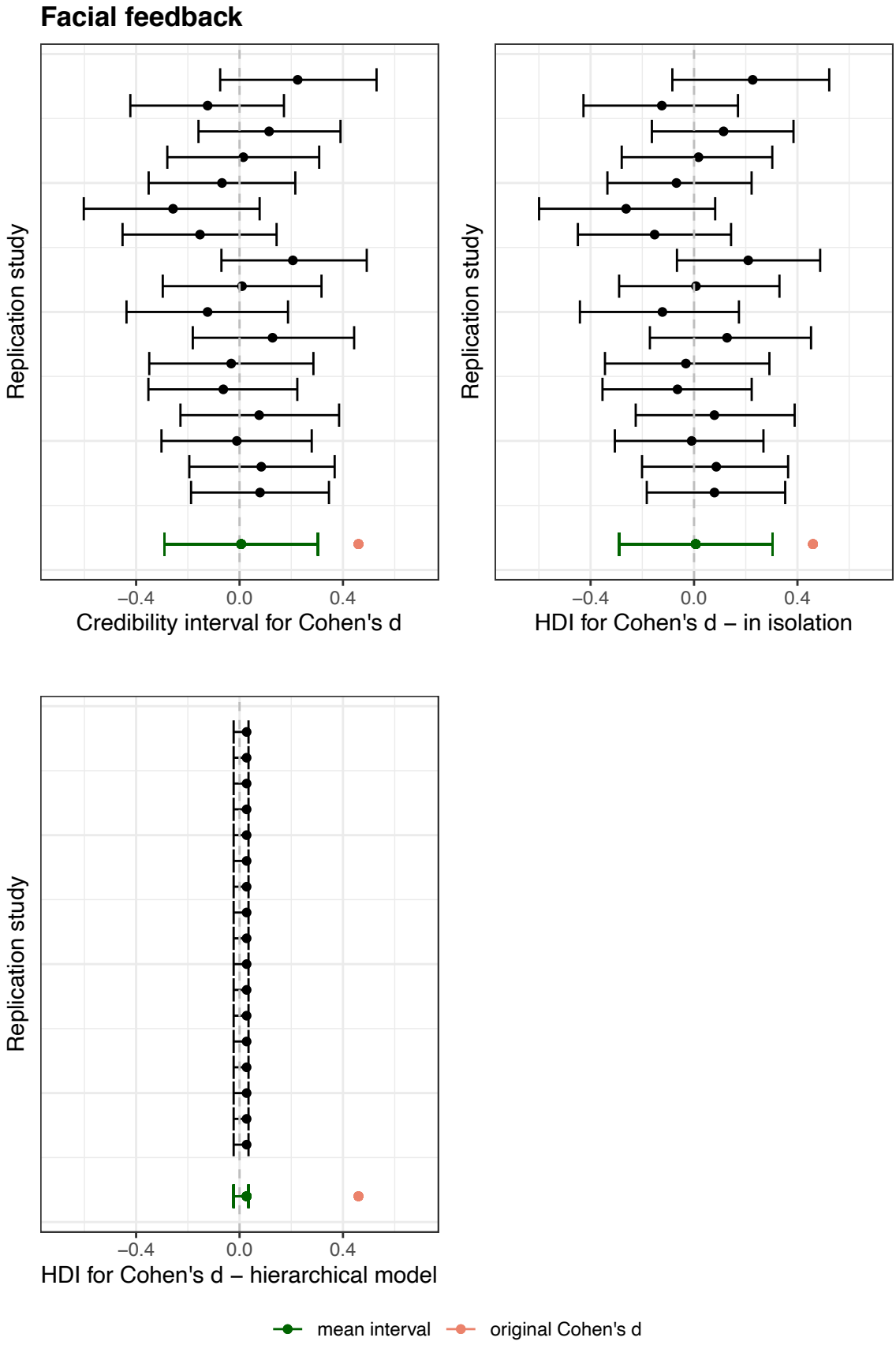

Figure D.1: Credibility interval for facial feedback hypothesis

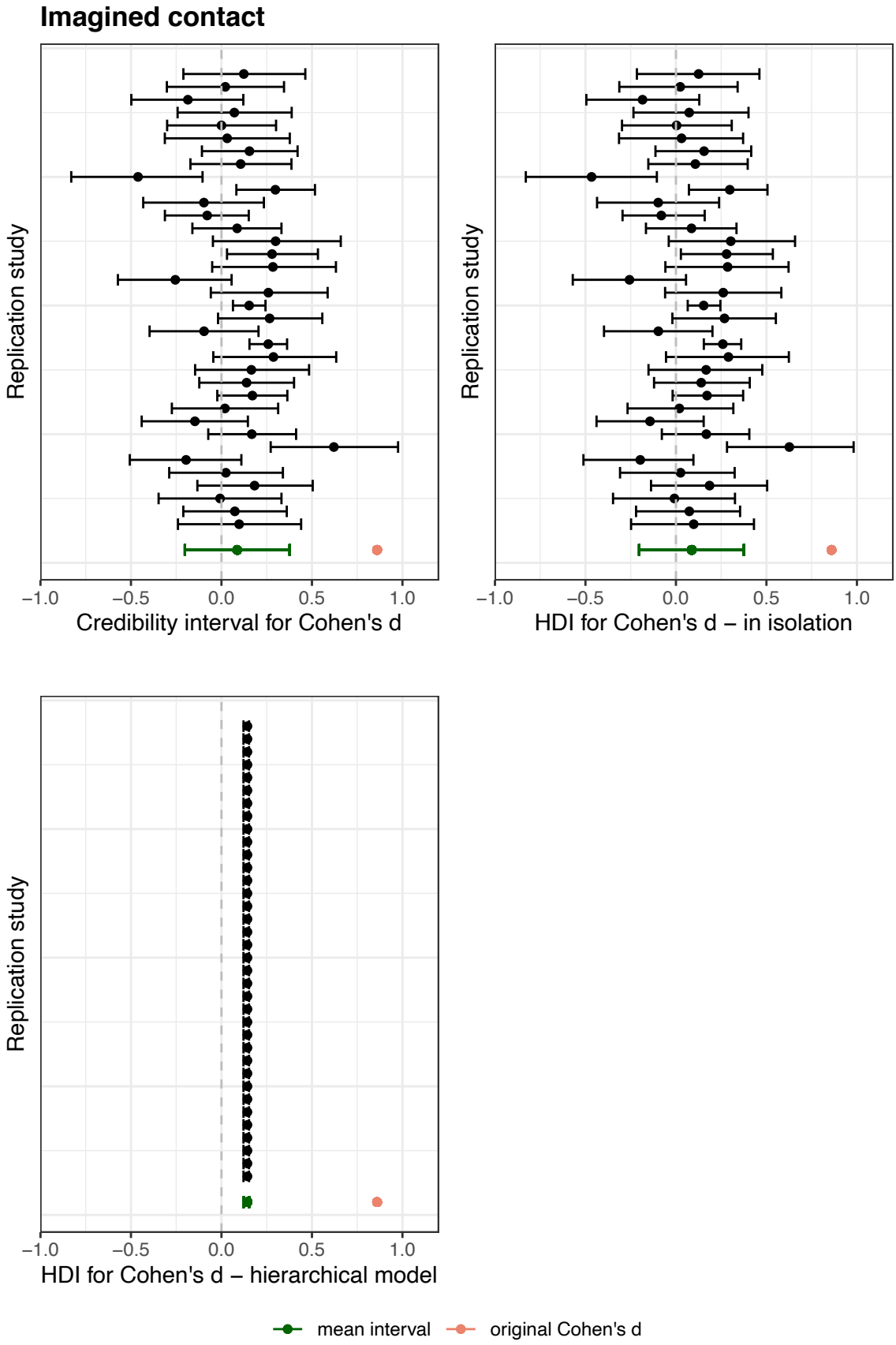

Figure D.2: Credibility interval for imagined contact hypothesis

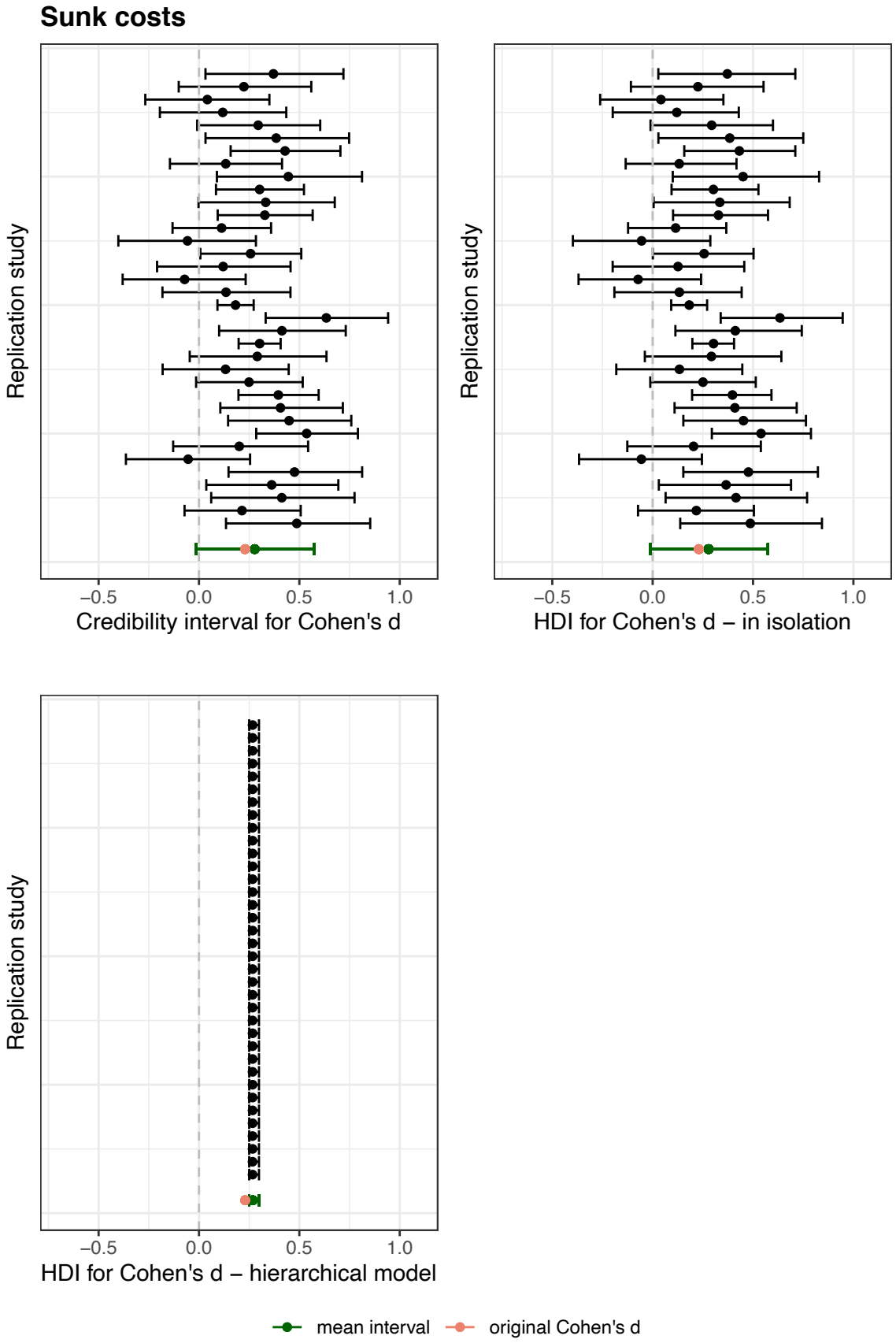

Figure D.3: Credibility interval for sunk costs hypothesis

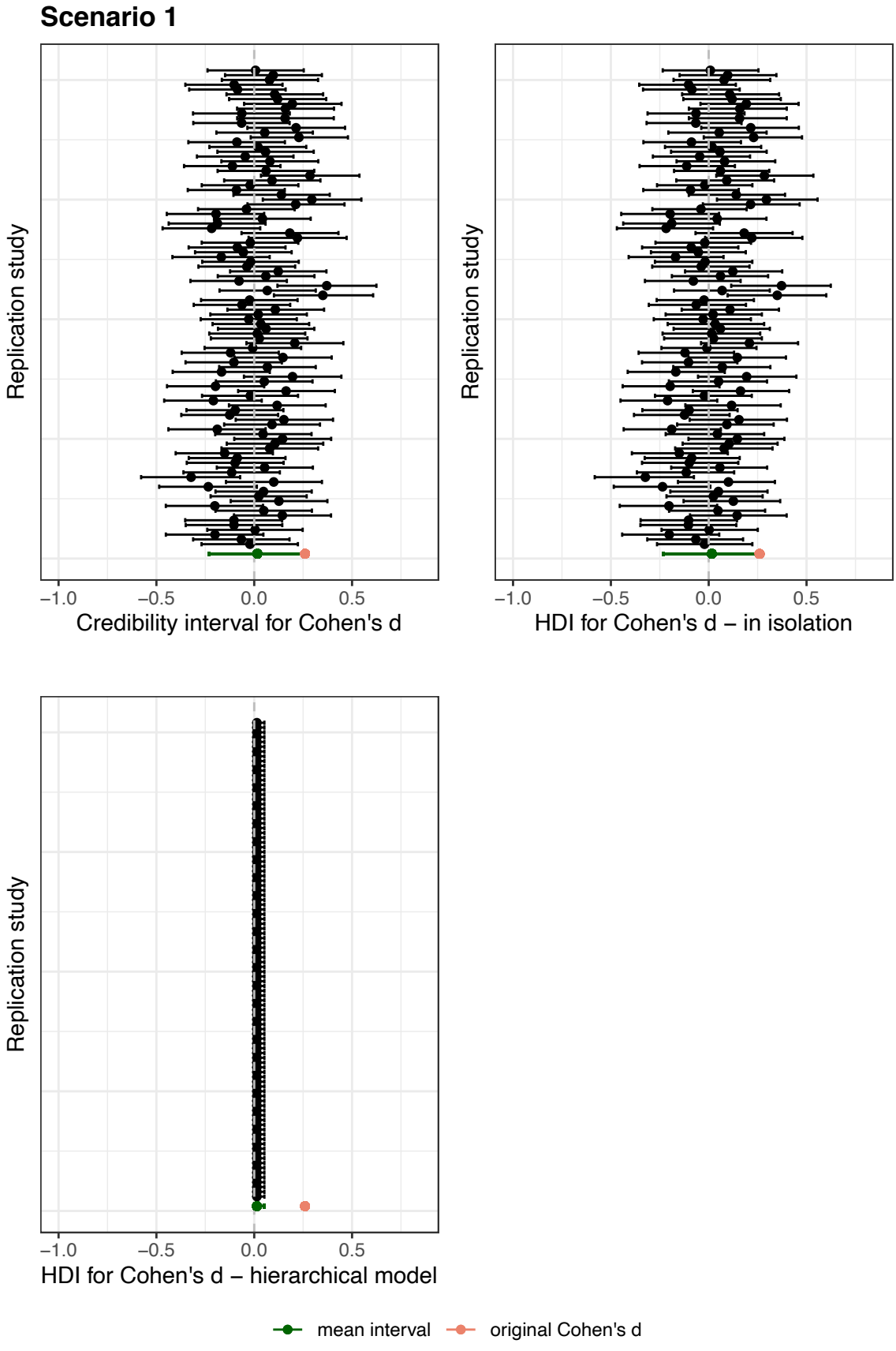

Figure D.4: Credibility interval for simulation data in scenario 1

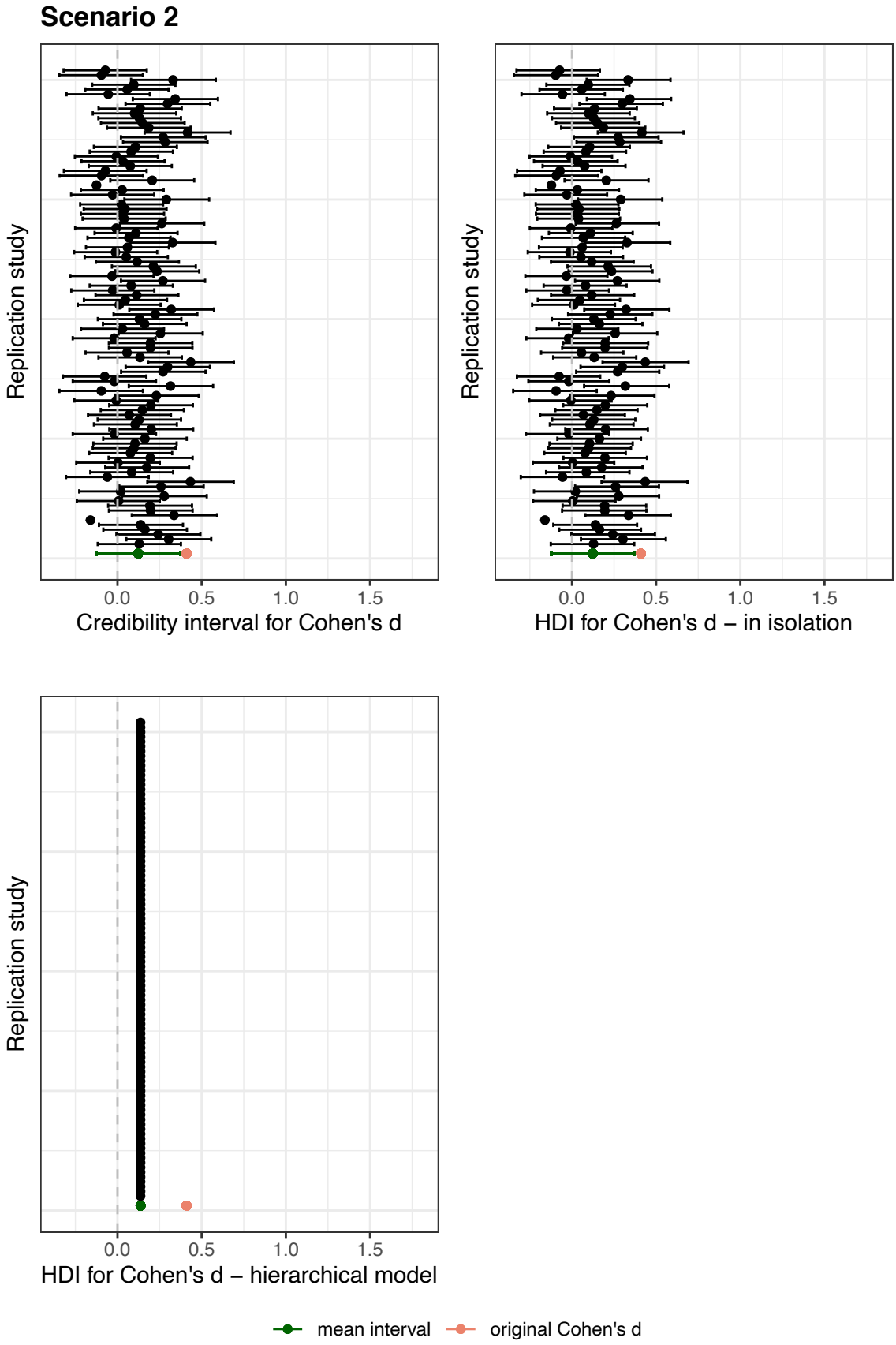

Figure D.5: Credibility interval for simulation data in scenario 2

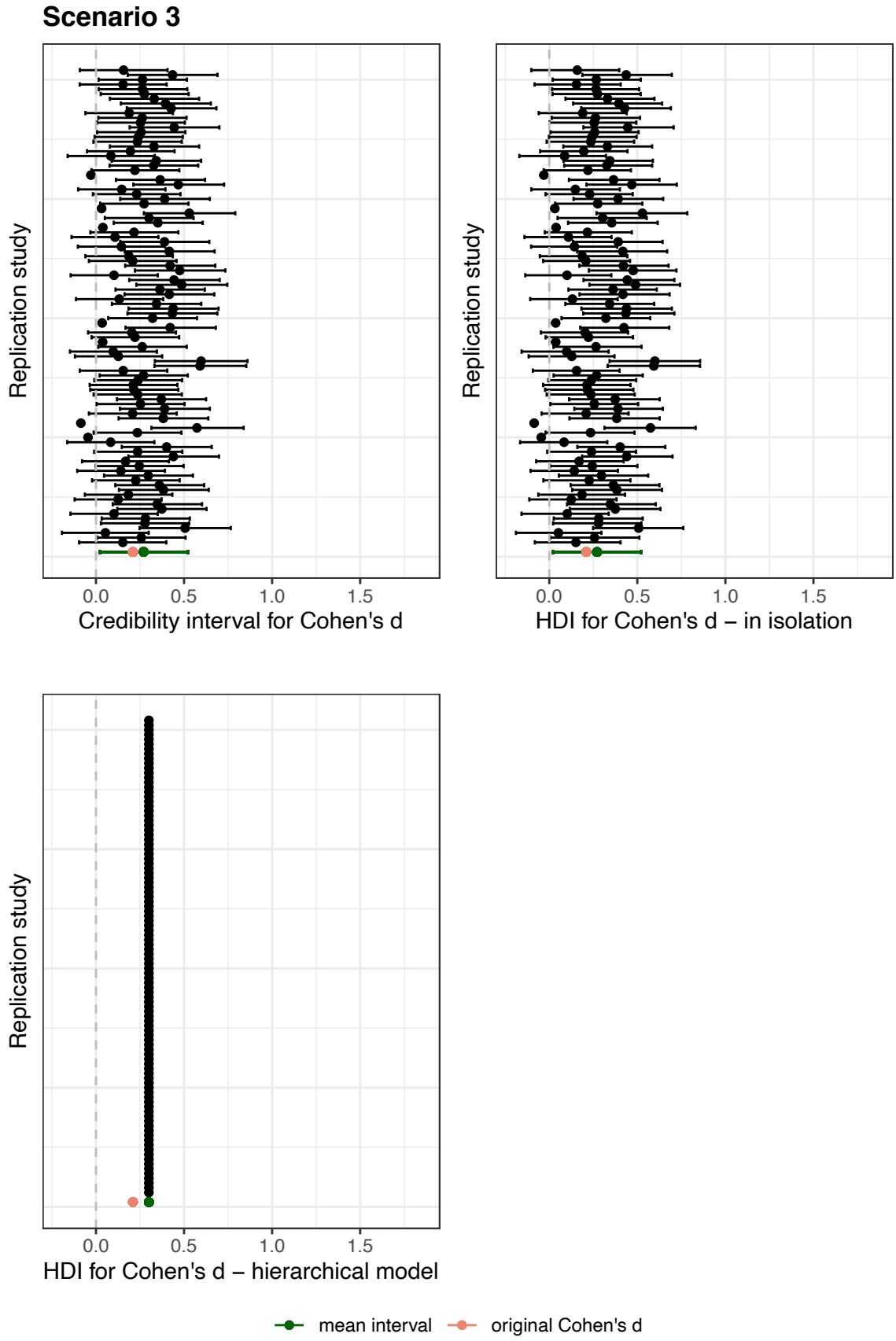

Figure D.6: Credibility interval for simulation data in scenario 3

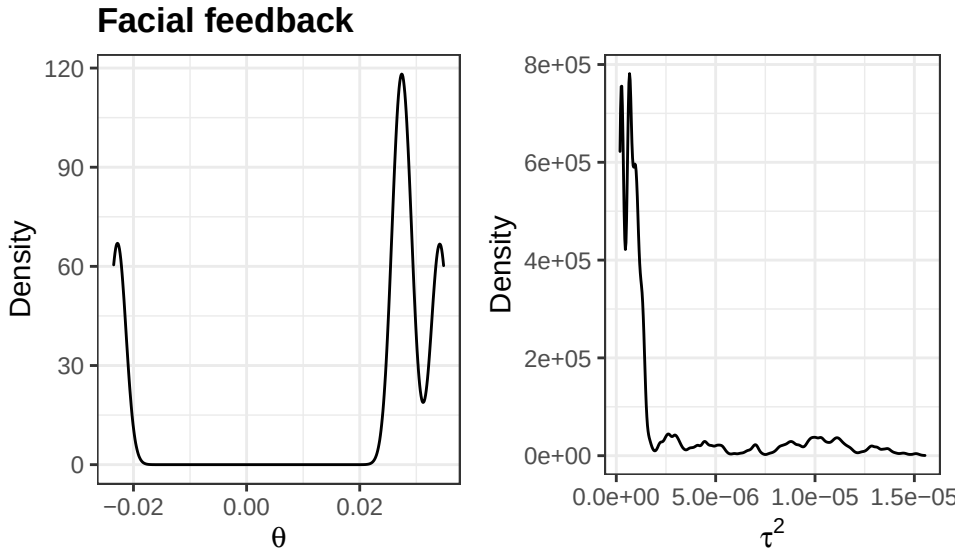

Figure D.7: Density of θ and τ^2 for facial feedback hypothesis

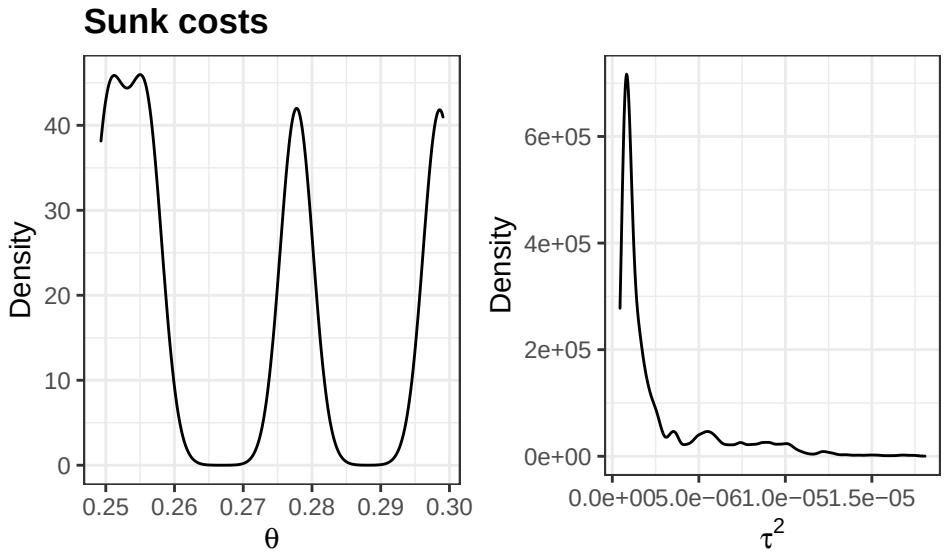

Figure D.8: Density of θ and τ^2 for sunk costs hypothesis

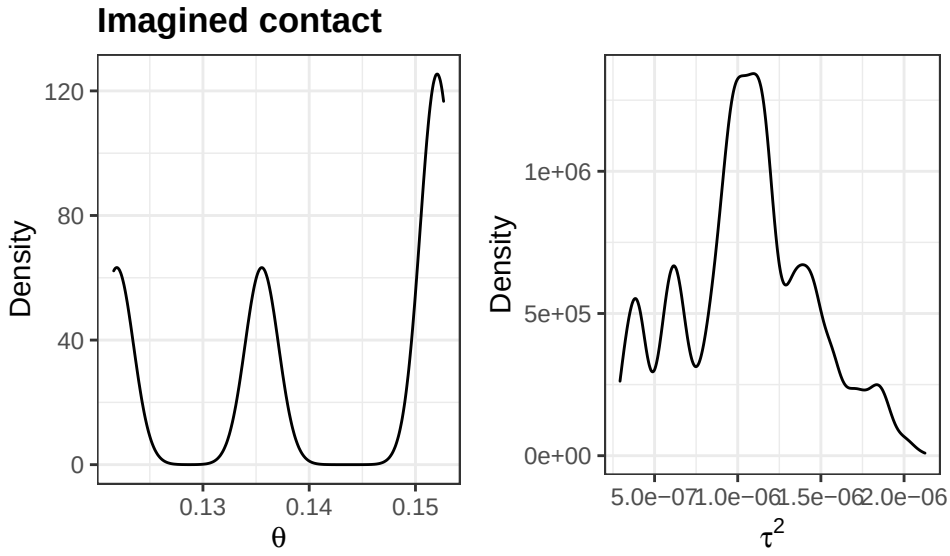

Figure D.9: Density of θ and τ^2 for imagined contact hypothesis

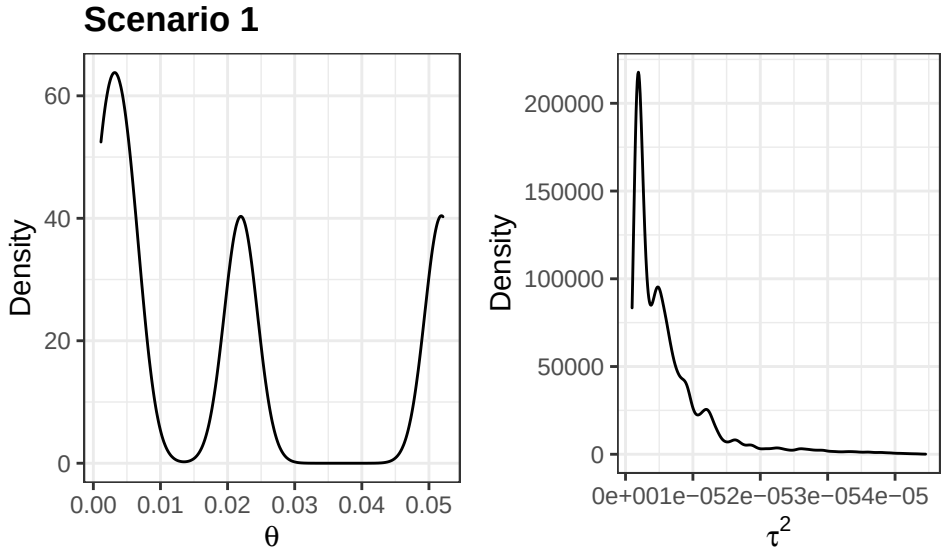

Figure D.10: Density of θ and τ^2 for scenario 1

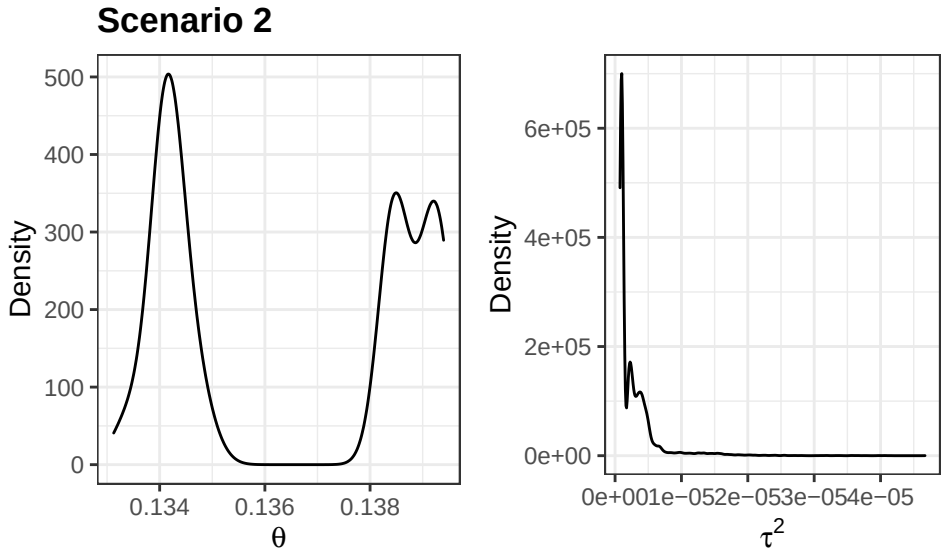

Figure D.11: Density of θ and τ^2 for scenario 2

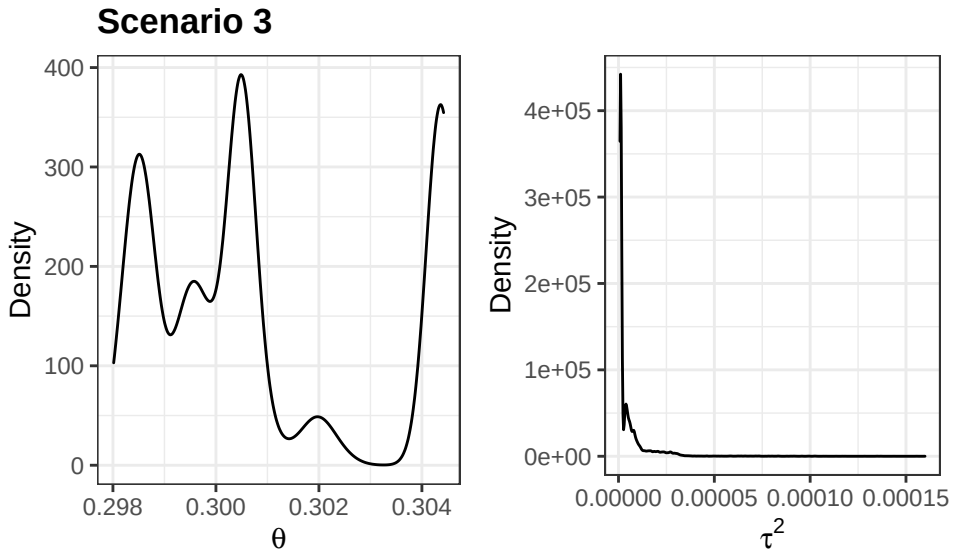

Figure D.12: Density of θ and τ^2 for scenario 3

Bibliography

Amrhein, V., Trafimow, D., & Greenland, S. (2018). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *PeerJ Preprints*, *6*, e26857v4.

Anderson, S., & Maxwell, S. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1.

Baumeister, R. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, *66*, 153–158.

Bayarri, M., & Mayoral, A. (2002). Bayesian analysis and design for comparison of effect-sizes. *Journal of Statistical Planning and Inference*, *103*(1-2), 225–243.

Bayarri, M., & Garcia-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, *94*(1), 135–152.

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815.

Berger, J., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 317–335.

Białek, M. (2018). Replications can cause distorted belief in scientific progress. *Behavioral and Brain Sciences*, *41*.

Bolstad, W., & Curran, J. (2016). *Introduction to Bayesian statistics*. John Wiley & Sons.

Bonett, D. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods*, *14*(3), 225.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Brandt, M., IJzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R., Grange, J., Perugini, M., Spies, J., & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224.

Center for Open Science. (2021). *Registered Reports: Peer review before results are known to align scientific values and practices.* Retrieved April 25, 2021, from https://www.cos.io/initiatives/registered-reports

Champely, S. (2020). *Pwr: Basic Functions for Power Analysis* [R package version 1.3-0]. https://CRAN.R-project.org/package=pwr

Chan, A.-W., Hrobjartsson, A., Haahr, M., Gotzsche, P., & Altman, D. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Jama*, *291*(20), 2457–2465.

Chawla, D. S. (2016). *How many replication studies are enough?* Retrieved May 11, 2021, from https://www.nature.com/news/how-many-replication-studies-are-enough-1.19461

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Del Re, A. (2013). *Compute.es: Compute effect sizes*. https://cran.r-project.org/package=compute.es

Diener, M. (2010). Cohen's d. *The Corsini Encyclopedia of Psychology*, 1–1.

Donnelly, S., & Kramer, A. (1999). Testing for multiple species in fossil samples: An evaluation and comparison of tests for equal relative variation. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, *108*(4), 507–529.

Dowle, M., & Srinivasan, A. (2019). *Data.table: Extension of 'data.frame'* [R package version 1.12.8]. https://CRAN.R-project.org/package=data.table

Dunlap, K. (1926). The experimental methods of psychology. *Powell Lecture in Psychological Theory, Apr, 1925, Clark University, Worcester, MA, US; Portions of this research were presented at the Powell Lecture in Psychological Theory at Clark University, April 21, 1925.*

Eden, D. (2002). From the editors: Replication, meta-analysis, scientific progress, and AMJ's publication policy. *Academy of Management Journal*, 841–846.

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2016). *Statistik: Der Weg zur Datenanalyse*. Springer-Verlag.

Ferguson, C., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561.

Garin, O. (2014). Ceiling effect. In A. Michalos (Ed.), *Encyclopedia of quality of life and well-being research*. Springer Netherlands Dordrecht.

Gelman, A. (2015a). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*, *41*(2), 632–643.

Gelman, A. (2015b). *The feather, the bathroom scale and the kangaroo*. Retrieved May 12, 2021, from https://statmodeling.stat.columbia.edu/2015/04/21/feather-bathroom-scale-kangaroo/

Gelman, A. (2016). The problems with p-values are not just with p-values. *The American Statistician*, *70*(10).

Gelman, A. (2018). Don't recognize replications as successes or failures. *Behavioral and Brain Sciences*, *41*.

Gelman, A., & Carlin, J. (2017). Some natural solutions to the p-value communication problem—and why they won't work. *Journal of the American Statistical Association*, *112*(519), 899–901.

German Research Foundation. (2017). *Replicability of Research Results*. Retrieved April 18, 2021, from https://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2017/170425_stellungnahme_replizierbarkeit_forschungsergebnisse_en.pdf

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International journal of endocrinology and metabolism*, *10*(2), 486.

Gollwitzer, M., & Schwabe, J. (2020). Context dependency as a Predictor of Replicability.

Goodman, S. (1999). Toward evidence-based medical statistics: The P value fallacy. *Annals of internal medicine*, *130*(12), 995–1004.

Goodman, S., Fanelli, D., & Ioannidis, J. (2016). What does research reproducibility mean? *Science translational medicine*, *8*(341), 341ps12–341ps12.

Goodman, S., & Greenland, S. (2007). Assessing the unreliability of the medical literature: A response to "Why most published research findings are false".

Griffin, J. M., Fuhrer, R., Stansfeld, S., & Marmot, M. (2002). The importance of low control at work and home on depression and anxiety: Do these effects vary by gender and social class? *Social science & medicine*, *54*(5), 783–798.

Gundersen, E. (2021). The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A*, *379*(2197), 20200210.

Held, L. (2019). The assessment of intrinsic credibility and a new argument for p< 0.005. *Royal Society open science*, *6*(3), 181534.

Held, L. (2020). A new standard for the analysis and design of replication studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(2), 431–448.

Held, L., Micheloud, C., & Pawel, S. (2020). The assessment of replication success based on relative effect size. *arXiv preprint arXiv:2009.07782*.

Held, L., Micheloud, C., & Pawel, S. (n.d.). *ReplicationSuccess: Design and Analysis of Replication Studies with ReplicationSuccess* [R package version 0.1.4]. https://rdrr.io/rforge/ReplicationSuccess/f/inst/doc/ReplicationSuccess.pdf

Hoaglin, D. (2016). Misunderstandings about Q and 'cochran's q test' in meta-analysis. *Statistics in medicine*, *35*(4), 485–495.

Hoffmann, S. (2017). *A hierarchical Bayesian approach to account for measurement error and dose uncertainty in the association between exposure to ionizing radiation and lung cancer mortality in the French cohort of uranium miners* (Doctoral dissertation). Université Paris-Saclay.

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A.-L. (2020). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines.

Hoijtink, H., van Kooten, P., & Hulsker, K. (2016). Why bayesian psychologists should change the way they use the Bayes factor. *Multivariate Behavioral Research*, *51*(1), 2–10.

Hunter, J. (2001). The desperate need for replications. *Journal of Consumer Research*, *28*(1), 149–158.

Husnu, S., & Crisp, R. (2010). Elaboration enhances the imagined contact effect. *Journal of Experimental Social Psychology*, *46*(6), 943–950.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Ioannidis, J. (2007). Why most published research findings are false: Author's reply to Goodman and Greenland. *PLoS Med*, *4*(6), e215.

Ioannidis, J. (2014). Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false. *Biostatistics*, *15*(1), 28–36.

Ioannidis, J. (2018). Why replication has more scientific value than original discovery. *Behavioral and Brain Sciences*, *41*.

Jager, L., & Leek, J. (2014a). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, *15*(1), 1–12.

Jager, L., & Leek, J. (2014b). Rejoinder: An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, *15*(1), 39–45.

Jeffreys, H. (1961). *The theory of probability*. Oxford University Press.

Jones, M., Coffee, P., Sheffield, D., Yangüez, M., & Barker, J. (2012). Just a game? Changes in English and Spanish soccer fans' emotions in the 2010 World Cup. *Psychology of Sport and Exercise*, *13*(2), 162–169.

Kauermann, G., & Hothorn, T. (2020). Lecture notes on Statistik IV: Wahrscheinlichkeitstheorie und Inferenz II.

Kelley, K. et al. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*(8), 1–24.

Kelley, K. (2020). *MBESS: The MBESS R Package* [R package version 4.8.0]. https://CRAN.R-project.org/package=MBESS

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

Kerr, N. (1998). Harking: Hypothesizing after the results are known. *Personality and social psychology review*, *2*(3), 196–217.

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

Lakatos, I. (1970). History of science and its rational reconstructions. *PSA: Proceedings of the biennial meeting of the philosophy of science association*, *1970*, 91–136.

Laraway, S., Snycerski, S., Pradhan, S., & Huitema, B. (2019). An overview of scientific reproducibility: Consideration of relevant issues for behavior science/analysis. *Perspectives on Behavior Science*, *42*(1), 33–57.

Liang, F., Paulo, R., Molina, G., Clyde, M., & Berger, J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423.

Lindley, D. (1957). A statistical paradox. *Biometrika*, *44*(1/2), 187–192.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2019). Replication Bayes factors from evidence updating. *Behavior research methods*, *51*(6), 2498–2508.

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). Bayestestr: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. https://joss.theoj.org/papers/10.21105/joss.01541

Marsman, M., Schoenbrodt, F., Morey, R., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A bayesian bird's eye view of 'Replications of important results in social psychology'. *Royal Society open science*, *4*(1), 160426.

Matthews, R. (2018). Beyond 'significance': Principles and practice of the analysis of credibility. *Royal Society Open Science*, *5*(1), 171047.

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

Moreno, E. (2005). Objective bayesian methods for one-sided testing. *Test*, *14*(1), 181–198.

Morey, R., & Rouder, J. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological methods*, *16*(4), 406.

Morey, R., & Rouder, J. (2018). *Bayesfactor: Computation of Bayes Factors for Common Designs* [R package version 0.9.12-4.2]. https://CRAN.R-project.org/package=BayesFactor

Morey, R., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics & Probability Letters*, *92*, 121–124.

Morris, T., White, I., & Crowther, M. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074–2102.

Musgrave, A., & Pigden, C. (2021). Imre Lakatos. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2021). Metaphysics Research Lab, Stanford University.

Nelson, L., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual review of psychology*, *69*, 511–534.

Nosek, B., & Errington, T. (2017). Reproducibility in cancer biology: Making sense of replications. *Elife*, *6*, e23383.

Nosek, B., & Lakens, D. (2014). Registered reports.

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

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, *45*(4), 867–872.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on psychological science*, *7*(6), 528–530.

Plesser, H. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in neuroinformatics*, *11*, 76.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Razali, N., Wah, Y. et al. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of statistical modeling and analytics*, *2*(1), 21–33.

Rosenkranz, G. (2021). Replicability of studies following a dual-criterion design. *Statistics in Medicine*.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638.

Rotello, C., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*(4), 944–954.

Rouder, J., & Morey, R. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*(4), 682–689.

Rouder, J., Speckman, P., Sun, D., Morey, R., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225–237.

Savage, L. e. a. (1961). The foundations of statistics reconsidered. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*.

Scheibehenne, B., Jamil, T., & Wagenmakers, E.-J. (2016). Bayesian evidence synthesis can reconcile seemingly inconsistent results: The case of hotel towel reuse [including Online material]. *Psychological Science*, *27*(7), 1043–1046.

Schmidt, F., & Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology*, *4*(1), 32.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L. L., Diermeier, D., Heinze, J., et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67.

Sidik, K., & Jonkman, J. (2019). A note on the empirical Bayes heterogeneity variance estimator in meta-analysis. *Statistics in medicine*, *38*(20), 3804–3816.

Signorell, A. e. m. a. (2021). *DescTools: Tools for Descriptive Statistics* [R package version 0.99.41]. https://cran.r-project.org/package=DescTools

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366.

Simons, D. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological science*, *26*(5), 559–569.

Smithson, M. (2003). Noncentral confidence intervals for standardized effect sizes. *Confidence intervals*, 33–41.

Snow, G. (2020). *TeachingDemos: Demonstrations for Teaching and Learning* [R package version 2.12]. https://CRAN.R-project.org/package=TeachingDemos

Spellman, B. (2015). A short (personal) future history of revolution 2.0.

Stan Development Team. (2020). RStan: The R interface to Stan [R package version 2.21.2]. http://mc-stan.org/

Strack, F. (2016). Reflection on the smiling registered replication report.

Strack, F., Martin, L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of personality and social psychology*, *54*(5), 768.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71.

Turner, R., Crisp, R., & Lambert, E. (2007). Imagining intergroup contact can improve intergroup attitudes. *Group Processes & Intergroup Relations*, *10*(4), 427–441.

van Aert, R. (2021a). *Web application snapshot method*. Retrieved June 17, 2021, from https://rvanaert.shinyapps.io/snapshot/

van Aert, R. (2021b). *Puniform: Meta-Analysis methods correcting for Publication Bias* [R package version 0.2.4]. https://CRAN.R-project.org/package=puniform

Van Aert, R., & Van Assen, M. (2017). Bayesian evaluation of effect size after replicating an original study. *PloS one*, *12*(4), e0175302.

Van Aert, R., & Van Assen, M. (2018). Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication. *Behavior Research Methods*, *50*(4), 1515–1539.

van Ravenzwaaij, D., & Etz, A. (2021). Simulation studies as a Tool to understand Bayes Factors. *Advances in Methods and Practices in Psychological Science*, *4*(1).

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417.

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://www.jstatsoft.org/v36/i03/

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q., Acosta, A., Adams Jr, R., Albohn, D., Allard, E., Benning, S., Blouin-Hudon, E.-M., et al. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Wagenmakers, E.-J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior research methods*, *48*(2), 413–426.

Welkowitz, J., Cohen, B., & Ewen, R. (2006). *Introductory statistics for the behavioral sciences*. John Wiley & Sons.

Wetzels, R., Raaijmakers, J., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic bulletin & review*, *16*(4), 752–760.

Wingen, T., Berkessel, J., & Englich, B. (2020). No replication, no trust? How low replicability influences trust in psychology. *Social Psychological and Personality Science*, *11*(4), 454–463.

Zwaan, R., Etz, A., Lucas, R., & Donnellan, B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*.