The Replication Crisis in Psychology: Statistical and Meta-Scientific Perspectives

- Kumulative Arbeit -

Inaugural-Dissertation
zur Erlangung der Doktorwürde
der
Philosophischen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität
zu Bonn

vorgelegt von
Christopher Harms
aus Goch

Bonn, 2021

Gedruckt mit der Genehmigung der Philosophischen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Zusammsetzung der Prüfungskommission:
Prof. Dr. Martin Reuter, Institut für Psychology, Universität Bonn
(Vorsitzender)

Prof. Dr. André Beauducel, Institut für Psychologie, Universität Bonn
(Betreuer und Gutachter)

Prof. Dr. Daniël Lakens, Human Technology Interaction Group, Eindhoven University of Technology
(Gutachter)

Prof. Dr. Ulrich Ettinger, Institut für Psychologie, Universität Bonn
(weiteres prüfungsberechtigtes Mitglied)

Tag der mündlichen Prüfung: 13. September 2021

Contents

Acknowledgement

> [A] PhD thesis [...] is part of a bigger entity. No PhD builds his own cyclotron as part of his thesis. No PhD orbits his own satellite to get his data.
>
> — Tukey (1969, p. 88)

A thesis is always the culmination of many people's work and influence. Therefore, there are many people I have to thank for their contribution to this work.

First of all, I thank my supervisor Prof. André Beauducel, who motivated me to start working on this interesting and relevant topic, and who provided me with the freedom to explore my own ideas. This includes allowing me to spend two months of my time in Eindhoven at the TU/e, where I was able to work with Dr Daniël Lakens, Anne Scheel, and Peder Isager, who have contributed to this work tremendously through the inspiration they are, the thoughtful discussions we had, and the chance to experience Dutch student culture. This work would have been very different without their support.

With chats, lunch and coffee breaks, as well as discussions on psychology, methods, and university politics, my colleagues at the Department of Psychology helped to make sense of times in which the PhD studies seemed to be neverending. This is a thank you to Laura, Nicolas, Anneke, and Vera.

My colleagues at SKOPOS, most notably Basti, supported this PhD by giving me the chance to focus on my research at the university as well as allowing me to spend the two months in Eindhoven. Without their openness and understanding, it would not have been possible to do both at the same time. Thank you, Elli, for proof-reading the final draft and having patience with my punctuation.

Having the chance to study what I was interested in would not have been possible without the upbringing, open-minded education, and the support by my parents, Hiltrud and Peter. They instilled in me the curiosity, the desire to learn, and the eagerness to pursue my goals, that made my academic endeavours possible. Last, but not least, this thesis would not have been possible without the support, understanding, and love by my wife, Nicola. Thank you for everything!

Zusammenfassung

Replikationen sind, in der vorherrschenden Sichtweise, ein wesentlicher Bestandteil von empirischen Wissenschaften. Das *Reproducibility Project: Psychology* hat jedoch gezeigt, dass systematische Replikationen in der psychologischen Forschung nur selten durchgeführt werden und viele Replikationsversuche vorherige Ergebnisse nicht replizieren können. Diese Veröffentlichung war ein wesentlicher Treiber für Diskussionen rund um die "Replizierbarkeitskrise" in der Psychologie. Seit mindestens 2012 beschäftigt die Frage nach der Replizierbarkeit, und damit der Zuverlässigkeit, von Studien die Psychologie. Im Rahmen dieser Krise wurden verschiedene Verursache unternommen, die Gründe zu identifizieren und die Zuverlässigkeit psychologischer Forschung zu verbessern. Im Rahmen dieser Dissertation werden die Ereignisse und Diskussionspunkte der letzten Jahre zusammengefasst und mit besonderem Blick auf wissenschaftstheoretische und statistische Aspekte beleuchtet. Bei der wissenschaftstheoretischen Perspektive ist insbesondere die Verbindung von Theorie und statistischen Hypothesen relevant, um die Auswertungen von empirischen Daten zu leiten. Bei der Auswertung von Studien werden in der Psychologie traditionell Null-Hypothesen Signifikanztests (NHST) verwendet. Dem gegenüber wird in dieser Arbeit die Bayes-Statistik erläutert und Bayes'sche Parameterschätzung und Bayes Faktoren mit frequentistischen Signifikanztests, insbesondere bei sogenannten Null-Effekten, verglichen. Mittels Bayes Faktoren lassen sich im Kontext der Bayes-Statistik Hypothesen prüfen, was insbesondere für den Vergleich von Original- und Replikationsstudien interessant ist: Die "Replication Bayes factors" werden dafür vorgestellt und für Mehrgruppen-Vergleiche hergeleitet. Anhand einer konkreten Replikationsstudie und der Literatur werden abschließend Empfehlungen für die Durchführung von Replikationsstudien gegeben. Hierbei sind insbesondere die gleichen Maßstäbe anzulegen wie bei Originalarbeiten. Die Arbeit endet mit einem Ausblick auf die weitere Entwicklung der Psychologie als wissenschaftliche Disziplin vor dem Hintergrund der Replizierbarkeitskrise.

Abstract

Replications are, in the predominant view, an essential part of empirical science. However, the *Reproducibility Project: Psychology* has shown that systematic replications are rarely carried out in psychological science and many replication attempts fail to replicate prior results. This publication was a major driver for discussions around the "replicability crisis" in psychology. Since at least 2012, the question of replicability, and thus reliability, of studies has preoccupied psychology. In the context of this crisis, various attempts have been made to identify the causes and to improve the reliability of psychological research. In this thesis, the events and points of debate of the last few years are summarised and examined with a special focus on meta-scientific and statistical aspects. In the meta-scientific perspective, the link between theory and statistical hypotheses is particularly relevant in guiding the analysis of empirical data. In the statistical analysis of psychological studies, null-hypothesis significance tests (NHST) are traditionally used in psychology. In contrast, Bayes statistics is explained in this paper and Bayesian parameter estimation and Bayes factors are compared with frequentist significance tests, especially for the case of so-called null effects. Bayes factors can be used to test statistical hypotheses in the context of Bayesian statistics, which is particularly interesting for the comparison of original and replication studies: The "Replication Bayes factors" are introduced for this purpose and derived for multi-group comparisons. Finally, recommendations for conducting replication studies are provided based on a specific replication study and the relevant literature. In particular, the same standards are to be applied as for original studies. The work concludes with an outlook on the further development of psychology as a scientific discipline against the background of the replicability crisis.

Chapter 1

Introduction

> We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence,' but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable.
>
> — Popper (2002, p. 23)

Many fields in science aim to observe consistent patterns in the world, to gain an understanding of the patterns and underlying mechanisms, and to provide a description of the world and its nature. These consistent patterns, from which rules can be deduced, might lead to theories which, in turn, can be tested. Showing that a rule can be applied independently of a single researcher or a single study is thus a cornerstone of the scientific method. Consequently, experiments, in particular, ought to be repeatable, show similar patterns under similar circumstances, and therefore lead to the same conclusions (Popper, 2002, pp. 23–24):

> Indeed the scientifically significant *physical effect* may be defined as that which can be regularly reproduced by anyone who carries out the appropriate experiment in the way prescribed.

If, on the other hand, we fail to repeatedly show the same patterns in different observations of the world, we might become increasingly sceptical about the natural law, "physical effect," or pattern in question. There is little disagreement about the notion that replication is thus, in general, a crucial part of scientific progress and the scientific method (see also Tukey, 1969). While in some fields, independent replication is the norm (esp. in the natural sciences), some fields do not value replication as much as others. Psychology, in particular, lacks replication research as an integral part of scientific development and does not have many replication studies in the published literature (Gigerenzer, 2018; Theodore D. Sterling, 1959; Tsang & Kwan, 1999). While this criticism has already been voiced for several decades, over the last decade, there has been increased awareness of this issue. Several failures to replicate previous findings and systematic replication projects have contributed to an ongoing, broad debate across the field. It has resulted in changes to research practices, publication policies, statistical practice, and methodology, and led to a rediscovery of previous researchers' writings on issues of replication, theory building, and appraisal, and statistical methods. The concerns about the difficult or impossible replication of several highly cited and important psychological effects have also led to several descriptive terms of the period such as "crisis of confidence" (Pashler & Wagenmakers, 2012), "replicability crisis" (Pashler & Harris, 2012), or "credibility revolution" (Vazire, 2018). Close to these terms, I will refer to the debate ongoing since 2010 as "replication crisis" throughout this thesis.

As both the replication crisis itself as well as research on the crisis are not limited to the field of psychology, the study of replication is more akin to what Faust & Meehl (2002) termed "meta-science": In its research questions and methods it is "research on research." For the present thesis, a focus is placed on psychology as a discipline, as it has been the field that has had the biggest challenge since 2012 to investigate itself. In psychology, the replication crisis has been covered broadly in many different journals, sub-fields, on conferences, increasingly on Twitter and personal blogs by researchers, and from a variety of viewpoints including philosophy of science, social psychology (by investigating incentives and biases among researchers), and statistics. Among commonly discussed aspects are the questions whether there is a "crisis," what the reasons for unsuccessful replications are, and how to improve psychology as a science in general. All of these are crucially important for the field but, at the same time, expansive and without a simple answer. The resulting meta-scientific research program can easily fill multiple dissertations and an increasing number of research groups focus on these questions. While a large amount of the debate takes place within the field of psychology, researching the replication crisis involves questions of philosophy, history, or statistics and involves scientific collaborations with those and other fields.

One theme has emerged repeatedly in the discussions about replicability already before the recent discussions in the 2010s: The way psychologists analyse data from experiments, field studies, or observational studies often poses statistical and epistemological problems not properly addressed in practice. It thus deserves careful investigation and critical examination. This thesis will give an overview on the replication crisis from different angles and provide a context for a more in-depth discussion of statistical practices in psychological research and the context of replications: In chapter 1, the recent history of the "replication crisis" will be outlined (section 1.1) and considered in the context of similar debates in other fields and previous years. It will also review the perspective of the philosophy of science (section 1.2). Chapter 2 will discuss two competing statistical frameworks empirical researchers can use to analyse data. Problems with the dominant framework, null hypothesis significance testing, are discussed. In particular, the analysis of "null effects" is discussed along Harms & Lakens (2018), and a statistical tool for the analysis of replication studies, the Replication Bayes factor, is introduced along Harms (2018). Chapter 3 will then focus more extensively on replication studies, give a taxonomy for replications, and outline how replication studies should be performed. As an example, a replication study on the "Rounded Price Effect" (Harms, Genau, Meschede, & Beauducel, 2018) will be discussed. The thesis will end with conclusions, summarising the discussion outlined therein, and by giving an outlook on changes to psychological science in general and replicability in particular in chapter 4.

This cumulative dissertation is based on three published journal articles which are woven into the fabric of the text and summarised in more detail in their respective subsections. Harms & Lakens (2018) (article I) exemplify different ways to statistically investigate null effects, i.e. hypotheses of no or negligible effect, using three different statistical frameworks. The first journal article will be considered in section 2.4 and put into the context of evaluating replication attempts. Harms (2018) (article II) introduces a particular way of using Bayes factors to evaluate the outcome of a replication study in relation to the original study. This method will be introduced in section 2.5.1 as one special case of Bayes factors and a particular method for the evaluation of replications. Lastly, Harms et al. (2018) (article III) is a close replication study which is used as an example in chapter 3. This chapter will also discuss the limitations and changes in requirements for replication studies are also discussed. All three articles are included as full-text in the appendix of the printed version.

1.1 The "Replication Crisis" in Psychology

Tracing back the origins of the term, a search on *Web of Science* reveals that Pashler & Harris (2012) have been the first to use the term "replicability crisis" in published literature. In their editorial to a special issue of *Perspectives on Psychological Science* (PPS), Pashler & Wagenmakers (2012) also speak of a "crisis of confidence." A term that was first used by Elms (1975), 37 years earlier, when he referred to increasing "self-doubts about goals, methods, and accomplishments" (Elms, 1975, p. 968) among researchers in social psychology after the field has grown for over two decades. When Pashler & Wagenmakers (2012) use the term, however, they refer to "an unprecedented level of doubt among practitioners about the reliability of research findings in the field" (Pashler & Wagenmakers, 2012, p. 528). This "doubt" is a result of the combination of different findings and studies leading up to the special issue in PPS. Nelson, Simmons, & Simonsohn (2018) have summarised five relevant developments between 2010 and 2012, which have fueled the concerns among psychological researchers:

1. Daryl Bem's paper on extra-sensory perception (ESP) (D. J. Bem, 2011), a paranormal phenomenon which is generally considered impossible. The article was published in the *Journal of Personality and Social Psychology*, one of the most prominent journals in psychology. While the paper and a subsequent meta-analysis (D. Bem, Tressoldi, Rabeyron, & Duggan, 2016) showed significant experimental differences and implied the presence of ESP, it is incompatible with common scientific knowledge about physics and psychology. Most researchers therefore have rejected Bem's results and considered it as a striking and convincing example of how common norms and practices can lead to spurious results. This led many researchers to question common practices and norms, especially statistical practices (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011).

2. Moreover, several cases of scientific fraud (which is – hopefully – not the primary reason for most non-replications) in psychology in the 2010s have also contributed to the magnitude of the "crisis of credibility" in psychology. Nelson et al. (2018) refer to the cases of Smeesters, Stapel and Sanna in particular (Harms, 2015; see also Simonsohn, 2013).

3. The authors' paper on "False-Positive Psychology" (Simmons, Nelson, & Simonsohn, 2011) has quickly become one of the most cited papers of the "credibility revolution."[1] The authors showed that, with some common techniques, they were able to make claims statistically significant. As an example, they illustrated how selecting variables and removing cases can lead to the result that listening to a specific song can make a person younger with a statistically significant result. They termed these techniques "*p*-hacking" as it allowed to generate smaller *p*-values than the data and hypothesis should produce. Around the same time, John, Loewenstein, & Prelec (2012) published a survey of researchers on "questionable research practices," which included several of the "*p*-hacking" techniques that Simmons et al. (2011) have used in their example. This includes reporting only dependent variables and covariate corrections "that worked." It is notable, however, that several of these "questionable research practices" have not only been common (for example, "failing to report all of a study's dependent measures" was admitted by 63.4% of the surveyed researchers, John et al., 2012, table 1, p. 525) but also recommended (see e.g. D. J. Bem, 2002). For some, the terms "*p*-hacking" and "questionable research practices" imply that researchers use them deliberately to misguide their recipients. Others have therefore suggested different terms, such as "garden of forking paths" (Gelman & Loken, 2013) or "researcher degrees of freedom" (Wicherts et al., 2016).

4. One of the first prominent non-replications of a long-standing effect in social psychology was the replication study by Doyen, Klein, Pichon, & Cleeremans (2012). The authors

[1] 3,519 citations on Google Scholar as of 29th May 2019.

tried to replicate the "elderly priming" paradigm published by Bargh, Chen, & Burrows (1996). Their failure to replicate one of the most prominent effects in social psychology cast doubt on the paradigm of "behavioural priming" (at least in the way it was described and explained until then) and started several critical inquiries into theories and paradigms from social psychology.

5. Lastly, the "Reproducibility Project: Psychology" [*RP:P*; Open Science Collaboration (2015)] is certainly one of the key papers in the discussion of replicability in psychology. In an attempt to systematically replicate a set of 100 studies from three top psychology journals, the collaboration found that only 39% of the original studies had been replicated successfully. While there are several concerns about the methodology of the *RP:P* that question the generalizability of the replication rate to the field of psychology as a whole, it showed that publication in a peer-reviewed journal (even with a high impact factor) does not guarantee replicability. While the *RP:P* certainly is one of the most prominent attempts at estimating the proportion of published studies in a field that can be successfully replicated, there have been also other projects for both certain journals (Camerer et al., 2018), selected findings (Ebersole et al., 2016; Richard A. Klein et al., 2014; Richard A. Klein, Vianello, Hasselman, Adams, & Adams, 2018), and specific subfields of psychology.

While Nelson et al. (2018) summarise the developments in psychological science, it is notable that psychology is not the only scientific field experiencing a period of doubt about the credibility of their field, their methods, and how they should move on. Econometrics faced similar issues before (Angrist & Pischke, 2010; Leamer, 1983), and medical researchers issued concerns at the beginning of the 2000s (Ioannidis, 2005, 2008). Concerns about the robustness of findings are not new in psychology either: Statistical power is and remains notoriously low in many studies (Button et al., 2013; Cohen, 1962; Maxwell, 2004; Sedlmeier & Gigerenzer, 1989), the practice of significance testing is regularly criticised (Gigerenzer, 2004, 2018; Meehl, 1967, 1990b), and publication bias remains pervasive (Carter & McCullough, 2014; Kühberger, Fritz, & Scherndl, 2014; Theodore D. Sterling, 1959; T. D. Sterling, Rosenbaum, & Weinkam, 1995) – to name just a few criticisms. While this thesis will focus on statistical practice and inference as one of the causes for non-replicable research, there is an ongoing discussion about other causes as well: Small sample sizes (Asendorpf et al., 2013a; Sedlmeier & Gigerenzer, 1989), selective, non-representative samples [mostly "WEIRD": western, educated, industrialised, rich, democratic participants; Henrich, Heine, & Norenzayan (2010); cf. Montag (2018)], low reliability of measurements (Asendorpf et al., 2013a), lack of theoretical frameworks (Meehl, 1978; Muthukrishna & Henrich, 2019), systematic biases and incentives in the scientific and publishing system against replications and "good" science (Chambers, Dienes, McIntosh, Rotshtein, & Willmes, 2015; Smaldino & McElreath, 2016). Moreover, publication bias, the "file drawer effect," the aforementioned questionable research practices, etc. lead to a systematically biased corpus of published literature in which large effect sizes and statistically significant results are overrepresented (Schäfer & Schwarz, 2019; Simmons et al., 2011; Theodore D. Sterling, 1959; T. D. Sterling et al., 1995).

On the other hand, the view that psychology is in "crisis" or that published failures to replicate previous findings is problematic, has been challenged on different grounds (D. T. Gilbert, King, Pettigrew, & Wilson, 2016a, 2016b; Iso-Ahola, 2017; Wolfgang Stroebe & Strack, 2014). D. T. Gilbert et al. (2016a), for example, questioned the methods and interpretations of the *RP:P*. They argued that many of the replications performed by the Open Science Collaboration (2015) contained "infidelities" presenting stark differences between the original and the replication (e.g. different populations and samples, in which the original effects cannot be expected to replicate) and that, considering only replications of high quality or high power, the replication rate is not as dramatically low as the original *RP:P* publication suggests. Notably, the response to D. T. Gilbert et al. (2016a)'s comment (C. J. Anderson et al., 2016) highlights

misconceptions in the comment and underlines that the *RP:P* only provides the first dataset on this issue and that "both optimistic and pessimistic conclusions about reproducibility are possible, and neither are yet warranted" (C. J. Anderson et al., 2016, pp. 1037–c). Wolfgang Stroebe & Strack (2014) more fundamentally challenged the notion that more replications are needed to corroborate the existing literature. In their article, however, they ignore the different problems outlined before in published literature, such as low power, publication bias, measurement error, and their effect on the inflation of effect sizes.

Projects that have attempted to estimate the "replication rate" across different fields of studies have received several criticisms. One the one hand, each replication might be considered not a faithful replication of the original study in question. As D. T. Gilbert et al. (2016a) correctly state, some of the replications in the *RP:P* are considerably different to the target of the replications – a concern that can only be countered in a single instance by adhering to certain standards (see chapter 3.2). On the other hand, the question remains whether a "replication rate" among a whole field does provide valuable scientific information about the state of a scientific field. The heterogeneity of studies within a single field can be large, with robust and very sensitive effects alike. Projects such as the *Many Labs* studies (Ebersole et al., 2016; Richard A. Klein et al., 2014; Richard A. Klein et al., 2018), which aim to investigate replicability, moderators, and heterogeneity of single effects by replicating the same study multiple times in different labs, provide a different and maybe more informative approach. The *Many Labs* projects, however, have also concluded that some published effects are difficult to consistently replicate and that "hidden moderators" (Wolfgang Stroebe & Strack, 2014) are rarely a sensible explanation (Richard A. Klein et al., 2018).

As science aims to be a self-correcting enterprise, the increased attention to issues of replicability and resulting changes to research practices can help to improve scientific practices. Much of the debate revolves around rigour, transparency, and trustworthiness of psychological research – three goals that are certainly laudable and fundamental to science. This thesis will summarise and discuss several aspects of the "replication crisis." However, there are many more changes currently under active discussion. Open science practices (Crüwell et al., 2019; Renkewitz & Heene, 2019), for example, will not be discussed extensively in this thesis, but also play an important role in moving the field forward by improving standard practices. While the focus of this thesis lies on the statistical practices discussed in the next chapter, there are close links between the meta-scientific perspective, the statistical perspective, and the perspective of the philosophy of science. Any recommendations for improving scientific methods should be grounded in philosophical principles – most of which are older than the current debate. This is a good indicator, that the current "replication crisis" is not the first era of doubt about findings and methods in psychological science. And likely not the last.

1.2 Perspective from Philosophy of Science

The quotes by Karl Popper at the beginning of the chapter strongly imply that replication studies are necessary. While there seems to be little debate about this position for the natural sciences, the replication crisis has sparked several discussions about the feasibility and necessity of replication studies in psychology. Any position in this discussion should be founded in the philosophy of science. Therefore, it is reasonable to give a very brief overview of the literature on philosophical frameworks seems reasonable.

1.2.1 Falsification and the Hypothetico-Deductive Framework

Current practice in psychology, both in training and in research, focuses nearly exclusively on a *hypothetico-deductive* framework along the lines of Popper (1935)'s stance on falsification.

It is the researcher's task to set-up experiments or observations to put a theory to a test (de Groot, 1969, p. 127):

> Anyone publishing a hypothesis should therefore indicate in particular how crucial experiments can be instituted that may lead to the refutation or abandonment of the hypothesis. The author of a theory should himself state which assumptions in it he regards as fundamental, how he envisages crucial testing of these particular assumptions, and what potential outcomes would, if found, lead him to regard his theory as disproven.

A theory can thereby never be confirmed, but it can be falsified, according to Popper. Popper's introduction of falsification was a stark contrast to the logical positivists before (Popper, 2002, p. 19):

> Thus inference to theories, from singular statements which are 'verified by experience' (whatever that may mean), is logically inadmissible. Theories are, therefore, never empirically verifiable. [..] These considerations suggest that not the *verifiability* but the *falsifiability* of a system is to be taken as a criterion of demarcation [between empirical and non-empirical statements].

Falsificationism has since become one of the most commonly used epistemological foundations for science. For Popper, however, falsification was primary a demarcation between science and pseudoscience in that scientific theory need to be falsifiable. That means a theory must be able to state boundary conditions which, if observed, would provide evidence against a theory and therefore falsify it. The more of such boundary conditions are stated, the more falsifiable, and the more empirical a theory can be seen (Dienes, 2008). To falsify, replication is needed (Popper, 2002, p. 66):

> We say that a theory is falsified only if we have accepted basic statements which contradict it [. . .]. This condition is necessary, but not sufficient; for we have seen that non-reproducible single occurrences are of no significance to science. Thus a few stray basic statements contradicting a theory will hardly induce us to reject it as falsified. We shall take it as falsified only if we discover a *reproducible effect* which refutes the theory. In other words, we only accept the falsification if a low-level empirical hypothesis which describes such an effect is proposed and corroborated.

Researchers shall, therefore, state *falsifiable* theories and aim to refute the theories through repeated experimentation and observation. Popper's framework, however, is not a model of how science operates in practice. Most notably, it does not explain how theories are conceived, developed, and possibly extended through observation. Furthermore, falsification in psychology is complicated by the fact that entities and outcomes are often probabilistic and not deterministic: for example, interventions that increase or decrease the probability of some behaviour. As will be considered later, effects in psychology are never exactly null. This easily leaves room for avoiding falsification in addition to employing "*ad hoc* an auxiliary hypothesis" (Popper, 2002, p. 19).

While most scientists would probably describe Popper's methodology as their "practiced methodology" (Fidler, Thorn, Barnett, Kambouris, & Kruger, 2018, p. 1), it is neither without extension nor without critics. For example, Mayo & Spanos (2006) have linked Popper's notion of putting hypotheses to risky tests with Neyman-Pearson's concept of statistical significance testing (see section 2.1.2) by introducing the concept of "severe testing." For *corroboration*, a hypothesis must pass a severe test, i.e. a test that has a high risk to refute the hypothesis. For psychology, in particular, Paul Meehl is one of the most important figures who has extended the Popperian philosophy (considering himself to be a "Neo-Popperian"), integrating

concepts from philosophers of science such as Imre Lakatos and Thomas Kuhn and applied it to psychological science. In his works, Meehl has voiced several pointed criticisms of research practices and theory development in psychology, especially in what he calls "soft psychology," for over four decades (e.g. MacCorquodale & Meehl, 1948; Meehl, 1967, 1990b). Meehl (1978, p. 806), for example, viewed the maturity of theories in psychology very critically:

> I consider it unnecessary to persuade you that most so-called "theories" in the soft areas of psychology (clinical, counseling, social, personality, community, and school psychology) are scientifically unimpressive and technologically worthless. [...] Perhaps the easiest way to convince yourself is by scanning the literature of soft psychology over the last 30 years and noticing what happens to theories. Most of them suffer the fate that General MacArthur ascribed to old generals – They never die, they just slowly fade away. In the developed sciences, theories tend either to become widely accepted and built into the larger edifice of well-tested human knowledge or else they suffer destruction in the face of recalcitrant facts and are abandoned, perhaps regretfully as a "nice try."

1.2.2 Meehl's Critique of Psychology

While Meehl's work offers much for the critique of psychology as a science still today, three closely related criticisms are of particular interest for the present work concerning replicability and statistical practices: First, the criticism on the use of significance testing in psychology. Second, the distinction (and lack thereof) between substantive and statistical hypothesis. And third, the extensive need for (untested) auxiliary hypotheses.

The debate around the use of statistical significance testing as it is used in practice today will be further elaborated in more detail (see section 2.1.3). In his critique of psychology, Meehl (1967) was already critical about the use of significance testing, especially in comparison to the use of the technique in physics.[2] Physical theories or laws commonly predict point values for physical entities. When aiming to measure these entities using experiments, physicists use significance testing to test the data against the predicted value. Most notably, the predicted value constitutes the *null hypothesis*, which the researcher seeks to falsify (Meehl (1990a) calls this "*strong use* of significance tests"). In contrast, psychological theories rarely predict point values. It is rather much more common that theories either predict a direction, a range of values, or (commonly) that "something is going on" (undirected hypothesis that some true effect is non-zero).[3] Meehl criticises this "*weak use* of significance tests" (ibid.) as in reversal to the actual Popperian ideals of falsification and making it very easy to claim a corroboration of a substantive theory (see also Gelman & Carlin, 2014).

The first concern is amplified by the lack of distinction between substantive theory and statistical hypotheses (see figure 1.1). In the published literature, significant results achieved in the way described before are considered a corroboration (or, worse, as proof) for a substantive theory. However, there rarely is an isomorphic mapping between theories and statistical hypotheses. As long as theories are as trivial as "group A will differ from group B," the undirected two-sample t-test might map to the substantive theory. As theories evolve and include more determinants of a phenomenon, however, the statistical hypothesis usually covers only a part of the theory. Meehl (1990a, pp. 116–117) goes on:

> One reason why psychologists in the soft areas naively think that they have strongly proved a weak theory by a few significant chi squares on fourfold tables

[2] He later repeated and extended this criticism several times (Meehl, 1978, e.g. 1990b, 1990a).

[3] The question whether this constitutes a theory and how often psychologists invoke such theories rather than resorting to even weaker formulations is beyond the scope of this thesis. There is evidence from studies published as registered reports, however, that in practice many "theories" in psychology fail even to state the simplest form of a prediction (Scheel, personal communication).

> is that in their education they learned to conflate *statistical significance* with the broader concept of *evidentiary support*. So they are tempted to believe that if there is nothing wrong with the experimental design, or in the choice of statistic used to test significance, they are "safe" in concluding for the verisimilitude of a theory.

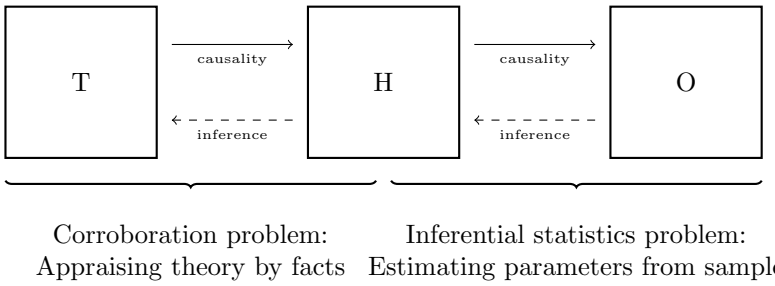


Figure 1.1: Dinstinction between substantive theory (T), statistical hypothesis (H), and observational data (D) as described by Meehl (1990a). *Note:* Figure reproduced from Meehl (1990a, p. 116, Figure 2).

Meehl (1990a) also highlights that the substantive theory consists logically of a conjunction of the substantive theory T as well as auxiliary hypotheses A (e.g. hypotheses about the operationalization of a construct, the reliability or validity of a measure, etc.), C_n particulars of the experimental design, and a *ceteris paribus* clause C_p. Following from the conjunction is the prediction that an observation O_2 follows observing O_1 (Meehl, 1990a, p. 109):

$$(T \land A_t \land C_p \land A_i \land C_n) \implies (O_1 \implies O_2)$$

When experimentally observing the implied observation $O_1 \implies O_2$, we corroborate the theory on the left-hand side. If, on the other hand, we do not observe the predicted pattern and thereby falsify the statement, we cannot falsify T alone. In fact, it is the conjunction in its entirety we have to reject and therefore cannot conclude whether the substantive theory or the auxiliaries are false. In practice, this underlines the need for both close and conceptual replications – as will be explained later – and requires replication researchers to carefully think about the goal and limitations of the particular study under investigation. Specifically, systematic replications involving the same substantive theory but different operationalisations (i.e. auxiliaries) are required to refute a substantive theory – of course under the assumption that the substantive theory is well formulated, falsifiable, and properly operationalised to begin with (see previous points).

In summary, the aforementioned three issues give rise to concerns about the use of statistical significance testing, the relationship between theories and statistical analysis, and underline the requirement for systematic replication from a Popperian perspective.

1.2.3 Alternative Frameworks

While rarely considered in psychological undergraduate courses as well as in the published literature in psychology, there are, however, other approaches to epistemology and inference that deserve attention. Fidler et al. (2018) have given a concise overview of different approaches in the context of evaluating null effects (an issue that will also be considered in section 2.4). Besides a Bayesian philosophy, in which observations are used to reallocate belief in hypotheses or explanations and which is not necessarily linked to Bayesian statistics (see

section 2.2), abductive inference or "inference to the best explanation" (Haig, 2005; Lipton, 2014) provide different approaches to evaluating or appraising theories.

All of the aforementioned positions and points of discussions rest on the assumption that human nature follows generalisable laws and that observations can be used to inform us about these generalisable laws. This assumption strongly informs the natural sciences, especially physics. Whether this assumption holds also for psychology is not without objection: Some psychologists and philosophers have argued that psychology is more akin to history (Lee J. Cronbach, 1957; e.g. Lee J. Cronbach, 1975; Gergen, 1973; Wolfgang Stroebe & Strack, 2014), in that it can make statements only about specific, singular events in time. These events, observed as phenomena and effects in our studies are contingent on many different factors, including contextual factors (e.g. time, place, political climate) and personal factors (e.g. individual experiences, live events, the interaction between subject and time/place), which are also "not time-invariant" ("irreversible units," see Schmidt, 2009, p. 92). Therefore, the induction of general laws about human nature is either impossible or elusive, and we should not expect previous findings to replicate in either direct replications (Wolfgang Stroebe & Strack, 2014) or even conceptual replications (Gergen, 1973). Meehl (1978) also considered this as part of the problem with the progress of psychology but came to a different conclusion. While the many possible influences on behaviour and cognition that are outside of an experimenter's control are indeed a challenge particular to psychology and not shared with most natural sciences, the history of psychology nevertheless shows several successes both in basic research in cognitive psychology (e.g. the Stroop effect) and in more applied, "soft" areas (e.g. the fundamental attribution error). These phenomena have been consistently repeated in independent replications and can therefore be considered to have some "lawful character." As a response to the Gergen (1973)'s positions, Schlenker (1974) laid down how his arguments fail to distinguish psychology from natural sciences: Physical observations are subject to specific times and locations, but the success of theories in physics resides in the abstraction of particular events. Physicists have identified patterns across such singular observations and derived abstract laws. Schlenker (1974) argues that Gergen (1973) has not provided an argument that such abstraction is impossible in psychology and thus psychology might not be so different from natural sciences.

This does not preclude the fact that there are typically contextual factors playing an important role that can hardly be controlled in an experiment. The question then is how such factors can be integrated into theory development. Looking at most published literature, effects and theories are described as being generalisable across a variety of stimuli, populations, settings, laboratories. It is often only after a close replication was not successful, that ad-hoc explanations such as the "hidden moderator explanation" are used to explain the different outcome in the replication. Even if these (previously hidden) moderators can be easily integrated into the theory, this again allows for avoiding falsification and questions the generalisability of the statements originally made. There is yet little consideration about how findings are presented as to be universally generalisable (cf. Simons, Shoda, & Lindsay, 2017), while there is increasing concern that the focus on so-called WEIRD samples is an inherent, but still often unrecognised problem (Henrich et al., 2010).

1.2.4 Summary

There are different frameworks in the philosophy of science to allow researchers to draw inferences from observations to theories. The frameworks most commonly used (Fidler et al., 2018) all consider replications to be important for the progress of science, to corroborate theories (Popper) or to increase belief in theories (Bayesian philosophy). While the more idiographic perspective seems to be at odds with scientific practices and established theories, it reminds psychologists about the constraints and difficulties in researching complex, multidimensional

systems such as humans. Moreover, it is the plurality of viewpoints and philosophies that allow for new ideas, creative theories, and further development of the field.

Scientific methods in practice should nevertheless be founded on principles derived from the philosophy of science. This seems to be a plausible consensus between both philosophers of science, except maybe for Paul Feyerabend's epistemological anarchism ("anything goes"; cf. Shaw, 2017), and empirical researchers. Due to the plurality of theories, empirical paradigms, fields of study, etc., it seems elusive to find a single philosophy of science that informs researchers on how they need to "do science." It seems rather more fruitful if there is a constant exchange of ideas and practices between philosophers of science and empirical researchers. Researchers face demands and challenges in their daily work which philosophers of science might want to incorporate to provide new guidelines on how these demands can be met while maintaining high scientific standards regarding the methods. On the other hand, the plurality of inferential frameworks allows for intellectual diversity that is necessary for the progress of science. In contrast to "anything goes," however, it should be the scientist's goal to employ a consistent framework in both theory development, statistical analysis, theory testing, and subsequently theory extension.

Chapter 2

Statistical Practice

The use of empirical data and subsequently statistical methods is a central part of psychological and, more generally, any empirical research. Undergraduate psychologists are trained in the basic understanding of statistics early on, and there is rarely a paper published in important journals that does not provide empirical data along with statistical analysis. So without a doubt, statistics is one of the most important tools any empirical scientist has available.

Statistical hypothesis testing is the dominant framework of analysing data, mostly in terms of "significance testing." Since its introduction by Ronald A. Fisher and its later extension by Jerzy Neyman and Egon S. Pearson, it has become the primary paradigm in psychology. It is so ingrained in the training of psychology students and the practice of researchers that there is nearly no paper in major journals that does not report a p-value. This practice, however, is not without criticism and has been the target of a heated debate among psychological practitioners, statisticians, and researchers from other fields. While some of the criticism is very general (e.g. Gigerenzer, 2004; Gorard, 2010; Meehl, 1990b), there are several specific concerns about the use and interpretation of significance tests. Within the context of the "replication crisis," increased scrutiny has been applied to the question of how researchers do and should do their statistical analysis (Wagenmakers, 2007; Wagenmakers et al., 2011).

The statistical framework for this kind of analysis is **frequentist statistics**. The naming refers to the understanding of probability within this context, which is based on the frequency of events in drawing repeated samples from a population. There exist other frameworks of statistics involving different understandings of "probability" which are not related to the concept of frequency. The second most relevant framework is **Bayesian statistics**. Both frameworks and their differences will be explained over the next sections. While the heated debate between statisticians of both camps might suggest that there is no middle ground between the two frameworks and schools of thought, one should bear in mind that researchers do not need to choose one side. As will be considered again in a later section, any statistical procedure asks specific questions and it is a researcher's job to select the procedure that is most appropriate for the answer of his substantial question – an approach Kass (2011) described as "statistical pragmatism."

2.1 Frequentist Statistics and Null Hypothesis Significance Testing

When Bachelor students in psychology learn their first inferential statistics, they are usually introduced to the concept of probability in terms of repeated sampling without replacement from an infinite population: A probability is, thus, a statement about the long-run frequency

of an event. This definition of probability (and its epistemological interpretation) is called "frequentism" and the corresponding framework "frequentist statistics," respectively. While the framework involves different procedures, estimators, and statistics, null hypothesis significance testing (NHST) is the most important tool in a researcher's toolbox and is widely applied in practice. It is also the central concern of most criticism about frequentist statistics. Therefore, the focus of the following section is on significance testing.

Significance testing is often perceived as a single, simple method to calculate a p-value and considering it a "significant finding" whenever $p < .05$. This perception has led to grave misunderstanding and misuses of significance testing, which do not fit the statistical and philosophical foundations (e.g. Gigerenzer, 2004; Greenland et al., 2016; Perezgonzalez, 2015). To clarify some of the misunderstandings, to highlight its impact on the replicability of studies, and to compare it to Bayesian statistics, the fundamentals of significance testing are presented along with the writings of Fisher (section 2.1.1) as well as Neyman and Pearson (section 2.1.2).

2.1.1 Fisherian Significance Testing

Statistician and geneticist Sir Ronald A. Fisher is considered the intellectual father of significance testing as it is used today. He popularized this approach in the first edition of his book "Statistical Methods for Research Workers" in 1925, but actually, Pearson (1900) is the first who published the idea of a deviation of observed data from theoretically expected data (Pearson, 1900, p. 158), which Elderton (1902, p. 155) summarised succinctly as the p-value:

> In other words we want to find out the probability P that in random sampling deviation-systems as great as or greater than that actually observed will arise.

Fisher later introduced several statistical tests based on this concept for application in, at the time, common research designs. For the case of what is now known as the "χ^2 test" he wrote (Fisher, 1950, p. 79):

> For any value of n, which must be a whole number, the form of distribution of χ^2 was established by Pearson in 1900; it is therefore possible to calculate in what proportion of cases any value of χ^2 will be exceeded. This proportion is represented P which is therefore the probability that χ^2 shall exceed any specified value. To every value of χ^2 there thus corresponds a certain value of P; as χ^2 is increased from 0 to infinity, P diminishes from 1 to 0.

In other words, the p-value represents the probability of data as extreme as observed or more extreme under a specified hypothesis ("any specified value"). The more extreme the observed data is, i.e. the farther away it is from the "specified value," the smaller this probability p will be. Fisher (1950, p. 41) called "critical tests of this kind [...] tests of significance" and used them to "discover whether a second sample is or is not significantly different from the first" (ibid.).

At what level of probability should one be surprised by the data? Fisher introduced the convention of $p < .05$ (Fisher, 1950, p. 80):

> In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If it is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05, and consider that higher values of χ^2 indicate a real discrepancy.

On the other hand, he also emphasised that, while such a level might be conventional, researchers need to carefully evaluate the context to choose a level of significance which is adequate for a given study (Fisher, 1956, p. 42):

> [...] for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.

In his examples, Fisher (1950) was mostly concerned with nil null hypotheses, i.e. null hypothesis of no difference or two samples coming from the same population. The mechanics of significance testing in the Fisherian sense, however, do similarly apply to null hypotheses of a quantitatively precise, non-zero difference. Furthermore, it can also be applied to directional hypotheses. The formulation of the null hypothesis (nil vs non-nil, undirectional vs directional) simply changes the expected distribution of the test statistic and therefore leads to a different calculated value for p.

To summarise the Fisherian perspective on significance tests (Perezgonzalez, 2015): After selecting a proper test and its statistic, the deviation from a null hypothesis (cf. Fisher, 1950: "specified value") is calculated as p-value. This p-value will then be compared to a certain "level of significance." While .05 is a convenient convention, it is no universal rule to be used in any case (see above). Fisher rather proposed to interpret it continuously as evidence against the null hypothesis (Bakan, 1966): Larger values indicate a larger disagreement of the data with the hypothesis. If the disagreement is large, a researcher is confronted with a binary decision: "*Either* an exceptionally rare chance has occurred, *or* the theory of random distribution is not true" (Fisher, 1956, p. 39). In the context of Fisherian testing, hypotheses are either true or false – Fisher (1956) rejected the notion that probability statements can be made about hypotheses or that such notion is even necessary (cf. chapter 2.2 on Bayesian statistics).

2.1.2 Neyman-Pearsonian Significance Testing

Neyman & Pearson (1933) built upon the significance testing framework introduced by Fisher and most notably added the concepts of alternative hypotheses and power. They formalised some aspects of significance testing more strictly and sought to find "rules to govern our behaviour [...] which [...] insure that, in the long run of experience, we shall not be too often wrong" (Neyman & Pearson, 1933, p. 291). That is, Neyman and Pearson's approach to statistical significance is one of *inductive behaviour* (Mayo & Cox, 2006). If a researcher finds a significant result, they should act is if the null hypothesis is false, or, conversely, if they find a non-significant result, they should act as if the null hypothesis is true. In repeated experiments on the same hypothesis with different samples, the researcher will "not be too often wrong" in the long run. This, already, is a stark contrast to Fisher's view of p being only a measure of evidence against the null hypothesis and a researcher's freedom to make a decision [see above; Fisher (1956), p. 39].

Formalising their rule of inductive behaviour, Neyman & Pearson (1933, p. 291) write:

> [...] to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts ; if $x > x_0$ reject H, if $x < x_0$ accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x \leq x_0$ or false when $x > x_0$.

Neyman & Pearson (1933) thereby recognise that the "true state" of a hypothesis, whether it is true or false, cannot be inferred from the hypothetical frequency of an event (observing data as extreme as or more extreme than the observed data) given the hypothesis. They concluded (Neyman & Pearson, 1933, p. 296):

> If we reject H_0, we may reject it when it is true ; if we accept H_0, we may be accepting it when it is false, that is to say, when really some alternative H_t is true.

On this ground, Neyman & Pearson (1933) introduced the concepts of two kinds of errors to the statistical testing framework: The "error of the first kind" for when rejecting H_0 when it is true, and the "error of the second kind" when failing to reject H_0 when it is false and some alternative H_t is true. The two long-run rates of errors are denoted by α and β, respectively. While missing from the Fisherian framework, the alternative hypothesis H_t is central to the Neyman-Pearson framework as its formulation allows to calculate the two error rates. To set up the significance test in the Neyman-Pearson framework, a researcher also needs to decide what error rates they wishes to achieve. They does so by choosing a level of significance α *a priori* along with the desired power $1 - \beta$ (e.g. 90%) and the smallest effect size of interest (SESOI) to calculate the number of participants needed to ensure the chosen long-run error rates (Cohen, 1988).

Most importantly, the value x_0, against which the observed statistic is tested, must be chosen before the test is run. That means that the "level of significance" is set *a priori*. Further, the "level of significance" is denoted by α, representing the long-run frequency of errors of the first kind. While the exact value of p can also easily be calculated in exactly the same way as in the Fisherian framework, it has a different meaning for Neyman-Pearson: It is the lowest α that would have led to a rejection of the null hypothesis. While informative to researchers, the behavioural rule requires a researcher to stick to the selected level of significance when analysing the data. Only then, the error rates are guaranteed over the long-run and an informed and consistent decision can be made based on the data.

Summarising the Neyman-Pearson approach to significance testing (Perezgonzalez, 2015), it also involves choosing an appropriate test statistic. Before collecting or analysing data, however, a researcher also needs to, first, select an appropriate level of significance, which determines their long-run rate of type I errors. Next, the desired power is chosen, which allows one to calculate the number of participants needed to test the hypothesis under investigation. After running the experiment and collecting data, the test statistic and associated p-value are calculated and compared to the pre-determined α level. Neyman-Pearson's rule of inductive behaviour now states that, if the test is significant ($p < \alpha$), the researcher should act as if the alternative hypothesis is true. When being strict, this rule can guarantee mathematically that the long-run error rates indeed hold as long as the assumptions of the framework are met.

Fisher (1955), however, rejected the Neyman-Pearson approach to significance testing furiously: He understood the proposed extensions to his introduction of significance tests as "acceptance procedures," which are more apt to commercial testing in a factory (Fisher, 1955, p. 69) and rejected their use for statistical analysis in scientific studies fundamentally.

From today's perspective, the Neyman-Pearson framework is an important extension of the concepts introduced by Fisher (1950). Considering the use of terms like "power" and "alternative hypothesis," it is also the framework used by most researchers when applying significance tests (Perezgonzalez, 2015). The framework has been adapted and extended, e.g. by Mayo & Cox (2006) and Mayo & Spanos (2006), who have used the error-statistical framework by Neyman & Pearson (1933) to introduce a consistent theory for inductive inference.

2.1.3 Current Use of Significance Testing

Frequentist significance testing makes several assumptions about the data and their origin (e.g. the sample is a fully random sample from a defined population), and the Neyman-Pearson framework sets up strict rules about applying tests and drawing inferences. Are those assumptions met in general and do researchers apply the framework correctly? Since their introduction, both the Fisherian and the Neyman-Pearsonian approach to testing have

been criticised for both their theoretical justification and their practice in research. While a full discussion of all points of criticism is beyond the scope of this thesis, it is notable that several arguments have reappeared over the last years also in the context of the replication crisis. In particular, some authors see the application of frequentist significance testing as a reason for low rates of replicability in psychology research. In this section, some of the arguments and criticisms are being discussed: The lack of power in psychological experiments, misunderstandings of core concepts, the misuse of significance testing, and subsequently the use of *p*-hacking.

Meehl (1990a) calls the way of using significance tests as it has been used in practice over the last decades "weak use" (Meehl, 1990a, p. 116, emphasis in original):

> Whereas the social scientist's use of chi square in a fourfold table, where H_0 is that "These things are not related," I call the *weak use*. Here, getting a significant result depends solely on the statistical power function, because the null hypothesis is always literally false.

The latter point is related to what Meehl (1990b) called the "Crud factor." Meehl noted that "it is highly unlikely that any psychologically discriminable stimulation which we apply to an experimental subject would exert literally *zero* effect upon any aspect of his performance" (Meehl, 1967, p. 109). The "Crud factor" is a fundamental criticism of point null hypotheses in any hypothesis testing framework.[1] However, a precise estimate and a proper definition are yet missing (Orben & Lakens, 2020). Generally, in this context, interval null hypotheses and well-justified sample sizes, based on the expected effect size or the smallest effect size of interest would be preferable to point null hypotheses. This issue is further discussed in chapter 2.4. However, point null hypotheses remain predominant in the use of significance testing since Meehl's statement.

As already discussed, the distinction between substantive theories and statistical hypotheses is fundamental for the inferential process. It is furthermore important to note that a substantive theory does not have a one-to-one mapping to a statistical hypothesis. For one, a substantive theory might consist of different operationalizations and aspects (auxiliary hypotheses in the sense of Meehl, 1990a). And secondly, even for a simple qualitative question such as "Is there a difference between the two groups?" different statistical hypotheses can be formulated. Hand (1994) showed in several instances how these different questions have been considered and how they might even provide different answers.

2.1.3.1 Low Power of Psychological Research

As noted before, *power* is a critical part of significance testing in the Neyman-Pearson framework. It is the long-run probability of a test to correctly reject a null hypothesis that is indeed false. While power can be thought of as a curve as a test has a certain power for any effect size, it is usually considered as a single probability value for a specific effect size (Cohen, 1988). The general recommendation is to select a sample size that has sufficient power to detect the *smallest effect size of interest* (SESOI) (Anvari & Lakens, 2020; Lakens, 2014). There is no universal rule on how much power a test should have to be considered "sufficiently powered" or "powerful enough." In practice, commonly a power of .80 to .90 is recommended (Cohen, 1962, 1988). This goal is rarely achieved in psychology: Cohen (1962) was the first who systematically investigated the statistical power in a set of studies – and found that researchers had only little chance to detect effects if they were not very large with a median power of .46. In follow-up studies, Sedlmeier & Gigerenzer (1989) and Button et al. (2013) found similar

[1] Meehl later softened this criticism, as there can be "precise nulls" in some experimental settings (Orben & Lakens, 2020, p. 5).

findings for different and newer sets of studies: Despite discussions among statisticians and methodologists, there is no evidence for an increase in statistical power over the last decades.

While low power primarily seems to be a problem when a non-significant result is obtained, it is also a problem for published significant studies when there is large measurement error, publication bias, or when questionable research practices are being used (Loken & Gelman, 2017; Wagenmakers et al., 2015). For a significant result, low power decreases the likelihood that the finding represents a true effect (Button et al., 2013). Furthermore, running an experiment with low power is an inefficient and ineffective use of samples, and there is little reason to conduct studies with low power (Perezgonzalez, 2015).

2.1.3.2 Misunderstandings and Misuses of Significance Testing

Several concepts of significance testing are regularly misunderstood. Nickerson (2000) has given the most concise overview of misunderstandings of null hypothesis significance testing. One of the major misunderstandings is the probability expressed in the p-value. As outlined concisely in Fisher's texts, but sometimes misrepresented in textbooks, p is the probability of observing data as extreme as or more extreme than the observed test statistic under the condition that the null hypothesis is true. It is *not* the probability of the null hypothesis given the data or the probability that the null or the alternative hypothesis is true. As briefly noted before, in frequentist statistics the hypothesis itself cannot be associated with a probability – it is either true or false and the actual state cannot be observed directly.

Another, related, misunderstanding concerns the conclusion whether the significant p-value implies that the alternative hypothesis is true (Nickerson, 2000, pp. 254–255). As much as a non-significant result does not imply that the null hypothesis is true (see below), a significant test itself does not provide this answer. It is more that the p-value can provide evidence *against* the null hypothesis without reference to any alternative hypothesis (Fisherian perspective) or that researchers should *act* upon the result as if the alternative hypothesis was either true or false (Neyman-Pearsonian perspective). This is, in fact, less a statistical misunderstanding but an epistemological one, which is certainly fueled by a misunderstanding of the foundations of significance testing.

As an extension to significance testing and an attempt to address those misunderstandings, some have tried to apply the "diagnostic screening" example to significance testing (Colquhoun, 2014; Ioannidis, 2005). The diagnostic screening example usually involves a population of patients that have or do not have a disease, e.g. cancer, and for which we use screening tests in order to find patients carrying the disease. Using sensitivity and specificity of the screening, we can calculate the probability that a patient has a positive screening test given he is either sick or healthy – very much like the outcome of a significance test given the null hypothesis. If we aim to find the probability that a patient is sick or healthy given the test outcome, i.e. to calculate the positive predictive value (PPV) or negative predictive value (NPV), we need to consider the base rate of the illness, which is commonly unknown or hard to properly estimate. For a clinical disease, prevalence or incidence can provide the necessary estimates, but for statistical hypothesis, this is a challenging task: How high is the rate of true null hypotheses? Since the space of possible hypotheses is difficult, if not impossible, to count, estimating the rate is equally challenging. Making some assumptions, Ioannidis (2005) nevertheless used the diagnostic screening example to estimate the rate of false findings in the research literature. Colquhoun (2014) called this the "false discovery rate" (FDR). Others have used different terminology for a similar approach (e.g. Bayarri, Benjamin, Berger, & Sellke, 2015). Mayo & Morey (2017), on the other hand, argue that the base rate for true hypotheses cannot be properly defined ("reference class problem"), and, therefore, the FDR does not have meaning in the context of significance testing. Further, a high PPV calculated in this way does not, per se, indicate strong evidence.

Over the short history of significance testing, several misunderstandings, misconceptions, and misapplications have arisen. Most importantly, a hybrid framework for significance testing has emerged in practice that combines aspects from both the Neyman-Pearsonian and Fisherian frameworks (Bakan, 1966; Gigerenzer, 2004; Nickerson, 2000; Perezgonzalez, 2015). This framework leads to running significance tests without planning sample sizes, calculating post-hoc power (cf. O'Keefe, 2007), interpreting a p-value loosely on its magnitude, and combining the interpretation with error rates. These interpretations are mostly inconsistent with both Neyman-Pearson (e.g. interpreting the p-value as a continuous measure) or Fisher (e.g. considering the power of a test), and violate the meaning of probability statements (e.g. when interpreting the p-value as the probability of the null hypothesis being true or false).

The problem often seems to be that researchers simply apply statistics as if there is a default way to do it and if this way works with whatever data or hypothesis there is. Gigerenzer (2004) called this the "mindless" application of significance tests in the "null ritual" (Gigerenzer, 2004, p. 588):

> 1. Set up a statistical null hypothesis of "no mean difference" or "zero correlation." Don't specify the predictions of your research hypothesis or of any alternative substantive hypotheses.
> 2. Use 5% as a convention for rejecting the null. If significant, accept your research hypothesis. Report the result as $p < 0.05$, $p < 0.01$, or $p < 0.001$ (whichever comes next to the obtained p-value).
> 3. Always perform this procedure.

The ritual implies that researchers analysing their empirical data commonly apply a statistical tool in a pre-set way without considering their choice of tool or underlying assumptions properly. Gigerenzer (2004), in particular, notes how prevalent the "null ritual" is in text books and introductory materials. While seemingly more nuanced perspectives are provided, examples often only cover the ritual as essence to statistical testing. Through textbooks and published articles, this "ritual" has become ingrained in research culture, mostly in psychological research, as the way to do statistics. By applying the "null ritual," researchers miss the opportunity to use the whole statistical toolbox to select a tool more appropriate to study design and statistical questions. The "null ritual" often is essentially described as a grave misrepresentation of the testing procedures introduced by Fisher or Neyman-Pearson, with its common use in science leading to spurious results. The core of this "ritual" seems to be confusion around, or a lack of, a deeper understanding of significance testing.

2.1.3.3 p-Hacking and Questionable Research Practices

The misunderstandings about significance testing have in part helped to establish practices that have the ability to further undermine the trustworthiness of test results. Simmons et al. (2011) have provided some illustrative examples of how practices that are not uncommon in psychological research can lead to nearly any result being statistically significant. They have termed these "p-hacking" because the goal in using those practices is usually to achieve significant p-values.[2] In one of their examples, the authors have shown how listening to a particular song will lead to participants being younger than before. They achieved this result mainly by selectively reporting outcomes and including/excluding covariates until the study "worked."

What Simmons et al. (2011) have done to achieve significant results has often been common practice. It is also what John et al. (2012) and Fanelli (2009) have termed "questionable

[2] Based on this observation, the authors have introduced a test sensitive to p-hacking: the p-curve analysis (Simonsohn, Nelson, & Simmons, 2014). See section 3.3.1.1 for an example.

research practices" (QRP). This includes selectively reporting outcomes and experimental conditions, collecting more data if results are not (yet) significant, rounding *p*-values to meet the significance criterion (e.g. .056 becoming .05), and (as a most severe case of QRP) falsifying data. Many of the practices are, however, common even when precise prevalence estimates differ (Fiedler & Schwarz, 2016; John et al., 2012). From a statistical perspective, these practices all lead to inflated error rates, meaning that the assumptions of significance testing are violated and resulting in more false positives and more false negatives than should be expected. This, combined with an incentive to publish significant results ("publication bias"), has led Ioannidis (2005) to consider "most research findings to be false."

This shows how important it is for researchers to use statistical tools in the way they were designed, check assumptions properly, and exercise professional judgment when using statistics. Pre-registration and openly providing analysis scripts are not a panacea but can shed light on the research progress and allows to more critically evaluate statistical analyses. Especially in a field where applied researchers perform statistical analyses on their own and usually without the help from statisticians, it allows outside experts to critically examine the analysis and results. The prevalence of questionable research practices or *p*-hacking highlights the need to change current practices.

2.1.4 Improving Statistical Practice

The largest group of statisticians, the *American Statistical Association* (ASA), has published a note on how to use and interpret *p*-values in 2016 (Wasserstein & Lazar, 2016), and gave five principles that should guide statisticians and researchers using significance testing and *p*-values in practice (Wasserstein & Lazar, 2016, pp. 131–132):

1. *p*-values can indicate how incompatible the data are with a specified statistical model.
2. *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

These principles are a brief summary of the basic principles of significance testing, but not limited to statistical interpretation. Especially points 2 to 4 refer to how a particular *p*-value needs to be considered in the context of the study and the reporting of said study. The ASA statement is a reminder what the *p*-value is not, but does not provide a guideline or positive examples for practitioners on how to use it.

The increasing criticism of null hypothesis significance testing both in its currently practised form and in its correct original way has led to several initiatives to change the current practice of analysing data through significance testing. While some favour Bayesian statistics as a complete replacement (see next section), others advocate for stricter rules in data analysis. As one way to improve the reliability of significance testing results, *pre-registration* has been repeatedly advocated (Nosek, Ebersole, DeHaven, & Mellor, 2018). In essence, *pre-registration* means to write down hypotheses, data collection plans, and analysis plans, including rules for data exclusion, the level of significance, and the desired number of participants, before running a study and create a time-stamped version of this document. It is later made available

alongside the study, so readers can compare the actually performed analysis against the preregistered one. This reduces the ways a researcher can make data-dependent decisions in the analysis which will result in inflated error rates.

Another initiative has advocated changing the default level of significance from .05 to .005 (Benjamin et al., 2018). Their argument is primarily based on Bayes factors (see section 2.2.2) and the statistical evidence that can be achieved by the significance levels. As a reply, Lakens et al. (2018) have instead called researchers to "justify [their] alpha," i.e. choosing a level of significance according to costs and benefits of different errors and making an informed decision. While they recognise the difficulties of a justification, their approach is indeed in line with Fisher (1956, p. 42) as quoted before.

Linking back to Meehl's notion of the "weak use" of significance testing, one might also rethink when to actually use hypothesis testing. As Gigerenzer (2004) has noted, it is currently used out of convention or as a ritual. Meehl (1990a) contrasts psychologists' use of significance testing to its use in physical sciences, where theories make point or range predictions and experimental observations are tested against these predictions in order to refute the theory and its predictions. This practice is much more in line with the hypothetico-deductive framework and a more consistent use of significance testing. In contrast to this "hard use," psychologists should perhaps be more interested in using statistics in order to *estimate effects* and *reduce uncertainty* through more precise measurement and robust experimentation. Along these lines, Amrhein, Greenland, & McShane (2019) have recommended to "abandon statistical significance" and "embrace uncertainty," for example by using Bayesian approaches to statistical inference (Amrhein, Korner-Nievergelt, & Roth, 2017; e.g. McShane, Gal, Gelman, Robert, & Tackett, 2019), as will be introduced in the next chapter.

2.2 Bayesian Statistics

While the frequentist statistical framework is often the only framework taught to students (and therefore the most commonly used one in published work), there are also other frameworks resting on different assumptions and using different definitions of "probability." Besides Bayesian statistics, which usually uses probability as a statement about *degrees of belief* (Edwards, Lindman, & Savage, 1963; Jaynes, 2003), there also exists a statistical framework involving quantum probabilities (Gomatam, 2012; Haven & Khrennikov, 2016; Srinivas, 1975). In essence, all statistical frameworks aim to quantify uncertainty. They differ with respect to how this uncertainty is interpreted and what entities have uncertainty associated. Even among proponents of Bayesian statistics, there are different schools of thought on how to build and justify models, how to interpret results, and how to make useful inferences from the results. Some of these differences will be further investigated later when comparing Bayesian to frequentist statistics.

The following section will briefly introduce concepts from Bayesian statistics as far as it is necessary to understand the basics for Bayes factors and points made in the discussion about using Bayesian statistics. Several textbooks provide a far more detailed introduction with illustrative examples (Gelman et al., 2013; Kruschke, 2014; McElreath, 2016). Since there are different schools of Bayesian statistics, any textbook introduction might emphasise a different aspect or provide different justifications for modelling practices.

For any Bayesian approach, the core is **Bayes' rule** about the relationship of conditional probabilities:[3]

[3]The rule is also known under the name "Bayes's theorem," but since it follows logically from the axioms, it is rather a rule than a theorem. Historically, it should also be noted that Reverend Thomas Bayes did not publish the formalised rule. It was Richard Price who rediscovered, edited, and posthumously published the work. Mathematician Simon Laplace later used and refined the rule and is therefore as much as important to the development of Bayesian statistics as is Thomas Bayes.

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

This rule can be deduced from the Kolmogorov axioms of probability calculus that provide a justification for using calculus with probability measures. In contrast, some Bayesians favour Cox' theorem (Van Horn, 2003) as an alternative set of rules to use Bayesian logic not only for statistics but also for logic (e.g. Jaynes, 2003). While this is another concern for the debate between philosophers and statisticians, it is rarely of concern for applied users of Bayesian statistics.

While Bayes' rule applies whenever conditional probabilities are used, Bayesian statistics use the rule to update model inferences through incoming data. As there are many schools of Bayesian statistics, there are different interpretations depending on how "probability" is understood. For the sake of introducing the relevant terms and providing an intuitive approach to Bayes factors, the interpretation of "belief updating" is used in the following section. In this context, the terms *prior* and *posterior* can be easily understood as *beliefs* about values or propositions. While this implies a subjective interpretation of probability (i.e. each researcher can hold different *beliefs*), there are other interpretations with little interest about individual beliefs (Gelman & Shalizi, 2013). A brief discussion on this topic will follow after the introduction.

The following example (adapted from Harms, 2018) illustrates the basic logic of using Bayes' rule for updating beliefs: Imagine a scientist, Sam, who is new to a particular field of research. She might begin with reading books, papers, or blog articles to get an understanding of the current state of the field and to learn more about some effect in particular. Before reading the first paper on a certain effect, she might be "uninformed," i.e. she believes that any outcome of an experiment is equally likely. After reading the paper, though, her beliefs on the outcome have shifted: As she has learned about a statistical relationship and an estimated effect size, she is now more inclined to believe that the effect is of this certain effect size she read about when running the particular experiment. However, Sam already knows that no single experiment is sufficient to provide an exact estimate of the effect size. So she reads another paper on this effect. In contrast to her first paper, she does not read the paper with an uninformed mind but will use her experience from the first paper as a starting point for reading the next paper. After finding a new estimate of the effect size, she will now combine the results from both papers in her own expectation for running the same experiment. That is, Sam uses incoming data (results from previous experiments) to inform her belief about this effect. In the context of Bayesian statistics, Bayes' rule is used to combine previous information (*prior* information) with new data (incoming through a particular *likelihood*) to inform an updated belief (*posterior* information).

Even when Bayesian statistics is not interpreted in the context of individual beliefs, the terms *prior*, *likelihood*, and *posterior* are used to denote the three ingredients to Bayesian analysis. With these terms, Bayes' rule can also be expressed as a statement about proportionality:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Comparing the two representations of Bayes' rule shows that likelihood and posterior are conditional probabilities. This becomes more transparent when mathematical terms are used. For each Bayesian analysis, a probability model $\mathcal{M}$ is given. The model posits the statistical relationship between observable (data) and unobservable entities (parameters), respectively. The parameters are commonly denoted as vector θ and the observed data as Y. The **prior** is now the unconditional probability for the parameters, $\pi(\theta)$. If real-valued parameters are assumed, i.e. that $\forall i : \theta_i \in \mathbb{R}$, the prior, $\pi(\theta)$, is a continuous probability distribution. The distribution provides a probability for each possible value of θ_i. This shows two fundamental properties that are distinct from the frequentist framework:

1. Parameters are random variables with a probability distribution. In frequentist models, parameters are considered to have a fixed, true value that is estimated from the data. Bayesian statistics aims to update the probability distribution of parameters considering the observed data.
2. Bayesian statistics do not result in a singular estimate for the parameter, but in an updated probability distribution for the parameter values.

The second property is apparent when considering the **posterior** distribution that results from applying Bayes' rule:

$$P(\theta|Y) = \frac{\mathcal{L}(Y|\theta)\pi(\theta)}{P(Y)} \tag{2.1}$$

The left-hand term, $P(\theta|Y)$, is the posterior distribution of the parameters θ *given the data* Y. It provides a probability distribution for the parameters after taking the observed data into account. $\mathcal{L}(Y|\theta)$ is the model likelihood for the data given the parameter values. The likelihood function is determined by the model, which will be explained in more detail in the next section with an example of Bayesian modelling. While prior and likelihood are chosen by the analyst, the term $P(Y)$ follows automatically. It is the unconditional probability of the data, which is hard to interpret and which can be rather seen as the necessary normalising constant: As the likelihood is not a probability distribution since it does not integrate to 1, the numerator of the term is also not a proper probability distribution. The **marginal likelihood**, $P(Y)$, is thus a constant to make the posterior integrate to 1. Therefore, it can also be calculated through integration of the numerator: $P(Y) = \int \mathcal{L}(Y|\theta)\pi(\theta)\ \mathrm{d}\theta$. The marginal likelihood, or **model evidence**, is of particular interest for Bayes factors, as will be seen in a later section.

The set-up of a Bayesian analysis offers different degrees of freedom: Not only the model itself, and thereby the model likelihood function, can be chosen by the analyst but also the prior distributions for the parameters. While the former is identical to the model specification in a frequentist model, the latter is a major cause for discussion between frequentist and Bayesian statisticians. Many critics of Bayesian statistics consider the prior as a possibility to have the Bayesian analysis result in any outcome the analyst desires. The argument that frequentist statistics does not have similar degrees of freedom falls short as has been shown through the discussion on p-hacking and other methods that render significance testing nearly uninterpretable or even invalid. It is, however, necessary that the choice of prior distributions is transparent, justified, and reasonable. There are different ways to justify the choice of a prior distribution: there are mathematical considerations (e.g. conjugate priors) for the distribution family and guidelines available for choosing prior distributions. An important tool, and a recommended step in a Bayesian workflow (Schad, Betancourt, & Vasishth, 2019), are simulations to show the impact of a chosen prior distribution on the resulting inferences (prior predictive checks).

There are different perspectives on the issue of subjectivity: While some Bayesian statisticians and Bayesian philosophers consider the subjective interpretation of probability an advantage (Berger & Berry, 1988; Sprenger, 2015), others have considered it to be not in line with the objective rigour required by science. The work of Andrew Gelman is mainly driven by a mechanical view, or a utilitarian perspective, in which the Bayesian framework offers the most appropriate tools for statistical analysis (Gelman & Hennig, 2015; Gelman & Robert, 2013; Gelman & Shalizi, 2013; Gelman, Simpson, & Betancourt, 2017; Gelman & Yao, 2020).

2.2.1 Bayesian Modelling

As an example to show the different ingredients to a Bayesian analysis and to prepare the introduction of Bayes factors as a tool for model comparison and hypothesis testing, consider

the case of linear regression: An outcome y_i is regressed onto a linear combination of, for example, two predictor variables x_{i1} and x_{i2} for an individual observation i. For the purpose of the example, consider predicting a person's weight (y_i) through their height (x_{i1}) and age (x_{i2}).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

ϵ_i are the error terms that are commonly assumed to be uncorrelated and random with expectation 0 and unknown standard deviation σ^2. The statistical goal of inference is to provide estimates for the parameters β_0, β_1, and β_2 using the available data. In frequentist statistic, commonly the *ordinary least squares* (OLS) estimate is used, which minimises the squared residuals.

For a Bayesian analysis, the model would need a formulation that includes prior distributions for the parameters. Different kinds of priors are possible, and an important part of developing a Bayesian analysis lies in the selection (and possible testing) of priors for parameters. A very broad distinction is between informative and non-informative priors: While non-informative priors aim at providing at least some regularisation for the estimates (Cawley & Talbot, 2007; Fahrmeir, Kneib, & Konrath, 2010), informative priors have a stronger influence on the resulting posterior distribution. The latter is only reasonable if there really is prior information available for a model quantity.[4] Furthermore, the particular details of a prior distribution should always be considered in the context of the model and the likelihood (Gelman et al., 2017): Whether a prior is informative or non-informative can, for example, depend on the scale of the parameters, and a reparameterisation of the model can lead to a different effect of the prior on the inference. *Prior predictive checks*, i.e. generating model predictions based on priors only and without data, can help to visualise and investigate the effect of priors on inferences. Generally, for informative priors, a defensible choice needs to be made, made transparent, and ultimately also checked for robustness. For the present example, weakly regularising priors are chosen that will shrink the parameter estimates to 0 when only little data is available. In a simple model, the prior is quickly overruled by the likelihood of the data. When only a few observations are available, one can argue that the priors' regularisation is important in order not to be led astray by single (possibly outlier) observations.

Extending the example model from above by adding prior information, the whole model definition could look like this:

$$\begin{aligned}
y_i &\sim \mathcal{N}(\mu, \sigma) \\
\mu &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \\
\beta_0 &\sim \mathcal{N}(0, 100) \\
\beta_1 &\sim \mathcal{N}(0, 100) \\
\beta_2 &\sim \mathcal{N}(0, 100) \\
\sigma &\sim \text{Cauchy}^+(0, 10)
\end{aligned}$$

This Bayesian model specification reveals an important point: While the frequentist formulation of the model considers the β parameters to be population parameters with a fixed value that is estimated through repeated random sampling, in the Bayesian model, on the other hand, the parameters are considered to be random variables with prior distributions. When applying Bayes theorem, observations and their likelihood in the model are used to find posterior distributions for the parameters. The posterior distributions are then informed by both

[4] Commonly, "model quantity" would refer to parameters. Note, however, that in a Bayesian model also transformed parameters, e.g. an effect size measure, has a posterior distribution. Furthermore, observable quantities can also be modelled as a random variable (e.g. to model missing data or measurement error) and have a posterior distribution.

Table 2.1: First five samples from dataset 'Howell1'.

height	weight	age	male
151.765	47.82561	63	1
139.700	36.48581	63	0
136.525	31.86484	65	0
156.845	53.04191	41	1
145.415	41.27687	51	0

the prior distributions and the observations. As a result, the Bayesian analysis does not provide a single parameter estimate, but a distribution for each parameter. Thereby, an analyst can see the posterior probability for ranges of parameter values.

To complete the example, consider the following data set `Howell1` provided by McElreath (2016), as shown in table 2.1.[5] Using this data, one can find a model to predict a person's weight based on their height and age. For the Bayesian model, we are using weakly regularising priors, i.e. priors centred on zero with a rather widespread. Figure 2.1 visualises the resulting parameter estimate from both the Bayesian model (blue circles) and the ordinary least square estimation (red triangles). The point estimates from OLS and the mode of the posterior distribution (maximum a posterior estimate) are nearly identical. Since only weak priors were used, enough data is present, and since the model is very simple, the estimates do not differ. For the Bayesian model, however, the posterior distribution for the parameters offers additional information compared to the point estimate of the OLS method – namely a 95% credible interval which is a summary of the whole posterior distribution. It is an automatic result of the Bayesian approach and available for any parameter, including the variance parameter. Frequentist methods can build confidence intervals, for example through bootstrapping. The Bayesian posterior, in contrast, offers more information as for any range of values a posterior probability can be calculated.

Software packages such as `brms` or `Stan` allow for easy and fast model building and estimation. The major challenge in using Bayesian statistics lies in the computation of the posterior distribution. Often, it cannot easily be calculated algebraically and has to be approximated through computational methods. The details of the computation are beyond the scope of the current work, but the constant increase in computer power and algorithms has helped Bayesian statistics to receive more and more attention. Software packages like `Stan` use sophisticated methods to draw samples from the posterior distribution through so-called Markov Chain Monte Carlo, MCMC for short, techniques (Gamerman & Lopes, 2006; Han & Carlin, 2001). For some cases, e.g. "Big Data" applications, even these methods cannot provide results in a reasonable time, and more approximate methods are available such as *variational inference*. An important part of using such software packages is to check the output not only for their statistical results but also for the computational accuracy. If the software package was not able to provide proper posterior samples, the results cannot be trusted and should be disregarded.

There is ongoing research on how to establish a reproducible and principled workflow when building, estimating, and checking Bayesian models using these software tools (e.g. Schad et al., 2019). An important part of the model-checking phase is to use predictions made by the model with and without data. The former is called a *prior predictive check*: Without observed data, what does the model and the priors, in particular, predict? If predictions are not sensible and not in line with either expectations or sensible mathematical considerations, the model or selected priors might not be a good choice. The latter is *posterior predictive checks*: After fitting the model and updating the joint posterior distributions, this kind of check is used to

[5] see https://github.com/rmcelreath/rethinking/blob/master/data/Howell1.csv.

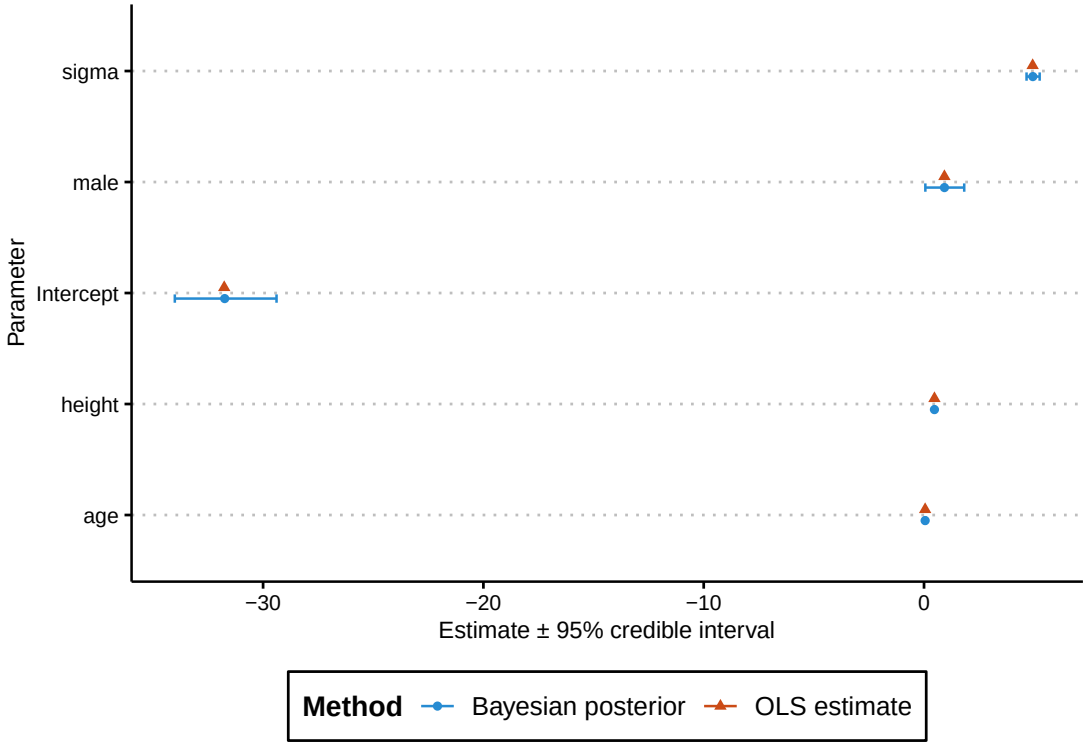


Figure 2.1: Visualisation of estimates for regression parameters from both the Bayesian and OLS model. Error bars for Bayesian results are 95% credible intervals based on the posterior distribution. Intervals for `height` and `age` coefficients are very small and therefore not visible in the plot. `sigma` represents the variance parameter of the regression models.

generate predictions for the observed outcome. The resulting predictions are then compared to the actually observed outcomes. Deviations between the predictions and the observations can highlight incompatibilities between the model and the observed data.

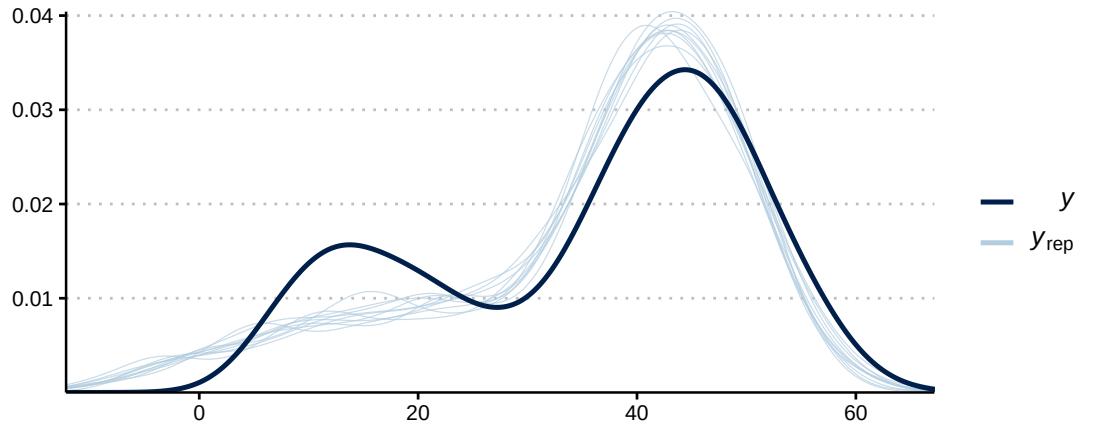


Figure 2.2: Posterior predictive checks for the Bayesian model of the `Howell1` data. The dark blue line is the kernel density estimate of the data, light blue lines are multiple replicated datasets from the posterior distribution of the model. The model does not properly predict the multi-modal structure of the empirical distribution.

Figure 2.2 visualises the posterior predictive checks for the model above. The light blue lines are predictions made by the model, which can be compared to the actual empirical distribution of the observed data. As can be seen in the figure, the model does not properly predict the multi-modal distribution very well. The researcher now needs to evaluate whether the deviation is minor for the purposes of his analysis (e.g. because they are mainly interested in means), or whether they need to adapt the model, e.g. by including additional predictors or changing the model family. Model development and model selection are beyond the scope of this introduction, but an iterative approach to developing a model and checking different model variations against the observed data is a different but rewarding approach to using statistics in analysing data.

Note that this kind of comparison between predictions and observed data can also be used for frequentist models. However, it is not common practice in psychology, as most researchers focus solely on the resulting parameter estimates by printing a coefficient table with significance tests. If at all, the model is checked by looking at R^2 or information criteria. While the latter can be useful for comparing and selecting models, visualisation and exploration of the model behaviour are rarely used. For many studies, especially in applied fields of psychology, the linear model is deemed as a useful and valid approximation of the data at hand – an assumption that often goes unchecked. More consideration of the statistical model used can improve the statistical inferences. Fields such as cognitive psychology are historically closer to a mathematical approach to statistics as more applied fields. Statistical modelling requires a different, possibly more advanced, statistical understanding that is rarely taught for both students and junior researchers.

2.2.2 Bayes Factors

The previous introduction of Bayesian modelling leads to the question: How to select between different, competing models? Different approaches to this challenge are possible, and there

is an ongoing debate about the adequacy of quantitative measures and principled guidelines. Information criteria are favoured by many modellers, and common criteria are the Bayesian Information Criterion (BIC) or the Widely Applicable Information Criterion (WAIC) (Vehtari, Gelman, & Gabry, 2017). Leave-One-Out-Cross-Validation (Vehtari et al., 2017) is another approach with an easily understandable background. For many Bayesian statisticians, these approaches are the most reliable and useful ones. In some cases, i.e. when comparing nested models (such as when comparing models with and without a specific parameter as in the case of a t-test), using information criteria is similar to using a likelihood ratio test (Dziak, Coffman, Lanza, Li, & Jermiin, 2018), showing an interesting link to frequentist methods. A different approach to information criteria, with a different philosophy, is the **Bayes factor**.

Bayes factors have garnered particular attention in psychology as they have been proposed as a direct and easy alternative to significance testing (Wagenmakers, 2007; Wagenmakers, Morey, & Lee, 2016). To better understand the use of Bayes factors, one needs to think as hypothesis testing as a problem of model selection: Hypotheses can be formulated as statistical models, and the tester's task is to select the hypothesis that is more in line with the data. At this point, frequentist statistics focus on the long-term error rates of the decision made for either hypothesis. The natural Bayesian approach would be to look at the posterior probability of a model. Having two competing models $\mathcal{M}_0$ and $\mathcal{M}_1$, a researcher might have some prior odds for the two models $\frac{\pi(\mathcal{M}_0)}{\pi(\mathcal{M}_1)}$. Prior odds of 1 would indicate that the two models are *a priori* equally likely, while prior odds of 2 would indicate that model $\mathcal{M}_0$ is two times more likely than $\mathcal{M}_1$. A Bayesian would now use the data according to Bayes' rule to update the prior odds in order to have posterior odds, which tell the researcher the comparative likelihood of the two models after having seen the data. Formally, with Y denoting the observed data:

$$\frac{P(\mathcal{M}_0|Y)}{P(\mathcal{M}_1|Y)} = \frac{P(Y|\mathcal{M}_0)}{P(Y|\mathcal{M}_1)} \cdot \frac{\pi(\mathcal{M}_0)}{\pi(\mathcal{M}_1)}$$

The factor by which the prior odds have to be multiplied is the **Bayes factor**. It is the likelihood of the data under the respective model – technically the model's marginal likelihood $P(Y)$ as described before (see equation (2.1)). The marginal likelihood is the normalising constant that is calculated by integrating over the numerator in the formula for the model's joint posterior distribution, i.e.

$$P(Y|\mathcal{M}_i) = \int \mathcal{L}(\theta_i|Y, \mathcal{M}_i)\pi(\theta_i|\mathcal{M}_i)\ \mathrm{d}\theta_i$$

It is therefore a real number for each model, and the Bayes factor is the ratio between these numbers from two different, competing models. When using Bayes factors for hypothesis testing, the question is how to set up the models in such a way that they represent the hypotheses in question. As with any Bayesian model, this involves both the model likelihood and the prior for the parameters. The modelling step itself provides, as discussed in the previous section, a wide range of possibilities to formulate theories as statistical probability models. Among others, Rouder, Speckman, Sun, Morey, & Iverson (2009) and Wagenmakers (2007) have advocated the use of Bayes factors in psychology by presenting "default Bayes factors" that resemble the traditional significance tests, such as t- or F-tests, most closely.

The competing hypotheses in a traditional independent samples t-test are the null hypothesis $\mathcal{M}_0$ that $\mu_0 = \mu_1$ and the alternative hypothesis $\mathcal{M}_1$ that $\mu_0 \neq \mu_1$.[6] Represented as a linear model, the hypotheses can conveniently be written as nested models

[6] In the context of hypothesis testing, the notation H_0 and H_1 respectively are more common. But as models and hypotheses are interchangeable in the present context, the notation using the letter $\mathcal{M}$ is more consistent.

$$\mathcal{M}_0 : \; y_i = \alpha + \beta x_i + \epsilon_i, \;\; \beta = 0, \epsilon_i \sim \mathcal{N}(0, \sigma)$$
$$\mathcal{M}_1 : \; y_i = \alpha + \beta x_i + \epsilon_i \;\; \epsilon_i \sim \mathcal{N}(0, \sigma)$$

For the Bayes factor, the two models would differ in their prior distribution for β: For the null model $\mathcal{M}_0$ the prior distribution for β would have its full probability mass at $\beta = 0$, i.e. zero is the only possible value for β in this model. Consequently, as no other values for β are possible in this model, the posterior distribution for β in $\mathcal{M}_0$ is identical to the prior. For $\mathcal{M}_1$, on the other hand, a more permissible prior distribution should be chosen. This is the primary challenge when using Bayes factors: The choice of prior has a strong impact on the inferences, as will be shown later. Note, however, that a similar decision must be made when using the Neyman-Pearson paradigm of significance testing: In order to maintain the properties of long-run error rates under the decision rule, the research has to choose an alternative hypothesis to calculate the necessary sample size. This also requires researchers to think carefully about the effect under investigation and their reasonable expectations based on theory and prior experience. While for frequentist testing, a point estimate is selected, the Bayesian tester will need to specify the reasonable distribution of the parameter. Section 2.4.5 will provide an example of how to use prior research to justify the definition of a prior distribution in the alternative model for a Bayes factor. Default Bayes factors (Rouder, Morey, Speckman, & Province, 2012; e.g. Rouder et al., 2009) have been suggested when prior information is scarce. More details on the selection of priors for Bayes factors will be discussed below.

When the two competing models are set up, the calculation of the Bayes factor requires the computation of the models' marginal likelihood. As this quantity involves the integral across the whole posterior distribution, an exact solution is rarely available. The aforementioned MCMC techniques can help, but usually only provide a limited number of samples from the posterior distribution, and using them can make for an unstable estimate of the marginal likelihood (Gamerman & Lopes, 2006). Bridge sampling (Gronau et al., 2017) and importance sampling are some of the approximate methods to calculate marginal likelihoods that will be considered in the context of the Replication Bayes factor below (see section 2.5.1). Software packages such as `BayesFactor` or `bridgesampling` can help to estimate the marginal likelihood for specific models.

The resulting Bayes factor is then the ratio between the models' marginal likelihoods:

$$B_{01} = \frac{P(Y|\mathcal{M}_0)}{P(Y|\mathcal{M}_1)} = \frac{\int \mathcal{L}(\theta_0|Y, \mathcal{M}_0)\pi(\theta_0|\mathcal{M}_0)\, \mathrm{d}\theta_0}{\int \mathcal{L}(\theta_1|Y, \mathcal{M}_1)\pi(\theta_1|\mathcal{M}_1)\, \mathrm{d}\theta_1}$$

As a convention, the ordering in the subscript shows which model goes into the numerator and denominator, respectively. Therefore, $B_{01} = 1/B_{10}$. Other notations are also common such as BF_{10} or K (Jeffreys, 1961). Note also that the models do not need to be nested meaning that θ_0 and θ_1 can be from distinct parameter spaces. In the context of simple hypothesis testing (e.g. Bayesian t-test), the models are simple and nested with $\theta = \theta_0 = \theta_1$. The difference in hypotheses is represented (as in the example above) in the prior.

2.2.2.1 Interpretation of Bayes Factors

The Bayes factor will be a number between 0 and $+\infty$ and can be interpreted in several ways. In contrast to p-values in the Neyman-Pearson framework, it is to be interpreted on a continuous scale and dichotomous interpretations are generally not favoured. Nevertheless, guidelines for verbally describing the results have been suggested. As explained in the introduction of Bayes factors, they are interpreted straightforwardly in the context of updating the ratio between model probabilities:

$$\frac{P(\mathcal{M}_0|Y)}{P(\mathcal{M}_1|Y)} = B_{01} \cdot \frac{\pi(\mathcal{M}_0)}{\pi(\mathcal{M}_1)}$$

On its own, the Bayes factor indicates *how much more likely the data (Y) is under the model M_0 when compared to the model M_1*. Importantly, it is not the posterior probability of the model/hypothesis! In order to learn about the posterior model probabilities, researchers need to make an informed decision on the prior odds. These can then be multiplied with the Bayes factor to reach the posterior odds. Consequently, when both models are equally likely *a priori*, the Bayes factor is equal to the posterior odds – but only in this case. It nevertheless suffices to report the Bayes factor (with the underlying models) as it contains all information necessary for making inferences.

To guide interpretation, Jeffreys (1961) and Kass & Raftery (1995) proposed a scale to interpret Bayes factors:

- $1 < B_{01} < 3$: "not worth more than a bare mention," because the data provide evidence for both models nearly equally.
- $3 < B_{01} < 10$: "substantial" evidence in favour of model M_0.
- $10 < B_{01} < 100$: "strong" evidence in favour of model M_0.
- $B_{01} > 100$: "decisive" evidence in favour of model M_0.

Conversely, these guidelines apply to $B_{10} = 1/B_{01}$ in the other direction. The verbal description helps to put a Bayes factor into perspective, but even with "decisive" evidence for model M_0, the other model can be *a posteriori* more likely when M_0 is very unlikely. When thinking about a hypothesis with a point null, for example, one might consider this null hypothesis to be practically impossible as nothing is ever truly zero (see "Crud factor" above). A Bayes factor of any size in favour of the null hypothesis might then not be enough to be convincing. This underlines the difference between the statistical inference ("What does the data tell?") and the substantial inference ("How convincing is the statistical result for my theory?").

2.2.2.2 Choice of Priors in Bayes Factors

An important part of the Bayes factors is, as described before, the prior distribution for the parameters of interest. Notably, Bayes factors are generally highly dependent on the specification of the models. In particular, through integrating across the whole parameter space in the marginal likelihood, the prior distribution has a strong influence on inferences. This has been a major concern for many researchers who seek a more "objective" way to statistical inference. On the other hand, this sensitivity is also a natural feature of Bayesian hypothesis testing and can be justified on two accounts: First, the priors should ideally be informed by theory. If the theory is vague about an effect size, so should necessarily do the priors for the Bayes factor. Conversely, if the theory is very specific about the effect, this needs to be reflected in the priors as well by reducing the spread of the prior distribution. Changes in the Bayes factor due to reformulation of a hypothesis with a strong prior therefore reflect the considerable change in the theory's prediction. Hence, sensitivity is desirable from this point of view (Vanpaemel, 2010). Secondly, it needs to be emphasised that significance testing is – from a Bayesian perspective – also subjective and sensitive to the specification of the models (Berger & Berry, 1988; Sprenger, 2015).

When choosing the prior for a Bayes factor, theory or knowledge from the literature should guide the formulation. One example for doing so is presented in the next section on null effects. Sometimes, however, prior information is scarce or exploration of the data is the primary goal of the analysis. For such cases, *default Bayes factors* with *default priors* have been proposed. For a Bayesian *t*-test and the underlying linear model, for example, Rouder et al.

(2009) proposed to place a prior on the effect size $\delta = \frac{\mu}{\sigma}$ with $\delta \sim \text{Cauchy}(\lambda, \tau)$. This prior results from using a Normal distribution centred on 0 for the effect size and an inverse chi-square prior for the variance, which is equivalent to directly placing a Cauchy prior on the effect size (Liang, Paulo, Molina, Clyde, & Berger, 2008). See figure 2.3 for a visualisation of the Cauchy distribution compared to a normal distribution. Additionally, a prior for the data's variance in the linear model, σ^2, has to be chosen. Rouder et al. (2009) suggested a non-informative $p(\sigma^2) = \frac{1}{\sigma^2}$. For these two priors, Rouder et al. (2009) use the name *JZS prior* to acknowledge the foundation on the works of Jeffrey, Zellner, and Siow – hence the acronym. This particular choice for a distribution is driven by practical and mathematical considerations: Generally, most effects under investigation are small and an effect size $d = .1$ is more reasonable to expect than an effect size $d = 1.5$, but a normal distribution will quickly reduce the probability allocated to larger effect sizes. Most default priors, therefore, use Cauchy distributions that have heavier tails and allow for larger/smaller values.

One might wonder why a distributional form for the prior has to be chosen at all – a uniform distribution or a normal distribution with infinite variance on the effect size (or any other parameter) would allow ignoring the question what effect sizes are considered reasonable. Consequently, this would, however, also mean that effect sizes of any size are equally likely. An implication that should be disputed. Moreover, this is the basis for Lindley's paradox (Lindley, 1957), where frequentist and Bayesian analysis provide contradicting results. While named a paradox, it is, in fact, a consequence of two different statistical models used: One in which any effect size is considered equally likely (frequentist approach) and one in which the likelihood of effect sizes is constrained through the prior (Bayesian approach).

The "default Bayes factors" as described for *t*-tests by Rouder et al. (2009) and also introduced for other common testing scenarios in psychology (Nuijten, 2012; Rouder & Morey, 2012; Rouder et al., 2012) are easy-to-use for researchers not yet accustomed to Bayesian analysis. They are designed to be closely related to their frequentist counterparts. In practice, default Bayes factors and significance tests yield similar inferences most of the time, but the statistical interpretation of the result differs: While the significance test makes a statement about error probabilities, the Bayes factor on the other hand is a statement in the context of Bayes' rule (see above) and is often interpreted in terms of "statistical evidence." The Bayes factor can be used to calculate posterior odds for competing hypotheses. The significance test does not allow for such interpretations. There has been work on error probabilities when using default Bayes factors (Gu, Hoijtink, & Mulder, 2016), but this approach has been discussed critically (Morey, Wagenmakers, & Rouder, 2016).

A major concern when using default Bayes factors is that it does not properly specify the alternative hypothesis for the hypotheses at hand: It is similar to using significance tests without setting the sample size *a priori* through a power analysis. As with frequentist analysis, a researcher should carefully specify and justify the hypotheses under investigation so the inferences are actually relevant to the substantial question asked (cf. figure 1.1). When using the framework of statistical modeling to carefully represent hypotheses as statistical models, the Bayes factor can be used as a hypothesis test to compare these models. The large benefit compared to "mindlessly applying significance testing" does not come from the Bayes factor, but from the additional thought of using models to transfer substantial questions into relationships between entities of probability. The Bayes factor, though, is one way to provide a consistent framework in the context of Bayesian statistics. If a researcher is more interested in error probabilities, the frequentist framework is a more appropriate one and can similarly make use of the additional clarity of statistical models.

While the best alternative to a default prior is to carefully choose a distribution with adequate parameters based on previous literature or theories, another option is to adjust the scale parameter of the Cauchy distribution, which changes the spread of the distribution, as can be seen in figure 2.4.[7] While the choice for the scale parameter should also be well justified, a

[7] Note that the Cauchy distribution does have neither mean nor variance.

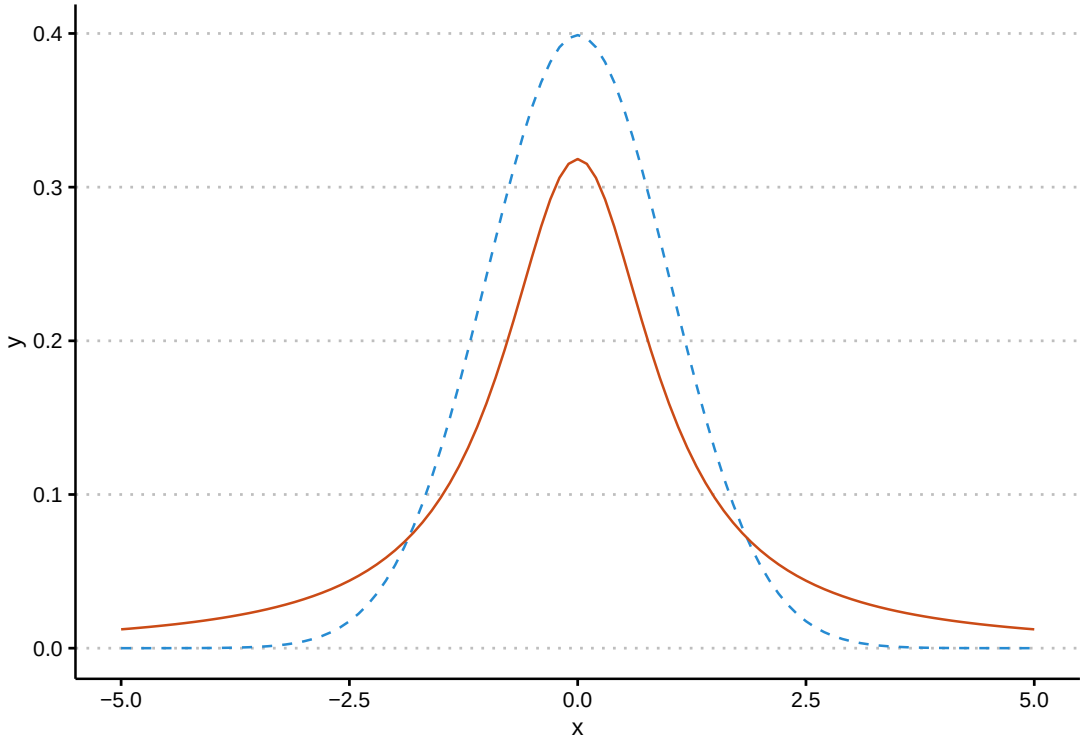


Figure 2.3: Plots of a normal (blue dashed line) and a Cauchy distribution (red solid line). Normal distribution has mean 0 and variance 1, Cauchy distribution has location 0 and scale 1. The Cauchy distribution has heavier tails, i.e. it has more probability mass for larger/smaller values compared to the normal distribution, that has its mass more centered around the mean.

sensitivity analysis can also be used to verify the inferences based on a default Bayes factor.

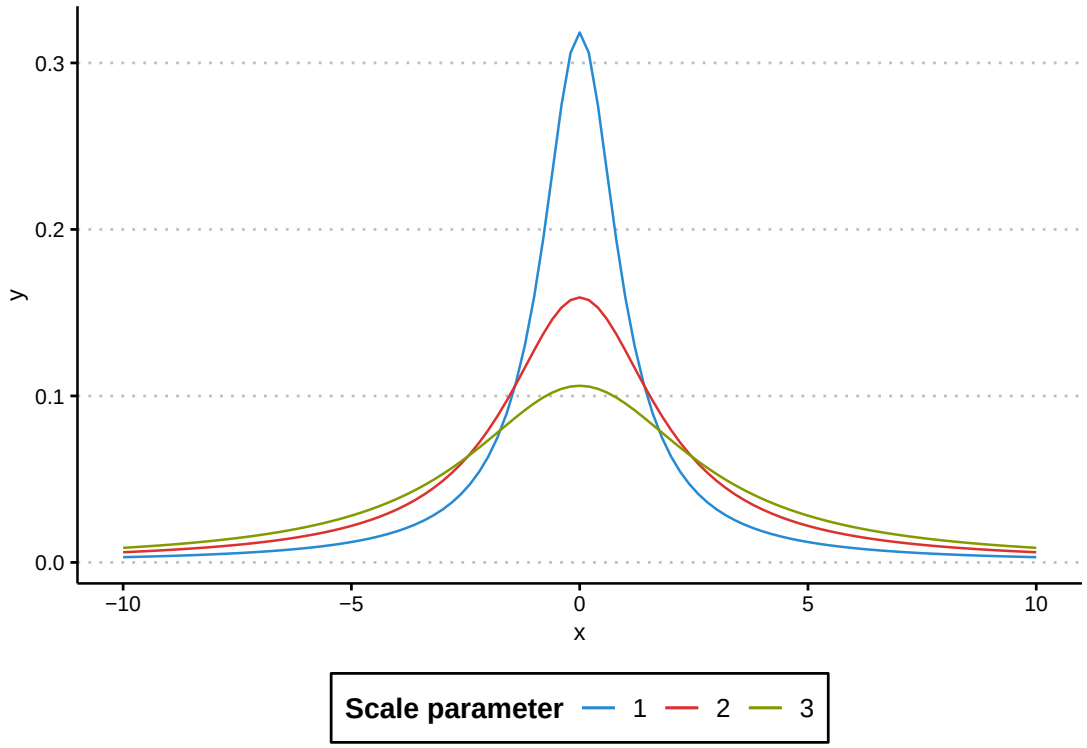


Figure 2.4: Plot of the Cauchy distribution with different scale parameters. While the distribution has neither mean nor variance, the scale parameter adjusts the spread of the distribution.

2.2.2.3 Sensitivity Analyses

As discussed before, Bayes factors are sensitive to the formulation of the priors, which is a concern for some. Besides a good justification for the selection of the prior distribution in the two compared models from theory or previous research, there are two practical options available:

1. One can perform a sensitivity analysis by calculating Bayes factors for a range of plausible (and implausible) priors (Sinharay & Stern, 2002), in default Bayes factors, for example, by plotting the resulting Bayes factor for different values of the scale parameter r (see figure 2.5). For Bayes factors involving normal distributions, one might consider different values of σ and/or μ. For sequential designs, a sensitivity analysis might show how the Bayes factor changes over accumulating data (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017).

2. If there is no theory or strong prior information about the specification of priors for the hypotheses, a fully Bayesian approach (see section 2.2.1) might be the more sensible alternative. With enough data and weakly regularising priors, the data will quickly overwhelm the prior and the models provide a regularised estimate for the parameters of interest. This shifts the analysis to an estimation framework, but the results from the Bayesian model can then inform further testing with Bayes factors or more appropriate

model selection tools such as *LOO* (Leave-One-Out cross-validation) or *WAIC* (widely applicable information criterion).

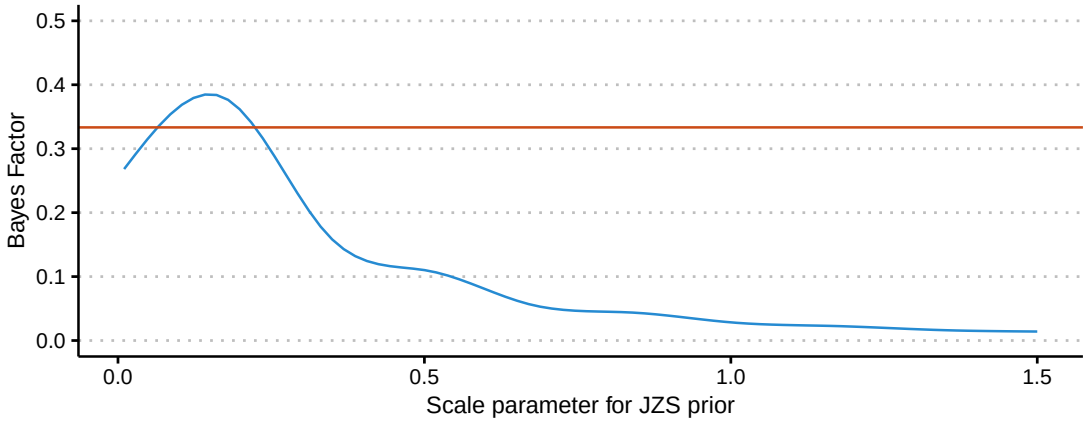


Figure 2.5: Example sensitivity analysis for Bayes factors. The plot shows the resulting Bayes factors in blue for scale parameter values between 0 and 1.5 (smoothed line for illustrating purposes). The red line denotes Bayes factors of $\frac{1}{3}$, indicating the conventional boundary for weak evidence. The analysis shows that the data consistently show evidence in favour of the denominator model, some of which might not be considered strong, however. Figure reproduced with data from the sensitivity analysis performed in Harms, Jackel, & Montag (2017).

The second option is another reminder that the choice of statistical analysis should be guided by the analytical goal. Bayes factors are a hypothesis testing, or model comparison, tool. If no theories or hypotheses can be formulated because the goal is to explore the data, other approaches might be preferred. Building a model on a first sample and testing the conclusions in a second or validation sample is one way to combine exploratory and confirmatory research. This point will be made again in the general discussion at the end of this thesis.

2.2.3 Other Methods and Discussion

Bayes factors and information criteria are both methods for model selection. When multiple candidate models are available, they guide the decision from which model inferences should be made. This approach gives a sense of certainty: The researcher has selected the best model, and inferences are based on the best model available and should be "better" than from the worse performing model. On the one hand, researchers should always be aware that their model is very likely not to be the "correct" model. Especially in simple hypothesis testing situations where true null effects are considered plausible models, researchers need to remember Box's dictum (Box, 1976, 1979):

> All models are wrong but some are useful

Some statisticians criticise the use of Bayes factors, for example because of their reliance on priors and their interpretation in the context of Bayesian updating. Gelman & Yao (2020) for example argue that Bayes factors focus too much the parameterisation of the model and the

priors. Small changes in the prior can have strong influence on the Bayes factor but little relevance to the parameter's posterior distribution, as has been discussed above.

Besides selecting for a single "useful" model, statistical modeling offers other ways to deal with multiple candidate models. Two such approaches are *model averaging* and *stacking*, which are summarised as methods for "ensemble learning" in machine learning (Hastie, Tibshirani, & Friedman, 2009, p. 605ff). Bayesian model averaging, for example, allows to consider all possible models at once and provides posterior probabilities for each model (Hoeting, Madigan, Raftery, & Volinsky, 1999; Raftery, Madigan, & Hoeting, 1997). While it also allows to make predictions across all models with each model weighted by their posterior probability, it is often used as an alternative to frequentist step-wise regression for variable selection (Wang, Zhang, & Bakhai, 2004). See Magnusdottir et al. (2019) for an example.

2.3 Application of Statistical Frameworks

Bayesian statistics, as introduced briefly here using Bayesian models and Bayes factors, offers another approach to analysing data in contrast to the predominant framework of statistical significance testing. They both are a useful addition to a researcher's statistical toolbox. Bayesian statistics has more technicalities and offers opportunities beyond the examples presented here. Hierarchical models providing partial pooling for estimates, structural equation models using priors for regularisation of path coefficients, Bayesian belief networks, or Gaussian process modelling are just four advanced tools that can be used in the context of Bayesian statistical framework. The availability of software tools such as `Stan`, `brms`, or `blavaan` have tremendously helped the increased use of Bayesian statistics within and outside science.

Fisher (1950) decidedly rejected the idea of Bayesian statistics (Fisher, 1950, p. 9):

> For many years, extending over a century and a half, attempts were made to extend the domain of the idea of probability to the deduction of inferences respecting populations from assumptions (or observations) respecting samples. Such inferences are usually distinguished under the heading of Inverse Probability, and have at times gained wide acceptance. This is not the place to enter into the subtleties of a prolonged controversy; it will be sufficient in this general outline of the scope of Statistical Science to reaffirm my personal conviction, which I have sustained elsewhere, that the theory of inverse probability is founded upon an error, and must be wholly rejected. Inferences respecting populations, from which known samples have been drawn, cannot by this method be expressed in terms of probability, except in the trivial case when the population is itself a sample of a superpopulation the specification of which is known with accuracy.

While differences between frequentist and Bayesian statistics are often hotly debated among statisticians and philosophers, both offer statistical tools that applied researchers can make use of. Both frameworks provide rules and guidelines on how the tools are to be used and their results are to be interpreted. As long as these rules are respected, a researcher does not need to make a final decision for or against one framework over the other. When analysing results both frameworks can be used side-by-side. It is important, however, that reasoning based on the frameworks is consistent. If one runs a significance test, for example, they need to reason about the control of error rates and how the outcome will be subject to the error rates resulting from power analysis and significance levels. They cannot interpret the p-value as statistical evidence on its own. On the other hand, if one uses a Bayes factor to quantify the statistical evidence in favour of a model of interest compared to another model, no statement about long-run error rates can be made. Moreover, both statistical interpretations are connected to epistemological questions (see chapter 1.2) that should guide interpretations. Researchers

should aim to be consistent in their use of statistical and epistemological frameworks when reasoning about data and their results.

2.4 Article I: Evaluating Null Effects

The following section is a summary of this thesis' first article:

- **Harms, C.**, & Lakens, D. (2018). Making Null Effects Informative: Statistical Techniques and Inferential Frameworks. Journal of Clinical and Translational Research, 3(S2), 7. https://doi.org/10.18053/jctres.03.2017S2.007

The content and several quotes are taken directly from this paper, which is also included in the appendix of this thesis.

Summary: While the debate around the replication crisis has brought increased attention to the question of "null effects," there remain several statistical misconceptions and problematic practices. Our paper aims to guide readers along different methods to make null effects more informative than a simple "non-significant" result. Using simulated example data, we explain equivalence testing, interpretation of Bayesian posterior distributions (ROPE procedure), and Bayes factors with an emphasis on how the conclusion of the results can differ – and how researchers can make use of these tools in order to draw better and more correct inferences from non-significant results. All approaches to investigating null effects need to be informed by previous theory or other external information if only the considerations of the researcher. In particular, we show how equivalence testing and the ROPE procedure on the posterior distribution are very similar procedures leading to similar conclusions and how the Bayes factor can be set up to compare different competing hypotheses.

As has been mentioned before, there is a common fallacy of interpreting non-significant findings as proof or indication that there truly is no effect. It is important to note that both from a Fisherian and a Neyman-Pearsonian perspective this interpretation is an invalid conclusion. A non-significant finding merely provides "absence of evidence," that is a lack of evidence in either direction. That is to say that a non-significant p-value does *not* provide "evidence of absence." This distinction is crucial (Altman & Bland, 1995) but often neglected, which might have contributed to the prevalence of "publication bias" (Theodore D. Sterling, 1959; T. D. Sterling et al., 1995). It certainly has created the perception that non-significant results are not informative among authors, reviewers, and editors (Neuliep & Crandall, 1990, 1993).

Aside from non-significant results in otherwise traditional testing scenarios, there are cases in which a researcher sets out to find null effects specifically. Two common scenarios where this could be the case are (Harms & Lakens, 2018, pp. 382–383):

> 1. [C]linical research, [where] it might be important to know if a cheaper or shorter treatment works just as well as a more expensive or longer treatment. Studies designed to answer such questions investigate non-inferiority (e.g., people in one group do not score worse than people in another group) or the statistical equivalence of different treatments (e.g., people in one group score the same as people in another group).
> 2. [Designing] a study that has the goal to demonstrate the absence of an effect because we aim to falsify theoretical predictions about the presence of a difference.

These two cases have in common that the way significance tests are usually applied cannot be used there. In particular, the null hypothesis is what the researcher wants to accept. There are, however, different ways of using the data to make more informed inference using both frequentist tests and Bayesian methods. In Harms & Lakens (2018), we have used simulated data as a worked example to show how the different approaches can be used. In this section, the example is re-used and summarised with references to the above sections introducing equivalence tests, Bayesian models, and Bayes factors.

2.4.1 Example Study

Imagine a study investigating whether mindfulness meditation (Brown & Ryan, 2003) improves pain intensity in patients with lower back pain (LBP). Since LBP is an increasingly common ailment in office workers, it has received a lot of attention in medical research and preventive medicine. Previous studies have investigated several treatments and interventions, both physical and psychological. For our imaginary study, in particular, we will recruit patients and randomly assign them either to an eight-week course in mindfulness meditation (treatment group) or to a waiting list condition (passive control group). The intensity of pain symptoms is measured through a 100mm Visual Analogue Scale (VAS) as is common practice in pain research. For each patient, the difference between pain intensity at the beginning and the end of the eight-week study phase is calculated. The treatment effect is the difference between the treatment and control group.[8] The traditional statistical approach for such an experiment is an independent t-test.

An *a priori* power analysis is required for any frequentist test and at best also stored as a pre-registration. Based on previous literature and common statistical practice, we choose an α level of 0.01 and aim for 90% power to detect an effect of $d = 0.3$. This constitutes the smallest effect we care about and that might have arisen in discussions with experts in the field. Based on these parameters, the required sample size for the study is exactly $n = 332$ per group. Figure 2.6 visualises the imaginary data for the two groups.[9]

2.4.2 Traditional Significance Testing

Performing a Welch's two-samples t-test, as would be the common practice, leads to a non-significant result ($t(661.63) = -1.64$, $p = .101$). The p-value indicates that the population difference estimated from the data is not extreme enough to warrant rejecting the null hypothesis of no effect (i.e. no difference between the groups). To repeat the point made before: This is "absence of evidence" and does not indicate that there is truly no difference between the two groups. In order to make this result more informative, we need to resort to other statistical tools.

2.4.3 Equivalence Testing

In the frequentist framework, one way to further investigate non-significant results is to ask whether the estimated effect size is consistent with an effect smaller than the smallest effect size of interest (SESOI). In many applied settings, researchers can give boundaries for effect sizes that are negligible for practical purposes. Equivalence testing allows to statistically test whether an effect size lies within some lower and upper bounds. Crucially, these equivalence bounds Δ_U (upper bound) and Δ_L (lower bound) have to be defined beforehand, that is, they should be part of a pre-registered study or analysis protocol.

[8] As also noted in the footnote in Harms & Lakens (2018, p. 383), the study design is best captured in a hierarchical model, but for the sake of this example, we will stick to simple statistical analysis.

[9] See the supplementary material to Harms & Lakens (2018) for the R code to generate the simulated data.

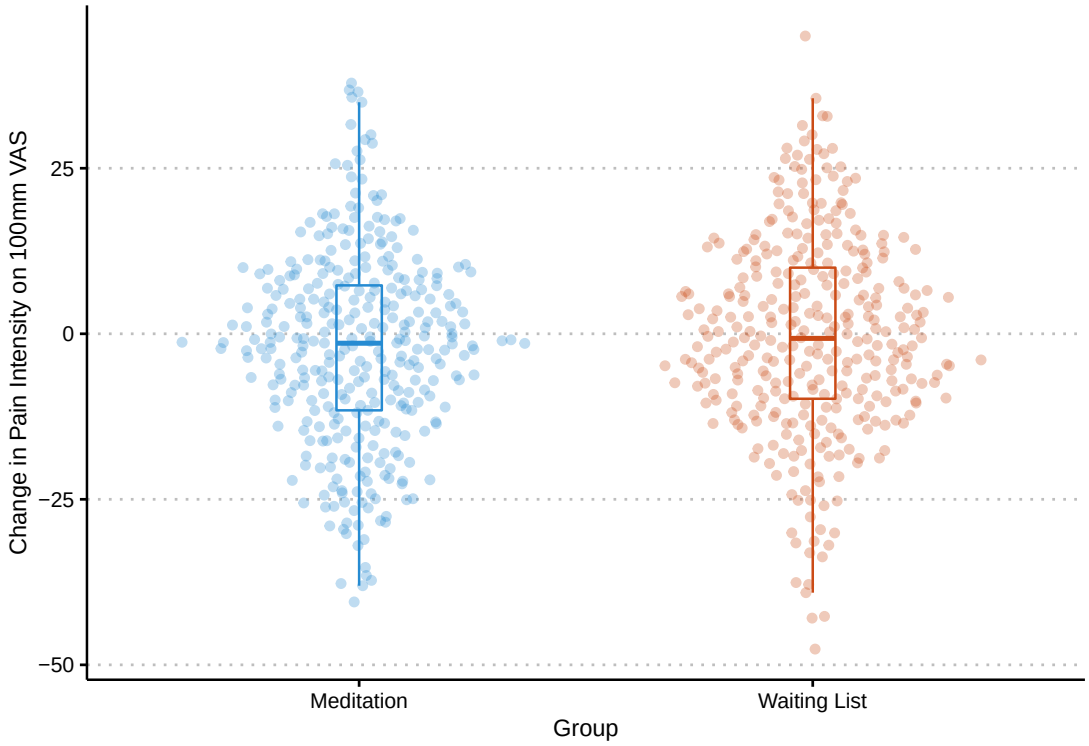


Figure 2.6: Plot of raw data for imaginary study. Each dot is a single case. Y-axis is dependent variable, i.e. change in pain intensity after eight weeks of either meditation class or being on waiting list. *Note:* Figure reproduced from Harms & Lakens (2018).

In medical research, the SESOI is often referred to as the "minimal clinically interesting difference" (MCID). For pain research, for example, a difference of 9mm on the 100mm VAS is considered an MCID because it is (on average) the difference that correlates to a subjective change in pain intensity perceived by patients (Wandel et al., 2010). For the equivalence test, two one-sided tests are performed using a significance level of 2α, or – equivalently – the $(1 - 2\alpha)\%$ confidence interval (CI) is compared to the equivalence bounds (Rogers, Howard, & Vessey, 1993).

- If the $(1 - 2\alpha)\%$ CI lies completely between the equivalence bounds (significant equivalence test), the estimated effect size can be considered equivalent to zero (or any other value around which the equivalence bounds were defined).
- If the $(1 - 2\alpha)\%$ CI lies completely outside the equivalence bounds, the null hypothesis of no equivalence could not be rejected. In most cases, however, this also means that the traditional significance test is significant.
- If the $(1 - 2\alpha)\%$ CI includes one or both equivalence bounds, the results are inconclusive. This happens especially when the estimated effect size is close to an equivalence bound.

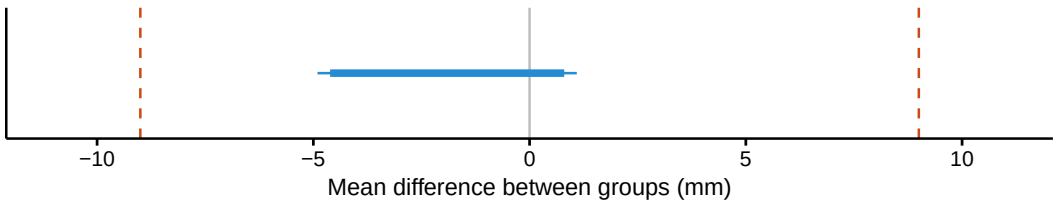


Figure 2.7: Visual representation of the equivalence test result. The horizontal axis is the mean difference between groups in millimetres on the Visual Analogue Scale. Red dashed vertical lines represent lower and upper equivalence bounds of ±9mm. The blue bar is the confidence interval for the traditional significance test (thick bar) and the equivalence test (narrow bar). Both intervals are well within the equivalence bounds. *Note:* Figure reproduced from Harms & Lakens (2018).

For the imaginary study in our example, testing for equivalence between $\Delta_U = +9$mm and $\Delta_L = -9$mm yields a significant result ($t_1(661.625) = 6.113$, $p < .001$; $t_2(661.625) = -9.397$, $p < .001$), indicating that the estimated effect size is equivalent to zero if we consider differences smaller than 9mm as equivalent (Figure 2.7). Following the Neyman-Pearson framework to statistical inferences, we know that we can act as if the difference is truly equivalent to zero – again, given the specified equivalence bounds – without being wrong too often in the long run.

More elaborate introductions and treatments of equivalence testing are available in Lakens (2017), Lakens, Scheel, & Isager (2018), or Meyners (2012).

2.4.4 Bayesian Model (ROPE Procedure)

Based on the principles for Bayesian models outlined above, one can also set up a model that closely resembles a traditional *t*-test but allows a Bayesian interpretation of the results. Kruschke (2013) introduced such a model, and Kruschke & Liddell (2017) formulated the ROPE (**R**egion **O**f **P**ractical **E**equivalence). While not considered a hypothesis test in the narrow sense, it is a decision rule for making inferences based on the posterior distribution of an effect size or of another parameter of interest. Kruschke (2018) defines the decision rule as:

> If the 95% HDI of the [parameter's posterior distribution] falls completely outside the ROPE than reject the null value, because the 95% most credible values of the parameter are all not practically equivalent to the null value. If the 95% HDI of the [parameter's posterior distribution] falls completely inside the ROPE then "accept" the null value for practical purposes, because the 95% most credible values of the parameter are practically equivalent to the null value. Otherwise remain undecided.

In practice, the ROPE procedure resembles closely the decisions made in equivalence testing above. The fundamental difference is the interpretation of the result along with the basis for this interpretation. While an equivalence test aims to control error rates in a frequentist framework, no such statement can be made (without additional assumptions) in the Bayesian context of the ROPE procedure. The ROPE procedure relies on the posterior distribution of a Bayesian model, where probability is distributed according to Bayes' rule. This is particularly important when looking at the width of the interval: While the 95% interval reminds one about the conventional 5% level of significance, the choice is arbitrary and does not have the same meaning. In Bayesian statistics, decision rules such as the ROPE procedure are merely utilities for researchers to interpret and make inferences about the posterior distribution. Many different approaches can be taken here: For example, instead of investigating the HDI and comparing it to the ROPE, one could also interpret the probability mass contained within the ROPE interval (Greenwald, 1975; an approach taken by Liao, Midya, & Berg, 2019). One can also establish a relationship between ROPE and Bayes factors by setting up models that represent the ROPE's decision rule (Liao, Midya, & Berg, 2020).

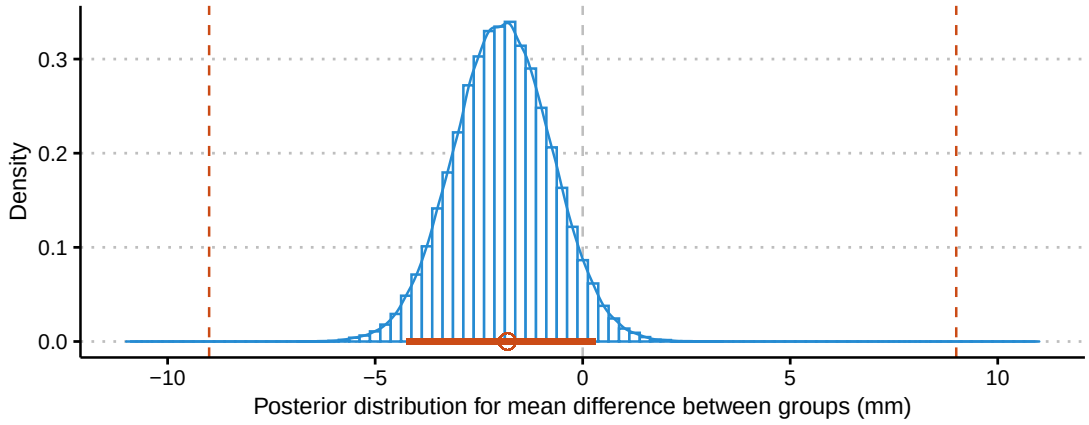


Figure 2.8: Histogram with superimposed density estimate for the posterior distribution of the mean difference in pain intensity rating change between groups. Posterior distribution is based on Bayesian *t*-test model (Kruschke, 2013). The thick bar is 95% Highest Density Interval, indicating the 95% most credible values for the mean difference. Circle in the interval is the most credible point estimate, the *Maximum A Posteriori* (MAP) estimate. *Note:* Figure is reproduced from Harms & Lakens (2018).

For the example study, the pre-defined model by Kruschke (2013) can be used to investigate the posterior probability for the difference between groups in mean change on the 100mm VAS. With the same reasoning as before, we choose a ROPE interval between -9 and $+9$. The resulting 95% HDI ($[-4.244; 0.317]$) lies well within these bounds (see Figure 2.8). We, therefore, declare a difference of exactly zero to be accepted for practical purposes. The ROPE

explicitly states that we do not make any inference about any other value within the ROPE interval.

Many advocates of Bayesian statistics argue that a major benefit of using posterior distributions for inferences is that it allows for nuanced and quantitative judgements and that it does not require researchers to make dichotomous decisions. While the ROPE procedure is designed to specifically make such a binary decision, it is important to see that the basis for the decision, the posterior distribution, allows for other decision rules – or none at all.

2.4.5 Bayes Factor

In the context of Bayesian statistics, hypothesis testing refers to the use of Bayes factors for selection between competing models (see section 2.2.2 above). As discussed before, when using Bayes factors for hypothesis testing, it is important to justify the use of prior distributions as Bayes factors are sensitive to them even with large samples. A range of "default Bayes factors" have been suggested for common scenarios in psychological research (Rouder & Morey, 2012; e.g. Rouder et al., 2012; Wetzels & Wagenmakers, 2012). For our present example, we did not choose a default prior but instead used past literature to inform our analysis. In a meta-analysis on psychological interventions for patients with chronic lower back pain, Hoffman, Papas, Chatkoff, & Kerns (2007) found an estimated effect size of $d = 0.62$ (95% CI: $[0.25; 0.98]$) for cognitive-behavioural therapy (CBT) when compared against a waiting list condition. As the outcome was also pain intensity on a 100mm VAS after treatment, we concluded that a meditation intervention should have an effect of similar size.

For the Bayes factor, we specify an alternative model with a normal prior distribution with $\mu = 0.62$ and $\sigma = 0.372449$ (calculated from the confidence interval). The null model is a point null hypothesis, so the prior has all its prior mass at 0. The Bayes factor, therefore, compares the hypothesis of absolutely no effect against the alternative that the effect is approximately in line with the meta-analysis by Hoffman et al. (2007). Both hypotheses might be wrong and both hypotheses can be formulated differently in mathematical terms, but the Bayes factor will allow us to conclude which of the models is more in line with the data at hand.

Calculating a Bayes factor for a t-test model (Gronau, Ly, & Wagenmakers, 2019) yields $\text{BF}_{01} = 2.945$, concluding that the data is 2.945 times more in favour of the null model compared to the informed alternative (see figure 2.9 for a visual representation of the Savage-Dickey-ratio, Wagenmakers, Lodewyckx, Kuriyal, & Grasman (2010)). When using thresholds to interpret the Bayes factor, this result can be considered borderline inconclusive. More data is required for a more definitive result (Schönbrodt et al., 2017). This result seems to contradict the results found previously with equivalence tests and the ROPE procedure, but the Bayes factor is asking a very different question: While equivalence test and ROPE were asking whether the effect is negligibly small, the Bayes factor was set up to investigate whether the data is more in line with either a true null effect or an effect close to previous effects of psychological interventions. When interpreting statistical results, it is crucial to consider the statistical and substantive questions asked by the chosen methods (Hand, 1994).

2.4.6 Summary

The example presented here from Harms & Lakens (2018) shows four different ways to investigate the outcome of a simple two-group interventional study. While the traditional t-test, which is commonly used in such scenarios results in a non-significant p-value, the other three approaches allow for a more nuanced interpretation. All proposed methods rely on the same underlying linear model but differ in assumptions and interpretation. While the frequentist equivalence testing aims to provide long-term error control in the Neyman-Pearson framework, the Bayesian methods provide posterior probabilities and use decision rules based on

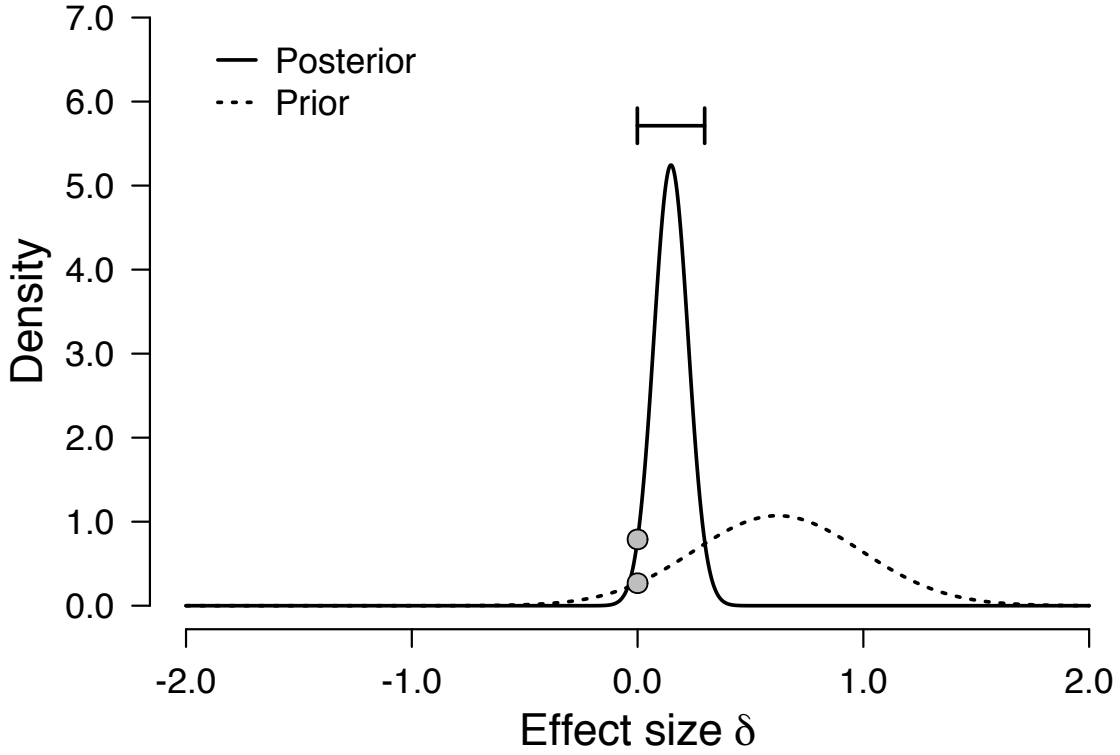


Figure 2.9: Visual representation of the Savage-Dickey-ratio (Wagenmakers et al., 2010). The Bayes factor can be understood as the ratio between posterior and prior at $\delta = 0$. *Note:* Figure is reproduced from Harms & Lakens (2018).

these probabilities (ROPE procedure) or use statistical evidence in terms of a Bayes factor to select between two hypotheses (Bayes factors). While there is some work on the error probabilities of Bayes factors (Gu et al., 2016), frequentist and Bayesian interpretations differ on a fundamental philosophical level. This, however, does not mean that a researcher cannot use multiple statistical approaches to analyse data to gain a nuanced interpretation of their results. Researchers must interpret statistical results in the context of the respective framework and employ a consistent epistemological framework when integrating different results (Fidler et al., 2018).

Comparing the results of the three analyses shows that equivalence testing and ROPE procedure lead to similar conclusions: For practical purposes, the study shows an effect size that is too small to be considered meaningful. Based on the results, one might consider mindfulness meditation not a suitable remedy for back pain. The Bayes factor, on the other hand, is more conservative by yielding an inconclusive result. This might seem contradictory. It is important to consider that the methods all ask different statistical questions and involve different model structures. For the Bayes factor, in particular, a very specific alternative was used and the result is only a statement about this specific comparison. Formulating different alternative hypotheses might lead to different results. This is true for frequentists as well: Would we have established a different alternative hypothesis for either the traditional or the equivalence test, or would have chosen different equivalence bounds, the results certainly will turn out different. This highlights the need to think about the assumptions employed when running statistical tests and how *a priori* decisions should be pre-registered. Lastly, the example shows that researchers can use both frequentist and Bayesian analyses side-by-side, compare their results and provide an interpretation resting on different assumptions.

A similar perspective can be taken when running replication studies. The next section will go into more detail on how to statistically evaluate replication studies. Null effects are a possible outcome for replication studies as well, so the methods discussed so far will again be referred to in the following section.

2.5 Evaluating Replication Studies

Conducting replication studies is an important part of empirical investigations in science. In section 3, replication studies will be discussed in more detail, especially considering the taxonomy of different kinds of replication studies. In any study replicating previous results, however, the statistical analysis for comparison with an original investigation is central, which is why this will be discussed below. There exist several approaches to statistically compare studies exist and it is generally recommended to use multiple approaches as they raise different statistical questions (S. F. Anderson & Maxwell, 2016; Asendorpf et al., 2013a; Brandt et al., 2014; Maxwell, Lau, & Howard, 2015). The most relevant approaches will be introduced briefly in this section and expanded by the introduction of a Bayes factor to investigate replication outcomes.

The most intuitive approach would be to compare the pattern of significant findings: A replication is considered successful if the same focal hypothesis test turns out to be statistically significant in both the original and the replication study (called "vote-counting"). This, however, poses several problems: First, a non-significant finding in the replication study must not be interpreted as "evidence of absence" (see above). Second, when the original study is significant, but the replication is not, this difference might not be significant in itself (Gelman & Stern, 2006). Third, comparing p-values completely ignores estimates of effect sizes. Reported effect sizes in original studies are commonly inflated due to publication bias, possibly in combination with measurement error (Ioannidis, 2008; Loken & Gelman, 2017). A replication study might yield a significant result, but – due to high power as replication studies tend to recruit more participants – the effect size might be much smaller than in the original study.

Depending on the original finding, more careful investigation is required to decide whether the replication can be deemed "successful" or not. Comparing the significance of results can be used as a first indication of the replication results, but should be followed by other techniques which take into account the effect size estimate and the associated uncertainty more strongly.

Such a technique was proposed by Simonsohn (2015); it is often referred to as "small telescopes" approach based on the paper's title. The starting point for the evaluation of the replication study is an effect for which the original study had 33% power to detect ($d_{33\%}$). While arbitrary, it can easily be considered a "small effect" in reference to the original study's primary effect size estimate. The effect size estimate from the replication study and the associated 95% confidence interval is then compared to the small effect $d_{33\%}$ of the original study. If the replication's effect size is smaller than $d_{33\%}$ and the confidence interval does not include $d_{33\%}$, the replication seems not to be in line with the original study's effect size or a smaller effect, even if the replication yields a significant result. On the other hand, a replication might be considered in line with a small effect if the effect size is smaller than $d_{33\%}$, but still includes the value. Replications with effect sizes larger than $d_{33\%}$, possibly significant, might then indicate a successful replication, i.e. a result that is very much in line with at least a small effect.

Simonsohn's "small telescopes" is essentially identical to running equivalence tests where the equivalence bounds are set to $d_{33\%}$, except that equivalence tests provide a more formal framework to interpret the result (Lakens, 2017). Furthermore, if there are conventions about what constitutes a lower bound for a relevant change in the outcome in question (such is sometimes the case in clinical areas, see *MCID* above), a replication researcher should also test the difference between the replication and the original study for equivalence in this way.

To consider the uncertainty of effect size estimates, confidence intervals have been advocated to be the mode of statistical inference in general, most strongly by Geoff Cumming (Cumming, 2012, 2014; Cumming & Finch, 2001). In the context of replications, LeBel, Vanpaemel, Cheung, & Campbell (2018) proposed a formal framework to compare the 95% confidence interval of the replication with the point estimate for the effect size in the original study. He uses the following terminology to describe the outcome:

- **Consistent vs. Inconsistent:** If the replication study's 95% confidence interval covers the effect size estimate of the original study, the replication is considered *consistent* with the original study, and *inconsistent* otherwise. This can be further qualified by describing the replication's estimate in relation to the original's with *smaller*, *larger*, or *opposite*.
- **Signal vs. No Signal:** Essentially the result of the significance test of the replication study. If it is significantly different from zero, the replication study is considered a *signal*, and *no signal* otherwise.

The two dimensions should be considered independently, thereby creating four different scenarios (see figure 2.10).

The approaches introduced briefly so far all stem from the frequentist perspective and rely on significance tests or frequentist confidence intervals. Bayesian approaches include, among others, meta-analytic methods by aggregating several studies in a multilevel model (Marsman et al., 2017) as well as Bayes factors. For the latter, different ways to set up a Bayesian hypothesis test can be imagined. A researcher could, for example, use Bayes factors to test the same hypotheses as a frequentist would to compare studies as described in the previous paragraph. Verhagen & Wagenmakers (2014) provided an overview of three different Bayes factors particularly suited to investigate replications:

1. Default Bayes factor for the replication study alone,
2. Bayes factor to test for equality of effect sizes, and

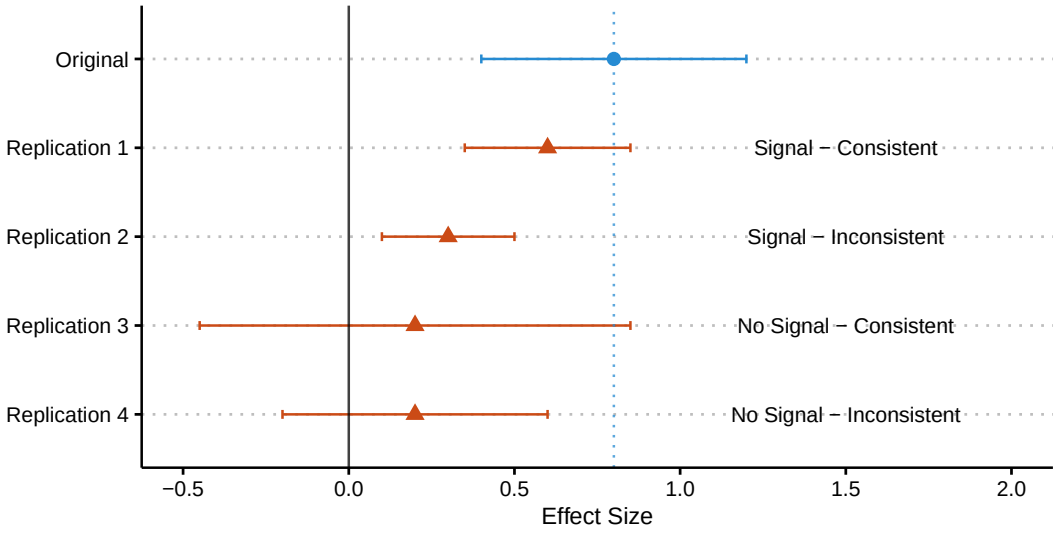


Figure 2.10: Different outcomes on the two dimensions proposed by LeBel et al. (2018). Estimates for effect sizes from the replication studies with their respective 95% confidence interval are compared to the original study's point estimate (blue, dotted line). If the replication's 95% confidence interval includes the original study's effect size, the replication is considered *consistent*, or *inconsistent* otherwise. When the replication is significantly different from zero, the replication contains a *signal*, when it is non-significant it is considered *no signal*. *Note:* This figure is reproduced based on figure 1 in LeBel et al. (2018).

3. Bayes factor for fixed-effect meta-analyses, which combines the data from original and replication study.

They also introduced a fourth option, namely the *Replication Bayes factor* (RBF), which is covered in more detail in the next section. As highlighted in the description of Bayes factors in general above (see chapter 2.2.2), Bayes factors compare two models which need to be specified properly. The model specification represents a particular statistical question under investigation. The different Bayes factors – as well as the different significance tests explained above! – therefore ask different questions: While the default Bayes factor merely tests the presence of an effect in the replication against the hypothesis of a nil effect (like a significance test in the replication study), the Bayes factor for equality of effect sizes tests the hypothesis of equal effect sizes in original and replication study against the hypothesis of different effect sizes. It is crucial to be explicit about the hypotheses being tested.

Both approaches, i.e. LeBel et al. (2018)'s two dimensions for replication outcomes and a Bayesian perspective on the statistical evidence, offer more nuanced perspectives on the outcome of a replication: In many cases, replications are considered to be either *successful* or *unsuccessful*. This is also the perspective some authors have taken when discussing the results of the *Reproducibility Project: Psychology*. This kind of thinking and labelling replications as "failed" easily suggests flaws with the replication study: That it was not performed properly or that it was not informative if it failed. However, replication studies with conflicting results or even consistent null results can be highly informative about the effect under investigation. In many cases replication studies are even *more* informative as they are performed under stricter rules, with more participants, and by researchers not as invested in the effect as an original researcher team might have been. This does not mean that each and any replication is valuable: There certainly are replication studies with low power, bad study design, little rigour, and not according to best practices for conducting a replication study. Such

replication studies are not informative and can be considered a failed attempt at replicating an original study. This, however, should be a judgement irrespective of the replication's outcome and solely be based on the study's design and setup. Considering the outcome of a properly planned and conducted replication study, using non-dichotomous descriptions on different dimensions (LeBel et al., 2018) or being carried out in more quantitative ways through Bayesian statistics (Marsman et al., 2017; Verhagen & Wagenmakers, 2014), allows for more nuanced discussions and can provide more valuable insights on the original effect and future directions.

Verhagen & Wagenmakers (2014)'s Replication Bayes factor is one way to extend common thinking by introducing a continuous measure of statistical evidence in form of a Bayes factor. As will be explained, the underlying models in the Bayes factor are one way to set up hypotheses about replication outcomes. In order to get a more holistic picture of a replication, several different statistical questions should be asked and investigated using proper tools. Moreover, a single replication study should rarely be the only deciding data point used to determine a theory's fate: As much as a single original study is not enough to establish a general psychological law, a single replication study is often not enough to counter a broad body of research. Evaluating the outcome of a replication study goes beyond a single statistical test: Researchers investigating an effect will need to consider both statistical as well as methodological aspects along with more qualitative information. Whether an original or a replication study was pre-registered, susceptible to publication or confirmation bias, had a sample diverse enough to provide some form of generality, et cetera, are information that need to be taken into account along with the outcome of a statistical procedure.

2.5.1 Article II: Replication Bayes Factors

The following section is a summary of this thesis' second article:

- **Harms, C.** (2018). A Bayes Factor for Replications of ANOVA Results. The American Statistician. https://doi.org/10.1080/00031305.2018.1518787

The content and several quotes are taken directly from this paper, which is also included in the appendix of this thesis.

Summary: The analysis of data from replication studies ought to happen through different methods and questions about the data. In this context, researchers can set up competing hypotheses about the outcome of a replication study and provide Bayes factors for the relative support for either hypothesis. Based on the "Replication Bayes factor" (RBF) introduced by Verhagen & Wagenmakers (2014) for *t*-tests, my paper outlines the general idea of the RBF and extends its application to settings in which ANOVAs are used for omnibus and interaction hypotheses, as is common in psychology. The RBF compares a successful replication (i.e. yielding an effect size estimate similar to the original study) with a failed replication (i.e. a null result). Using simulations and examples, the paper investigates the behaviour of the RBF for *F*-tests in different scenarios, with particular consideration of order effects that are not represented in reported *F*-values.

In this section, the idea of the Replication Bayes factor is elaborated for the case for *t*-tests (Verhagen & Wagenmakers, 2014) and then further extended to *F*-tests from studies involving ANOVA designs.

Understanding the Bayes theorem from a perspective of Bayesian belief updating, where one starts with some knowledge about the world or an effect under investigation and one updates

this knowledge using new evidence, allows to think of Bayes factors also in the context of replication studies: A researcher is informed by previous studies on an effect of interest, performs a replication study and uses the data from the replication to update their belief about the effect. The updated belief or knowledge is represented in the resulting posterior distribution. More concretely, this means that the posterior distribution of the effect size from the original study becomes the prior distribution for the replication study. The posterior distribution of the replication study then is the updated belief about the effect size after both the original and the replication study. A Bayes factor set up in this way informs about the evidence in the replication study and how much researchers should shift their beliefs considering the data from the replication study. The Bayes factor taking this particular view was proposed by Verhagen & Wagenmakers (2014) and termed *Replication Bayes factor*.

Phrasing the analysis in terms of "belief" is one way to understand Bayes theorem and Bayes factors (see above), but it is not necessary to interpret it this way. As explained before, the Bayes factor compares two competing models and can also be interpreted in this narrow way. The Replication Bayes factor sets up the two models in an elegant way by considering the posterior distribution of an original study. It essentially compares two possible replication outcomes as models and calculates the respective Bayes factor: Either the replication shows a "consistent signal" by yielding an effect size comparable to the original study (later denoted by H_r), or the replication yields an effect size close to zero and thus is inconsistent with the original study (H_0). One could construct many different hypotheses and models to be compared in a Bayes factor for replications (see above), but the Replication Bayes factor is most closely in line with common interpretations of replication studies. The two possible outcomes are represented as models in the Bayes factor which Verhagen & Wagenmakers (2014) introduce figuratively as positions held by different virtual researchers: The *proponent* (H_r) holds a belief in the original study and expects the replication study to give a similar estimate for the replication as in the original study. The *sceptic* (H_0), on the other hand, is sceptical of the original study and thus expects the replication study to yield an effect size of zero.

Formally, the two models differ in their priors for the parameters of interest. While the sceptic's position is formally a point null hypothesis that assigns all of its probability mass at $\delta = 0$ (where δ represents the effect size measure of interest), the proponent's position is informed by the posterior distribution of the original study, i.e. $\delta \sim p(\delta|Y_{\text{orig}})$. When raw data from original studies are available, the posterior distribution can result from reproducing the original analysis in a Bayesian way. In practice, however, most papers only report summary and test statistics from frequentist significance tests and do not publish raw data. The Replication Bayes factor is therefore designed to take only summary statistics from frequentist analyses into account. Usually, this means frequentist *t*-, *F*-, *r*-, or χ^2-statistics with associated sample sizes, degrees of freedom, and *p*-values. The challenge now lies in setting up a model with minimal additional assumptions incorporating this data and representing the positions outlined above. When raw data from the original study is available, a more elaborate analysis can be performed, e.g. incorporating the data from both studies in a multi-level model (Marsman et al., 2017).

While Verhagen & Wagenmakers (2014) introduced the Replication Bayes factor for one- and two-sample *t*-test results and Boekel et al. (2015, Appendix A) for correlations, many studies in psychology investigate differences between multiple groups and their interactions by running ANOVAs. It seems therefore desirable to extend the Replication Bayes factor to *F*-tests reported in these studies. In the remainder of the section, this extension is outlined and briefly compared to the *t*-test case. For further details on the latter case, please refer to Verhagen & Wagenmakers (2014) and Harms (2018).

2.5.1.1 Replication Bayes Factor for F-tests

The construction of the Replication Bayes factor involves the following steps:

1. Select an effect size measure of interest.
2. Find the posterior distribution for an original study reporting F-tests.
3. Describe the marginal likelihood of the two models H_r and H_0.
4. Estimate the marginal likelihoods to calculate the Bayes factor.

While Cohen's d seems to be an obvious choice for the comparison of two groups (Verhagen & Wagenmakers, 2014), there is a wide range of effect sizes for ANOVA designs (Steiger, 2004). Since the Replication Bayes factor is supposed to be calculated from commonly reported summary statistics (observed F-value, associated degrees of freedom (df_{effect}, df_{error}), number of groups (k), and sample size (N)), an effect size measure with convenient relationships with these numbers is desirable. Cohen's f^2 has these relationships as it can be calculated through

$$f^2 = \sqrt{\frac{df_{effect}}{N}(F_{obs} - 1)}$$

and has simple relationships both to other effect size measures and to the non-centrality parameter λ of the non-central F-distribution. The non-central F-distribution is the distribution of observed F-values under alternative models and simplifies to the central F-distribution at $\lambda = 0$ (the null hypothesis). The non-centrality parameter λ and Cohen's f^2 are related through

$$\lambda = N \cdot f^2 \tag{2.2}$$

Effect sizes in the context of ANOVA studies often have the problem that they are not easily comparable between different studies, especially when designs differ (Steiger, 2004). In the context of replication studies, however, studies of very similar designs are usually compared (i.e. direct and close replications, see chapter 3). Concerns about the interpretability of f^2 are therefore minor in the present case (Steiger, 2004). Another limitation to be aware of is that f^2 is only applicable to fixed-effect ANOVAs and (approximately) equal cell sizes. This also limits the application of the Replication Bayes factor for F-tests as defined here to such cases. For within and mixed designs, the relationship is more complex, and the Replication Bayes factor would need to be calculated in a different way.

Based on the assumptions of frequentist statistics, the likelihood for Bayes factor models follows from the distribution of F-values under the frequentist alternative hypothesis. The likelihood of an observed F-value F_{obs} with associated degrees of freedom and sample size N is then given by the non-central F-distribution, resulting in the following mixture for the model likelihood:[10]

$$\mathcal{L}(Y|f^2) = F_{df_{effect};df_{error};N \cdot f^2}(F_{obs})$$

For the posterior distribution of the original study, we also need a prior distribution, $\pi(f^2)$. While one might resort to the literature before the original study to construct an informed prior, a more universal approach seems to be desirable for the Replication Bayes factor. Verhagen & Wagenmakers (2014) have used a uniform distribution for the t-test case. Flat priors are generally regarded as problematic in Bayesian inference: A uniform prior implies that an effect size between $3 \leq f^2 \leq 4$ is as probable as an effect size between $0.5 \leq f^2 \leq 1.5$. Considering our knowledge about psychological effects in general, this is – in most practical cases

[10] Note that the parameter of interest and the parameter the likelihood function conditions on is f^2 and not F_{obs}!

– highly unlikely. The uniform prior, however, represents the result of a frequentist test most literally. Despite the improper prior distribution, the posterior is a proper distribution since all parameters in the model are observed and larger than 0, and the sample size is $N_{\text{orig}} \gg 1$.

With this information, the posterior distribution of an original study reporting an F-test can be formulated as

$$p(f^2|Y_{\text{orig}}) \propto \mathcal{L}(Y_{\text{orig}}|f^2)\pi(f^2)$$

The constant factor to the proportionality statement is the marginal likelihood, $p(Y_{\text{orig}})$, which is the above term integrated over all values of f^2:

$$p(Y_{\text{orig}}) = \int \mathcal{L}(Y_{\text{orig}}|f^2)\pi(f^2)\ \mathrm{d}f^2$$

The posterior distribution of the original study, $p(f^2|Y_{\text{orig}})$, was named Λ^2 distribution by Lecoutre (1999). In contrast to the Λ' distribution in the case of t-tests (Verhagen & Wagenmakers, 2014), it cannot be easily approximated by a normal distribution. This introduces additional complexity in the estimation of the Bayes factor, as will be detailed below.

For the proponent's perspective on the replication study, the original study's posterior becomes the replication's prior: $\pi_{\text{rep}}(f^2|H_r) = p(f^2|Y_{\text{orig}})$. This leads to the marginal likelihood for the proponent given by

$$\begin{aligned} p(Y_{\text{rep}}|H_r) &= \int F_{\text{df}_{\text{effect,rep}};\text{df}_{\text{error,rep}};N_{\text{rep}}f^2}(F_{\text{rep}})\pi_{\text{rep}}(f^2)\ \mathrm{d}f^2 \\ &= \int F_{\text{df}_{\text{effect,rep}};\text{df}_{\text{error,rep}};N_{\text{rep}}f^2}(F_{\text{rep}})p(f^2|Y_{\text{orig}})\ \mathrm{d}f^2 \end{aligned} \tag{2.3}$$

For the sceptic, in contrast, the prior $\pi_{\text{rep}}(f^2|H_0)$ is 1 at $f^2 = 0$ and 0 everywhere else, simplifying to

$$\begin{aligned} p(Y_{\text{rep}}|H_0) &= \int F_{\text{df}_{\text{effect,rep}};\text{df}_{\text{error,rep}};N_{\text{rep}}f^2}(F_{\text{rep}})\pi_{\text{rep}}(f^2)\ \mathrm{d}f^2 \\ &= F_{\text{df}_{\text{effect,rep}};\text{df}_{\text{error,rep}}}(F_{\text{rep}}) \end{aligned} \tag{2.4}$$

Calculating the ratio between equations (2.3) and (2.4) gives the Replication Bayes factor

$$\text{B}_{\text{r0}} = \frac{\int F_{\text{df}_{\text{effect,rep}};\text{df}_{\text{error,rep}};N_{\text{rep}}f^2}(F_{\text{rep}})p(f^2|Y_{\text{orig}})\ \mathrm{d}f^2}{F_{\text{df}_{\text{effect,rep}};\text{df}_{\text{error,rep}}}(F_{\text{rep}})} \tag{2.5}$$

The challenge lies in the calculation of the integral. While an algebraic solution is yet unknown, one might get a sufficiently precise estimation through numerical estimation. Verhagen & Wagenmakers (2014) used what is known as the *Monte Carlo estimate* for the Replication Bayes factor, which is given by drawing M samples from the posterior-turned-prior distribution $p(\delta|Y_{\text{orig}})$ and calculating the average:

$$\begin{aligned} \text{B}_{\text{r0}} &\approx \frac{1}{M}\sum_{i}^{M} \frac{t_{df_{\text{rep}},\delta_{(i)}\sqrt{N_{\text{rep}}}}(t_{\text{rep}})}{t_{df_{\text{rep}}}(t_{\text{rep}})} \\ \delta_{(i)} &\sim p(\delta|Y_{\text{orig}}, H_r) \end{aligned}$$

While for the t-test case, the samples can be easily drawn from a normal distribution (Lecoutre, 1999; Verhagen & Wagenmakers, 2014), this, however, is not possible for the posterior distribution of an F-test study. Markov Chain Monte Carlo (MCMC) techniques are commonly

used in Bayesian statistics to draw samples from posterior distributions of an unknown form (see above). For calculating Bayes factors, the Monte Carlo estimate for marginal likelihoods as used by Verhagen & Wagenmakers (2014) is known to be inefficient and unstable, especially in cases where prior and likelihood do not mostly overlap (Bos, 2002; Gamerman & Lopes, 2006). While *bridge sampling* (Gronau et al., 2017; Meng & Wing, 1996) can be used to approximate Bayes factors in very complex models, *importance sampling* is a good compromise between efficiency and stability. Instead of drawing samples from the posterior-turned-prior distribution directly, an importance density $g(f^2)$ is used and an adjusted likelihood term is averaged. For the marginal likelihood of the proponent's model, H_r, this leads to (Harms, 2018, eq. 13):

$$p(Y_{\text{rep}}|H_r) \approx \frac{1}{M}\sum_{i}^{M}\frac{p(Y_{\text{rep}}|\tilde{f}_i^2)p(\tilde{f}_i^2|Y_{\text{orig}})}{g(\tilde{f}_i^2)}, \tag{2.6}$$

$$\tilde{f}_i^2 \sim g(f^2) \tag{2.7}$$

Ideally, the importance density $g(f^2)$ is close to the function that is to be approximated. For the Replication Bayes factor, one can draw samples from the posterior distribution by using MCMC techniques and construct a half-normal distribution by calculating mean and standard deviation from these samples. The half-normal distribution is easy for drawing samples and can be quickly calculated at any point without the need for approximation methods. While the normal distribution is not close to the posterior distribution of an F-test study, it is close enough for efficient importance sampling. One of the requirements for importance densities (Gronau et al., 2017) is that it should have fatter tails than the normalised posterior distribution. This is the case in the present example (Harms, 2018, fig. 1). The Replication Bayes factor is finally calculated by drawing random samples from the importance density, calculating the approximated marginal likelihood in (2.7) and dividing it by the marginal likelihood of the sceptic's model, (2.4).

As a Bayes factor, the Replication Bayes factor informs us about the ratio between the two models' marginal likelihood and tells us how much more in line the data is with one of the models. It is important, however, to underline that the Replication Bayes factor does not provide posterior odds of a successful replication as the posterior probability of a model depends on the prior model probability (or the prior odds). Nevertheless, the Replication Bayes factor provides a valuable perspective to the evaluation of replication studies by providing a metric of statistical evidence in the context of two elegantly defined models that capture common notions of "replication success." As many psychological studies investigate differences between multiple groups, the extension to F-tests seemed necessary. Table 2.2 compares the main ingredients of the Replication Bayes factor for the t- and the F-test version.

Table 2.2: Comparison of model elements in Replication Bayes factors.

	t-Test	F-Test
Effect size measure (parameter of interest)	δ	f^2
Prior for original study	Uniform	Uniform
Posterior of original study	Λ' distribution	Λ^2 distribution
Marginal likelihood H_0	$t_{df_{\text{rep}}}(t_{\text{rep}})$	$F_{df_{\text{effect}},df_{\text{error}}}(F_{\text{rep}})$
Marginal likelihood H_r	see Verhagen & Wagenmakers (2014, eq. 6, p. 1461)	see eq. (2.3)

2.5.1.2 Simulations and Examples

For using the Replication Bayes factor in practice, one needs to understand the behaviour of the Bayes factor in different settings. The paper presents three simulations to show (a) the behaviour in three general settings, (b) differences between the Monte Carlo and the importance sampling estimate, and (c) the Replication Bayes factors for t- and F-tests when used in two-group settings where both tests can be used. The simulation results are summarised in the next sections.

2.5.1.2.1 Simulation 1: General behaviour of the Replication Bayes factor In order to understand the behaviour of the Replication Bayes factor, it is useful to simulate different outcomes and look at the resulting Bayes factor. When does the Replication Bayes factor yield strong results, when is it not able to provide informative results? The results further facilitate interpretation in individual cases. For the simulation, several combinations of sample size (between 25 and 100) and effect sizes (between 10^{-5} and 0.7) in both original and replication studies have been created and the resulting Bayes factor have been calculated. The results of this simulation study can be seen in Figure 2.11 (from Harms, 2018, fig. 2, p. 7).

In general, the figure shows that the Replication Bayes factor is increasing – more in favour of the proponent's model implying $f^2_{orig} \approx f^2_{rep}$ – with increasing effect size estimates in the replication study. This underlines a first and important underlying assumption: If the replication study yields a larger effect size estimate than the original study, it is considered evidence in favour of successful replication, implicitly assuming that the theory predicted an effect size "at least as large as in the original study." When the theory informs the researcher that the effect size should be in a certain range and larger effect sizes are counter-evidence, the Replication Bayes factor cannot give results to answer the question of successful replication properly. Such expectations or prior information need to be represented in the definition of the models. This observation also shows the need for careful comparison of original and replication study in light of the theory under investigation, independent of the statistical procedure (see above).

The simulation also shows that larger replication samples provide more evidence leading to the Replication Bayes factor having a larger value. When the original effect size is small, the Replication Bayes factor is close to 1, because the sceptic's and the proponent's model are very similar. Therefore, even large replication studies with effect size estimates close to zero are not informative enough to distinguish between the two models. Again, this underlines the need for a more nuanced comparison of the studies in light of the theory or research setting: Researchers should use further information on the effect and define bounds for reasonable effect sizes to make a judgement on the outcome of the replication. When even small effect sizes matter in practice, small changes between effect size estimates are likely to be important to researchers as well. This can guide researchers in setting up models that more finely distinguish between relevant and non-relevant effect size estimates (cf. equivalence testing and ROPE procedure).

2.5.1.2.2 Simulation 2: Differences between estimation techniques As discussed in the introduction of the Replication Bayes factor, the Monte Carlo estimate used by Verhagen & Wagenmakers (2014) is generally considered to be unstable and inaccurate. For practical purposes, however, it is useful to investigate the severity of this problem. The second simulation was used to compare the Monte Carlo estimate to the importance sampling estimate that has been advocated for above. In the context of the Replication Bayes factor, the critical question is how much original and replication need to disagree for the Monte Carlo estimate to become unstable in a relevant order of magnitude. For this simulation, again, different combinations of original and replication studies were generated. The original sample size was fixed at 15,

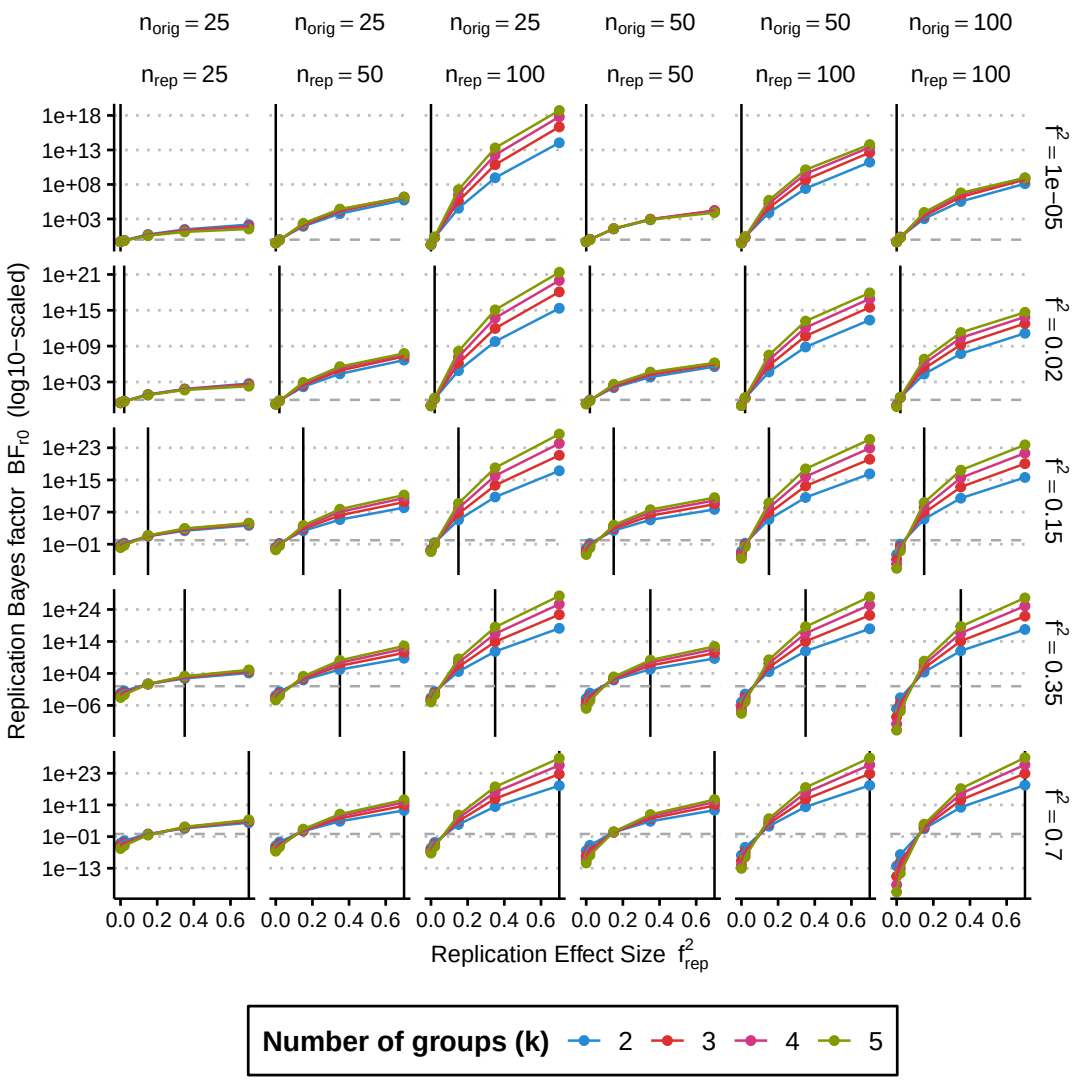


Figure 2.11: Results from simulation 1. The figures show the behaviour of the Replication Bayes factor in different scenarios. Individual columns are different sample sizes in both original and replication study, rows are f^2 effect sizes in original studies between 10^{-5} and 0.7. In each plot, the x-axis shows the effect size in the replication study with the original estimate indicated by a black vertical line. The y-axis is the resulting Replication Bayes factor and coloured lines show the number of different groups. Please refer to the text for a summary of the results.

replication sample size varied between 15 and 100. Original effect sizes were very large (between Cohen's $d = 1$ to 5), while replication effect sizes were small (between $d = 0$ and 0.5). These results are not unrealistic but exaggerated for illustrative purposes. For each study pair, the Replication Bayes factor was calculated using both the Monte Carlo estimate and the importance sampling estimate. All simulations used the Replication Bayes factor for t-tests for simplicity. Results are analogous for the F-test case. Figure 2.12 shows the result of this simulation (from Harms, 2018, fig. 3, p. 8).

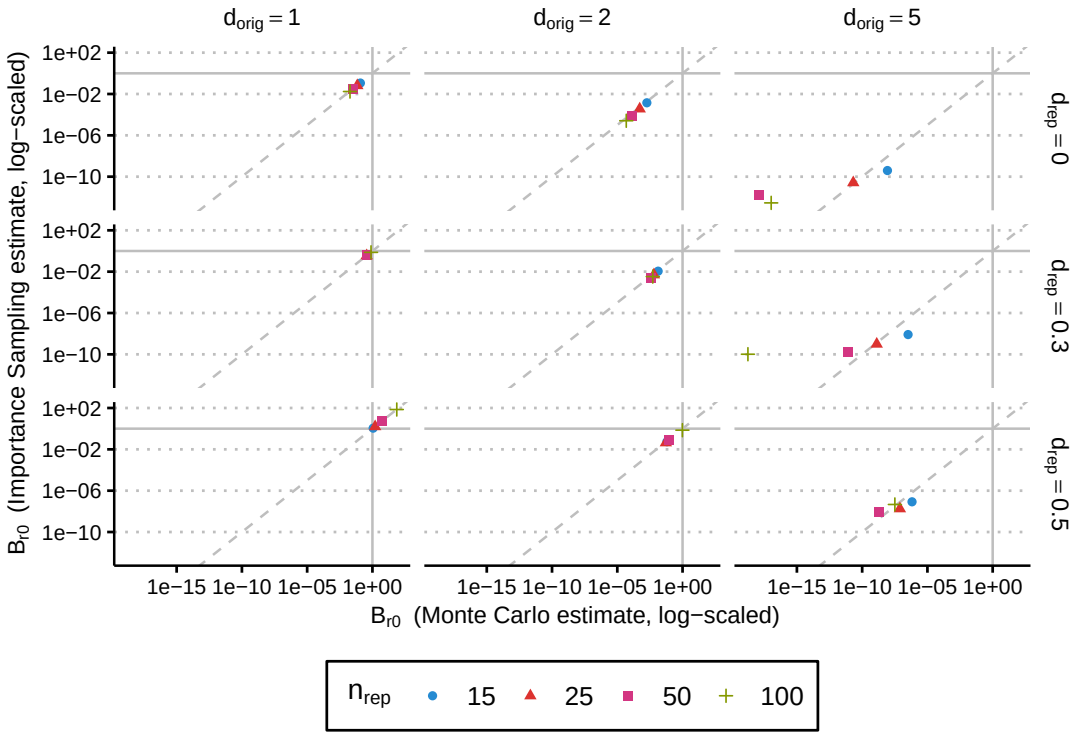


Figure 2.12: Results from simulation 2, showing the differences between Monte Carlo and importance sampling estimates. Large differences exist in cases where original and replication studies yield very different effect sizes. While the interpretation would not change, the importance sampling provides the more precise estimate of the Bayes factor. *Note:* Figure reproduced from Harms (2018).

The simulation shows that importance sampling and Monte Carlo estimate differ substantially in extreme cases, i.e. only when the replication effect size was $d_{\text{rep}} = 0.5$ and the original effect size was $d_{\text{orig}} > 2$. It is not only the case that the difference in estimates is substantial enough, in this case, to be visible from the descriptive statistics alone, but also the general interpretation of the Bayes factor would not change.

For the Replication Bayes factor with its very simple models, the different estimation techniques are in practice not as relevant as previous literature might suggest. For more complex models, the difference can and will be larger. Nevertheless, importance sampling has the advantage to provide a more robust estimate, which is why it should be preferred. On modern computers, the computational requirements are moderate for the Replication Bayes factor, and the `ReplicationBF` package (see below) makes it easy to make use of importance sampling when calculating the Replication Bayes factor.

2.5.1.2.3 Simulation 3: *t*- vs. *F*-test in two-group designs For data from two independent groups, traditional *t*-test and ANOVA will yield the same inferences and the same *p*-value since $t^2_{obs} = F_{obs}$. This offers the chance to investigate the Replication Bayes factor's behaviour when used on the same data, once using the *t*- and once using the *F*-statistic. Does it yield the same result? As the third simulation shows, this is not the case. While this is not directly evident from Figure 2.13 (Harms, 2018, p. 9, fig. 4), investigating the relationship between the two calculated Bayes factors shows that they are near-perfectly correlated across all scenarios ($r = 0.999$), but the ratio is about 2.21. That means the Replication Bayes factor for the *t*-test is about two times larger than the Replication Bayes factor for the *F*-test on the exact same data. How can this be? The observed *t*-statistic carries more information because it contains a sign indicating the direction of the effect. By squaring the *t*-value, we remove this information and the Replication Bayes factor using the *F*-value cannot take the direction into account. Therefore, the statistical evidence in favour of one model or the other can only be lower (Harms, 2018, p. 7).

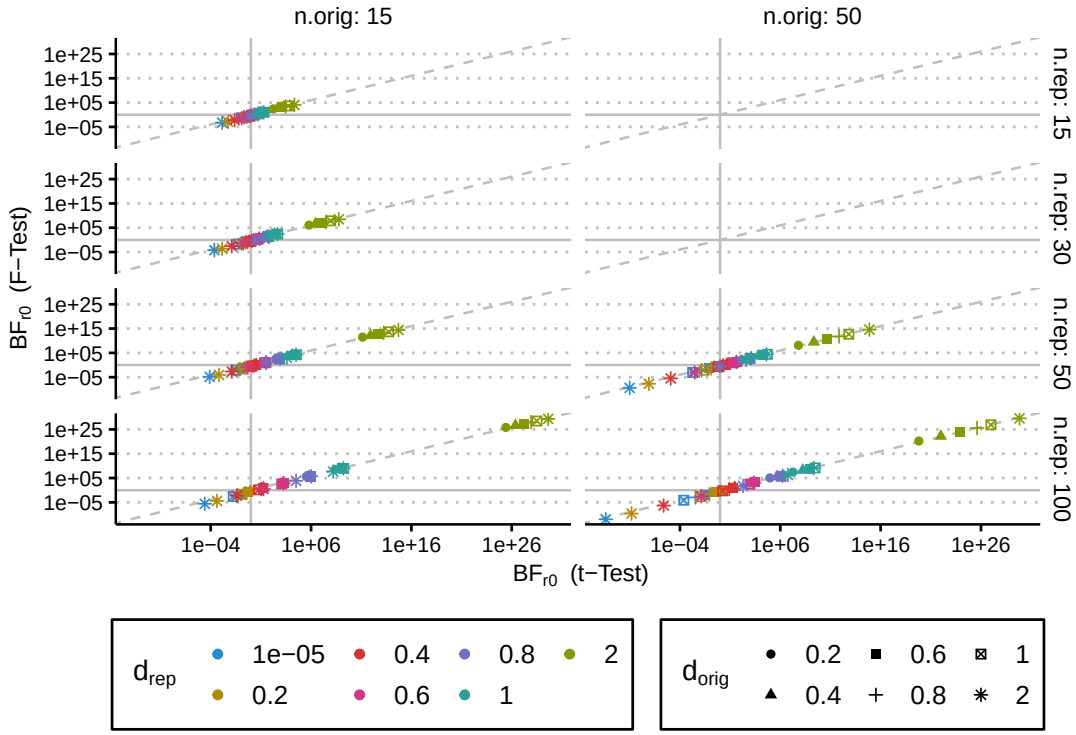


Figure 2.13: Results from simulation 3, comparing the Replication Bayes factors for *t*- and *F*-tests on the same data set. The dashed grey line is equality between both estimates. Axes are log-scaled. Bayes factors are nearly the same, but the estimate for the *t*-test is about two times larger than the Replication Bayes factor for *F*-tests. *Note:* Figure reproduced from Harms (2018).

2.5.1.2.4 Example: Order effects in ANOVA settings In the paper, three examples are presented. The first two examples show how to apply the Replication Bayes factor for *F*-tests in two different replication studies. The third example investigates a problem occurring when using summary statistics in scenarios with multiple groups, which will be summarised below. As noted before, the *F*-statistic of a common ANOVA setting does not convey information

about the order of effects. As an omnibus hypothesis test, it merely represents the information whether there is any difference. Therefore, qualitatively different results can produce the same F-value and therefore the same statistical inference, while their interpretation would differ wildly. Since the Replication Bayes factor, as presented here, is designed to only take the frequentist test-statistic as reported into account, it too cannot consider the ordering of the effect either.

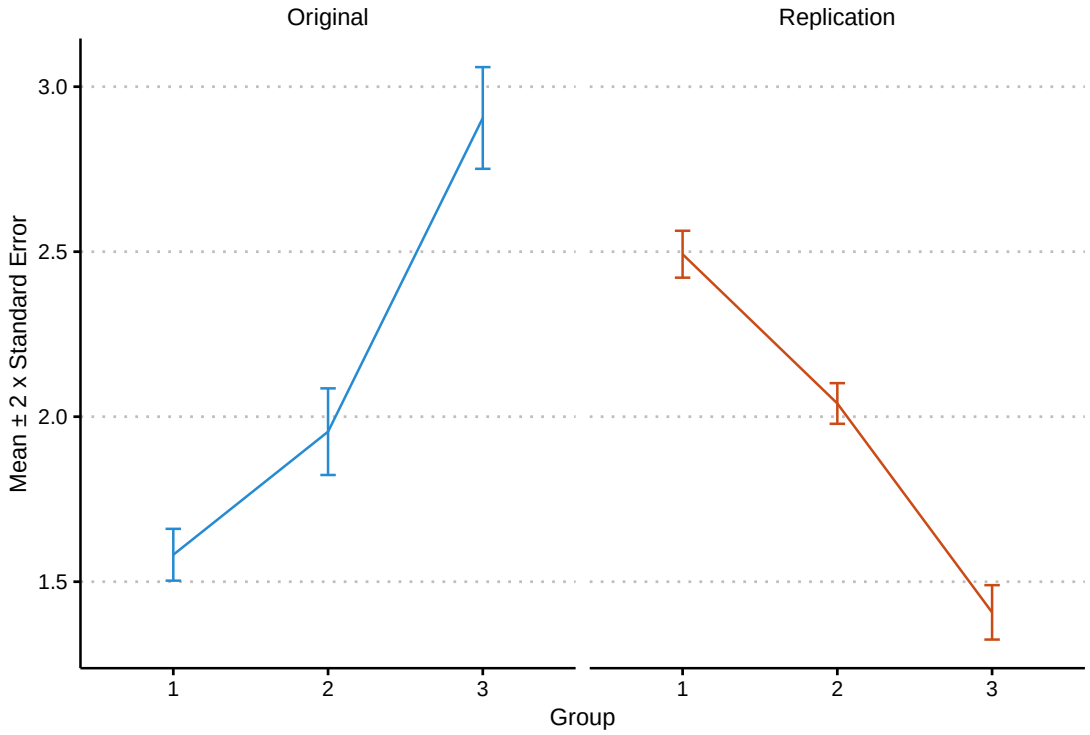


Figure 2.14: Imaginary results of a replication study yielding the same observed F-value for a very different pattern of results. The Replication Bayes factor will be in favour of a replication consistent with the original study. This underlines the need for investigating contrasts in addition to the omnibus result. *Note:* Figure reproduced from Harms (2018).

Imagine an experiment in which three groups are compared and 15 participants per group are recruited and randomly assigned to an experimental condition. For the original study, imagine the true (and in reality obviously unknown) population means of the three groups are $\mu_1 = 1.5$, $\mu_2 = 2.2$, and $\mu_3 = 2.9$, respectively with a standard deviation $\sigma = 1$ for all groups. Running a one-way ANOVA on a simulated data-set with these parameters yields a significant result for the obtained sample ($F(2, 42) = 7.91$, $MSE = 0.88$, $p = .001$, $\hat{\eta}^2_G = .274$). The left side of figure 2.14 visualises the result. A replication study aims to investigate this original effect in a new sample with 30 participants per group. The replication study, too, yields a significant results for the omnibus F-test ($F(2, 87) = 7.60$, $MSE = 1.17$, $p = .001$, $\hat{\eta}^2_G = .149$). Looking at the results (see the right plot in figure 2.14) and comparing them to the original study clearly shows that the replication provides a very different result: While group means are similar, they are ordered in the opposite direction. While this might be simply a coding error or a difference in the experimental setup, it might as well be the result of an exact replication. In any case, the F-statistic does not convey this interpretation. Subsequently, the Replication Bayes factor also does not take the pattern of the result into account and shows

evidence in favour of a successful replication ($B_{r0} = 38.261$). This contradicts the sensible interpretation of the results, which would consider the replication contradicting the original result.

This highlights the need for researchers not to rely on a single statistical value, but to take descriptive results as well as different statistical key figures into account. In the particular case of the example, post-hoc t-tests or planned contrasts can make the statistical test more informative. For example, the difference between groups 1 and 3 is significant in both the original ($\Delta M = -1.32$, 95% CI $[-2.02, -0.63]$, $t(20.81) = -3.95$, $p = .001$) and the replication study ($\Delta M = 1.09$, 95% CI $[0.49, 1.68]$, $t(56.77) = 3.64$, $p = .001$). But as is evident in the sign of the t-statistic, the difference is in opposite direction and the Replication Bayes factor for the t-test will correctly indicate evidence against a successful replication ($B_{r0} = 0.0015$). When investigating F-tests, researchers should be careful when making conclusions. One statistical way would be to use a directed hypothesis in replication studies where original studies have shown particular patterns that have become part of the theory or phenomenon under investigation. It is unlikely to assume that the original study's result in the present example is "yes, something is going on," but rather that the particular pattern is the conclusion of the study. The replication study should therefore test the hypothesis whether the particular pattern has been replicated.

2.5.1.3 The `ReplicationBF` Package

Along with the paper (Harms, 2018), a package for R (R Core Team, 2019) is available at https://github.com/neurotroph/ReplicationBF. The package allows to calculate Replication Bayes factors for all scenarios outlined before, i.e. t- and F-tests.

```
> install.packages('devtools')
> devtools::install_github('neurotroph/ReplicationBF',
    dependencies=TRUE)
> library(ReplicationBF)
> rbf <- ReplicationBF::RBF_Ftest(27.0, c(3, 48), 52,
                                  3.2, c(3, 33), 37)
> rbf$bayesFactor
[1] 0.02785122
```

By convention the results of the packages are presented in favour of the proponent's model (i.e. B_{r0}).

2.5.1.4 Summary and Discussion

In summary, the Replication Bayes factor works as intended, but it should be used carefully with respect to the particular statistical questions asked. It is better thought of as a basis for adapting the models used in the Bayes factor to the hypotheses under investigation. There is, for example, ongoing research on how to include order constraints in Bayes factors (e.g. Mulder, 2016).

In this section, the interpretation of replications was rather dichotomous between "successful" and "failed" replications. This narrow view not only reduces the Replication Bayes factor to a binary decision criterion but also limits a detailed investigation of replication studies. The next chapter will more extensively discuss replication studies and how to interpret the outcomes of a replication. When using more elaborate ways to describe such outcomes, the Replication Bayes factor can present one statistical piece of evidence among different ways to analyse the data. The continuous nature of Bayes factors allows for a nuanced interpretation, also

indicating a lack of evidence of either model. As explained before, significance tests and p-values do not provide such information, but can also offer statistical information to be used when interpreting both original and replication studies.

2.6 Conclusion

This chapter briefly outlined some of the historical aspects to the current use of the dominant statistical framework in psychology, null hypothesis significance testing. With improvements in computational capabilities and a renewed interest among statisticians and applied researchers, Bayesian statistics have received increasing attention over the last years. This attention brings both theoretical developments as well as new software packages for researchers to make use of. While most published research still relies on significance testing, exploring Bayesian statistics offers additional methods useful for the analysis of data from psychological studies.

Focusing on the replication crisis in psychology, using Bayesian statistics will not solve replicability once and for all. Looking at the root causes of the increasingly widespread doubt about published study results and theories, the misuse of statistical procedures and their resulting misinterpretation is only one part of the problem. Some researchers do advocate in favour of Bayesian statistics replacing traditional methods (McShane et al., 2019; cf. Savalei & Dunn, 2015), but Bayesian methods can be misused and misinterpreted as much as frequentist methods. It seems more imperative to increase transparency (cf. pre-registration, open data, and open materials) and improve the way statistical models are set up and used. Statistics need to be used in a way that allows to properly answer the questions asked by the substantive theory. If the statistical analysis does not represent the researcher's questions, it does not actually provide the information the researcher seeks. Learning about the technical details in statistics can help users to understand what questions a method can answer and how the methods can be altered in order to fit a new question. Building probability models using Bayesian statistics has several advantages over running frequentist significance tests as provided by a software package of choice.

Overall, there is no need to decide on one way of doing statistics. Both statistical approaches, frequentist and Bayesian, can be used side-by-side. The important task for researchers lies in choosing the tools that allow for an answer to their statistical question asked and interpreting the statistical results in a principled way that allows for valid inferences on their substantive questions. A significance test might be appropriate for checking data against a well-informed theoretical prediction. If theories are not mature enough to provide such predictions, more exploratory analyses, developing statistical models, or descriptive statistics might be a better-suited alternative. For mature theories that involve a range of boundary conditions and proposed mechanisms, model-building should also include proper model checking and, probably, extension. When doing inference on the statistical results, uncertainty is a valuable statistical property that has to be considered in the interpretation: It can provide a better understanding of model performance and which quantities cannot reasonably be relied upon. Bayesian statistics provide a natural framework for thinking probabilistically about uncertainty. Statistical uncertainty about model parameters is, however, only one kind of epistemological uncertainty. Ultimately, empirical researchers who want to analyse their data on their own should improve their statistical training. The replication crisis has sparked a range of introductory papers for this. More interdisciplinary discussion between domain experts and statisticians can further help to provide instructional guidelines for applied researchers.

Chapter 3

Replication Studies

A central issue of the "replication crisis" (Pashler & Harris, 2012) is, obviously, replication studies. Historically, the proportion of original research being replicated in independent studies has been very low (Smith, 1970; Theodore D. Sterling, 1959). Replication studies are generally considered not worthwhile to be published as they are said not to provide "novel insights" and to have done little to further scientific knowledge (Neuliep & Crandall, 1990, 1993). Interest in replications and publications of replication studies have increased since the 1990s (Makel, Plucker, & Hegarty, 2012; but also see Evanschitzky, Baumgarth, Hubbard, & Armstrong, 2007) and the replication crisis has further fueled interest in replications. New publishing formats such as *Registered Replication Reports* [Simons, Holcombe, & Spellman (2014); RRR] also helped to direct interest towards replication research as they provided a way to ensure publication of replication studies even if they do not yield significant results. At the very least, the replication crisis has increased awareness that replications are rarely done systematically and that there are many things to consider when conducting a replication (e.g. S. F. Anderson & Maxwell, 2016; Brandt et al., 2014). Zwaan, Etz, Lucas, & Donnellan (2017) recently provided a "manifesto for reproducible research" and advocated that systematic replication should become the norm and to put more emphasis on *direct* replications as well.

To introduce a common basis for describing replication studies, the following section will introduce a taxonomy of replication studies. Moreover, based on recent recommendations and a published example, guidelines for conducting replication studies are proposed. Finally, the third paper of this thesis is summarised as an example of a close replication study (Harms et al., 2018).

3.1 Taxonomy of Replication Studies

Over decades of replication research, several different terms have emerged to describe replication studies. Unfortunately, there is currently no consensus or precise language used universally across different fields, subfields, or teams of researchers. This is most apparent in the use of the terms *replicability* and *reproducibility*. For the present work, it will be primarily referred to as *replications* and *replicability* where we aim to replicate findings in independent studies. In contrast, *reproducibility* shall refer to the goal of using existing data to reproduce the statistical results and interpretations of an original study. Some authors (e.g. Tsang & Kwan, 1999) have considered both concepts in combination and they are indeed closely linked. Reproducibility is, in general, a way of scientific housekeeping (Asendorpf et al., 2013b) and a necessary condition for replicability: If an original finding is not to be trusted, because the results and interpretations as stated in a publication cannot be reproduced using the same data

and same analysis, we ought not to expect a replication to show the same results leading to the interpretation in the original study. Analysing existing data from a previous study using a different statistical method constitutes a *re-analysis*, which shall herein be considered as part of reproducibility. While fraud seems to be the most obvious example of non-reproducible and likely non-replicable studies, it should be noted that recent fraud cases used manipulated data. Therefore, the analysis might have indeed been reproducible, but the underlying data were not real and successful replication should not be expected.

Therefore, in the following, the term *replicability* is considered in a narrow sense: That is when conducting a new study to collect new data and to analyse this data to shed more light on the original finding. Such new studies, i.e. replication studies, can vary in their similarity to the original study. To describe this closeness, several classifications or taxonomies have been proposed and used before (e.g. Brandt et al., 2014; LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018; Lykken, 1968; Schmidt, 2009). While this thesis is not aimed at "reinventing the wheel," it is important to make a clear definition of the terms as a common language to describe replication studies within the context of this thesis. The aim here is to provide a classification of replication studies that is useful to describe different studies and that provides a basis for discussion. It aims to summarise and compromise previous work and to establish clarity in this thesis.

On which dimensions should replications be considered? Schmidt (2009) based his terminology on five different functional goals of replications (Schmidt, 2009, p. 93):

1. To control for sampling error (chance result),
2. To control for artifacts (lack of internal validity),
3. To control for fraud,
4. To generalize results to a larger or to a different population,
5. To verify the underlying hypothesis of the earlier experiment.

Each goal requires the replication study to focus on a different aspect of design and analysis. To control for sampling error, for example, it is important to keep the population and sampling plan identical between the original and replication study. The fifth goal, verifying the underlying hypothesis, in contrast, allows for more flexibility in the experimental setup.

Schmidt (2009) proposed a system of eight classes based on Hendrick (1991) to describe "research reality." The classification of replications is based on which classes are kept constant and which are systematically varied. Based on this system, one could extend the classes by describing more specifically what aspects can and need to change to provide an even more detailed classification of replications. A key aspect of Schmidt (2009)'s categories of replications is the insight that direct replications are considered "low risk": They have aimed purely at fact confirmation, and both a corroborating and a falsifying outcome are relevant for the advancement of the theory. Conceptual replications, on the other hand, are high risk: If they successfully generalise the previous findings, the theory is advanced. If the conceptual replication fails, however, it is difficult to conclude whether the original finding was false or whether the conceptual extension was not successful. While being a high-risk endeavour, conceptual replications are necessary for a deeper understanding of a phenomenon as well as further theory development. In the classification provided above, exact, direct, and close replications can be considered as this kind of "low risk" undertaking.

Other authors (e.g. Brandt et al., 2014; Lykken, 1968) have taken a more practical look and distinguished types of replications primarily through the closeness of different dimensions. Based on this range of previous works, table 3.1 summarises four different categories of replications ranging from nearly identical studies ("exact replications," see below) to highly different studies ("conceptual replication"), which aim to investigate the boundary conditions of and possibly extend previous findings.

Table 3.1: A proposed taxonomy for replication studies as a basis to describe replication studies.

Type of Replication	**Summary**
Exact Replication	New sample from the same population is drawn and the study is carried out with the exact same tools. Theoretical concept: Usually impossible to carry out as even smallest changes to the experimental setup are inevitable (e.g. weather, events, time of day, ...). See also: Lykken (1968): "literal replication"; Tsang & Kwan (1999): "exact replication"
Direct Replication	Using the same material and experimental setup a study is conducted by a different researcher targeting the same population. See also: Schmidt (2009): "direct replication"
Close Replication	Based on the reported experimental setup, materials are re-created and a replication is carried out. Experimental design and underlying hypotheses remain unchanged. See also: Lykken (1968): "operational replication"; Brandt et al. (2014): "close replication"; Tsang & Kwan (1999): "empirical generalization"
Conceptual Replication	Using a different experimental setup and/or population, the underlying theory, idea, or construct of a study is put to test. See also: Lykken (1968): "constructive replication"; Brandt et al. (2014) and Schmidt (2009): "conceptual replication"; Tsang & Kwan (1999): "conceptual extension" / "generalization and extension"

More fine-grained distinctions are possible and offered by some of the referenced works. The goal for this taxonomy is primarily pragmatic: A common language to describe different degrees of closeness between an original and a replication study. For practical use, however, "exact" (Tsang & Kwan, 1999) or "literal replications" (Lykken, 1968) are generally considered impossible as long as the original researcher does not simply collect more data (Feest, 2018; Lykken, 1968; Wolfgang Stroebe & Strack, 2014). Two experimental situations cannot be identical. Contextual factors, such as time of day, recent events (both globally and affecting participants personally), weather, etc. are outside of an experimenter's control and it is generally considered possible that each and any of these contextual factors influence the effect under investigation.

Other taxonomies also include other categories of replications, which are better grouped under the question of "reproducibility" (see above). Tsang & Kwan (1999), for example, use *checking of analysis* (the same dataset is analysed using the same measures and analysis; reproducibility in a narrow sense) and *reanalysis of data* (the same dataset is analysed using different measures and analyses; reanalysis or reproducibility in a wider sense). As explained before, reproducibility of the original analysis is not part of "replicability" within the meaning of this term as it is used within this thesis.

In particular, a study can fall into several categories at the same time. For example, if parts of the experiment are reproduced in the manner of a *direct* replication, but some additional conditions are added to investigate a previously unpublished aspect of a theory or finding, therefore being a *conceptual* replication. If a single study covers both aspects, and the conceptual part does not interfere with the direct part, it can be considered to be both a direct and a conceptual replication at the same time.

This examples also shows that a clear distinction is not always possible and that one can have different perspectives on the same study. This should not distract from the fact, however, that taxonomy as proposed here is a useful tool to set a common language for describing replication studies. It also underlines the importance of both, direct replications as well as close or conceptual replications (Tsang & Kwan, 1999). For the remainder of this work, and the next section, in particular, the above taxonomy is used to describe replication studies.

3.2 Guidelines for Replication Studies

Several authors have also previously proposed guidelines for planning, designing, conducting, and evaluating replication studies (e.g. Brandt et al., 2014; LeBel et al., 2018; Zwaan et al., 2017). The following points, which will be addressed in this section, are a summary of different papers and include some novel aspects that have received little attention so far:

1. Selecting studies and effects of interest
2. Selecting focal hypotheses
3. Re-analysing original results
4. Contacting original authors
5. Planning the replication study
6. Pre-registering the replication
7. Running the replication study
8. Analysing data

3.2.1 Selecting Studies and Effects of Interest

There are several justifications why a certain study might be a candidate for a replication attempt. In many cases, researchers are interested in replications within their own academic domain. One scenario could be, for example, to ensure that previous findings are robust before starting a new line of research in this field. While seemingly a reasonable approach, replicating previous findings and publishing the results of the replications are rarely done in practice. The *Reproducibility Project: Psychology* selected studies from three major journals for systematic replication to estimate the rate of replicability across a whole field. A practice that has been criticised (e.g. D. T. Gilbert et al., 2016b). After publishing replications of single studies, possibly with a result contradicting the original study, the selection is often criticised for being driven by personal goals or for being ad-hoc and unscientific. Such claims are hard to dismiss as they are often without a rational basis themselves. Curiosity or doubt about an original effect can also be a reason for running a replication study and, as outlined before, replication should be part of the standard process in empirical sciences (see above and Zwaan et al., 2017). In practice, however, the selection of target studies is a problem of resource allocation: If a meta-researcher has limited time, money, and personnel, which studies should they focus on for replications to maximise the generation of new knowledge? There is ongoing research on quantifying "*replication value*," i.e. the added value a direct or close replication would provide (Coles, Tiokhin, Scheel, Isager, & Lakens, 2018; Field, Hoekstra, Bringmann, & Van Ravenzwaaij, 2019; Isager, 2018, 2019). This offers both an avenue for replication researchers who focus on replicating studies in systematic ways to advance robustness of findings in general as well as researchers who need to decide and justify which particular study on a given issue they should replicate.

Beyond the practical considerations that come with selecting a study for replication, researchers should also cautiously appraise the underlying theory. In psychology, formal theory development and appraisal, however, have played a minor role in recent years [see also chapter 4; Fiedler (2017); Glöckner, Fiedler, & Renkewitz (2018)].

3.2.2 Selecting Focal Hypotheses

After selecting a paper or theory for replication, researchers need to decide which experimental setup (e.g. in the case of multi-study papers) and which focal hypothesis they aim to replicate For direct or close replications, the target of replication is usually a single focal statistical hypothesis. Depending on the goal of the replication (see also Schmidt (2009)'s functional goals) one might also decide to replicate multiple effects, outcomes, or statistical tests in a single replication study. When selecting single effects and statistical tests, a good justification is helpful – not only for the sake of the argument but also for providing a basis on which future replications can target the same or similar effects. Selecting an effect size or a specific statistical test as a target for the replication and transparently justifying the selection, e.g. in a pre-registration, ensure that the reported statistical results in the replication are not cherry-picked to allow for a certain interpretation (cf. *p*-hacking, chapter 2.1.3.3). For systematic replication studies, such as the RP:P very general rules can be useful to reduce possible bias in the selection of a target for replication (Open Science Collaboration, 2015, pp. aac4716–2):

> By default, the last experiment reported in each article was the subject of replication. [...] The key result had to be represented as a single statistical inference test or an effect size. In most cases, that test was a t test, F test, or correlation coefficient.

When the replication aims to validate a specific effect, a researcher might have better reasons for choosing a very specific test or effect size estimate as a target for the replication.

3.2.3 Re-Analysing Original Results

As discussed above, the reproducibility of the original analysis is generally a necessary condition for expecting replicability. As original studies rarely provide their raw data openly, reproducibility cannot always be checked in practice. Meta-analytical techniques, such as *p*-curve (Simonsohn et al., 2014; Simonsohn, Simmons, & Nelson, 2015), *z*-curve (Schimmack & Brunner, 2017), or R-Index (Schimmack, 2012) can nevertheless provide insights into possible biases in the reported statistics. For example, studies with a skewed *p*-curve indicating selective reporting of significant results, or multi-study papers reporting only significant results (leading to a low R-Index) should inform a replicator that the original findings might contain biases and that they should investigate which kind of biases a replication can address. This can guide the experimental design, sampling plan and sample size, as well as the preregistration. It is important to stress that these statistical techniques cannot ultimately indicate intentional biases, fraud, or misconduct and that these techniques are still under development and discussion (Erdfelder & Heck, 2019; Morey, 2017). Statistical techniques to investigate biases, however, inform a researcher about possible caveats before naïvely replicating an original study at face value.

3.2.4 Contacting Original Authors

There is widespread expectation that replication researchers should contact original authors and inform them about their replication attempt and their results before publishing a report on the replication. From a meta-scientific perspective, there is little reason to make this mandatory as a publication should contain all information necessary for a replication and replications are part of the normal way of science. In practice, however, contacting the original authors can be beneficial for replicators as the experience shows that experimental designs might deviate from the reported setup. This might be due to human error, vague descriptions,

or word limits in the methods section. As data and materials are also made openly available only rarely, contacting original authors can help to provide such materials for replicators. Furthermore, contacting original authors might also provide a chance to use the raw data of the original study for more elaborate statistical analyses (see last point below). Original authors ought to be experts on both the reported effect and the methods employed. They might then also be able to provide additional expertise in planning a study that conveys the underlying theory most efficiently and appropriately. However, if original authors do not respond, if materials and data are unavailable, or if researchers are not cooperative, this must not diminish the value of a replication: Original publications need to be as detailed as possible, and researchers ought to be able to replicate results based on published research alone.

3.2.5 Planning the Replication Study

The design of a replication study is easily guided by the intention underlying the replication: Whether a direct, close, or conceptual replication is the most adequate approach is determined by the goal of the study (see taxonomy above). The type of replication then determines how closely materials and experimental setup have to match the original study and how much freedom in the design is acceptable. As discussed before, depending on the study in question, there can also be a mix of direct, close, and conceptual aspects. For example, a researcher could include an additional treatment group with a different operationalisation of the construct in question or add a different population to the study. For transparent reporting and adequate analysis, such additions should be included in both the pre-registration (see next recommendation) and the statistical analysis.

When planning the replication study, special consideration should be given to the issue of sample size. When replicating previous work, published effect sizes cannot be used at face value for power analysis. Published effect size estimates tend to be inflated due to publication bias, measurement error, and *p*-hacking (Loken & Gelman, 2017; Theodore D. Sterling, 1959). Taylor & Muller (1996) propose an approach to power calculation using confidence bounds on power. More universally, a replication should include more participants than the original study and Simonsohn (2015) recommends using 2.5 times as many participants as a rule of thumb. As with original studies, replication studies should justify their sample size some way or another in their pre-registration.

3.2.6 Pre-Registering the Replication

One of the simplest and most effective recommendations for countering selective reporting and post-hoc theorizing (Kerr, 1998) is pre-registration. Pre-registration requires investigators to lay out study design, hypotheses, analysis plan, sampling plan, and study materials before conducting the study (T. Hardwicke, 2016; Nosek et al., 2018). Several online websites offer ways to securely store a pre-registration with a timestamp, so readers can transparently see when and how the study was pre-registered and compare to the final study report. Examples for such websites are AsPredicted.org or the Open Science Framework. A pre-registration's primary goal is to increase transparency on the scientific process and to support primarily confirmatory research. This, obviously, does not protect against fraud by running a study, posting the pre-registration online after the results are known, and reporting the study a couple of weeks later pretending the registration was actually done prior to the study. Nevertheless, pre-registering a replication is a powerful tool for many experimental studies to combat problems with reporting and analysis that have been persistent over the last years.

A counter-argument to pre-registration is often that researchers are afraid that it prohibits exploratory research. On the contrary, however, pre-registration forces researchers to be explicit about parts of a research programme that are confirmatory versus exploratory. Exploratory

research is essential for theory development, and pre-registration helps by clearly identifying which part of a study is exploratory and which part is confirmatory (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

Registered Reports (Chambers et al., 2015; Chambers & Mellor, 2018; T. E. Hardwicke & Ioannidis, 2018) is a publication format that incorporates pre-registration in the publication process: Instead of only including a final report, the submission to a journal includes the study rationale along with the method section. Moreover, it is reviewed before the study is run. Peer reviewers make their comments regardless of the study outcome and can provide remarks on the study design at a time when feedback on the design can still be considered. If the manuscript is accepted at this stage, it receives an "in-principle acceptance" that also allows researchers to publish studies irrespective of the study outcome. This effectively combats both selective reporting and post-hoc reasoning as well as publication bias. Registered Replication Reports (Simons et al., 2014) is a similar format specifically for replication studies.

For replication studies, there should be little concern about pre-registration as the target study is already a pre-registration for the largest part of direct or close replication. In such a replication, researchers aim to have a similar design to the original study and therefore the original study already is some kind of pre-registration. Nonetheless, a formal pre-registration of the design and analysis plan helps to, for example, clarify deviations between original and replication study.

3.2.7 Running the Replication Study

Replication studies often suffer from similar problems as original studies such as non-representative or WEIRD samples (Henrich et al., 2010). Projects such as the *Psychological Science Accelerator* (Moshontz et al., 2018) or *StudySwap* (Chartier, Riegelman, & McCarthy, 2018) can help to increase the sample's size, diversity, and representativeness. Especially in replications aiming to generalise findings to larger or different populations, a, possibly international, collaboration can help to improve the study. Apart from this, running a replication study is similar to running an original study. With respect to the pre-registration researchers should take note of possible deviations from the registered study protocol. As unforeseen challenges might arise in any empirical study, they should be addressed during the study and transparently reported in the final article.

3.2.8 Analysing Data

The statistical analysis of a replication study should comprise three aspects: First, the replication study should be analysed as a study on its own. Second, the results from the replication study should be compared to those of the original study. Third, if the data from the original study is available, the data from both the replication study and the original study should be analysed jointly. This it allows for more in-depth analysis when a multi-level model is used to model potential differences between the two (or more!) studies.

The analysis of the replication study on its own should follow the pre-registered analysis plan (see above) as well as general recommendations for analysing empirical data. Researchers might want to consider Bayes factors with a properly defined alternative model (see chapter 2.2.2) or a fully Bayesian analysis. Chapter 2.5 offered different perspectives on the second aspect mentioned above, i.e. comparing the replication study with the original study statistically. In general, a replication researcher might compare the studies with respect to different questions such as

- Do descriptive results show similar patterns?

- Was the original study able to detect an effect as small/large as in the replication study with 33% power ("small telescopes," see Simonsohn, 2015)?
- Is the replication's effect size in line with the original study's effect size (LeBel et al., 2018)? For example, by using Replication Bayes factors [see section 2.5.1; Harms (2018); Ly, Etz, Marsman, & Wagenmakers (2017); Verhagen & Wagenmakers (2014)].
- What effect size estimate does a meta-analytical approach yield?

It is generally advisable to disregard vote counting, or simply comparing the pattern of statistical significance, as this can lead to misleading results. Instead, researchers should ask multiple questions using different tools (S. F. Anderson & Maxwell, 2016; Maxwell et al., 2015). For replication studies, it is just as important as for original studies to think about the theoretical question one asks and how this question can be translated in a statistical framework (Hand, 1994; Meehl, 1990a).

The joint analysis of data from both the original and the replication study allows researchers to model differences between the studies as fixed or random factors in a multi-level or mixed-effects model (Gelman & Hill, 2007; Yarkoni, 2019). This allows more nuanced inferences but requires the original data to be available. Meta-analytical models, however, can also be used on summary or test statistics. Such an approach also generalises to more than one replication study. Marsman et al. (2017) offered an example of different Bayesian methods to analyse a range of replication studies for inferences about the replicability of social psychological studies.

3.2.9 Summary

Replication studies are as diverse as original studies, and the previous recommendations will not fit for any replication study. For example, the requirements for replications of an international, multi-centred study on psychotherapy effectiveness are very different from the requirements for replications of laboratory experiments in cognitive psychology. The recommendations presented within this thesis should, nevertheless, be general enough to be applied to a wide range of areas and incorporate current recommendations. The most fundamental recommendation is to think clearly about the purpose of a replication study and to formulate a question to be answered through a replication study. Decisions about replication design and analysis often follow more or less directly from this. Pre-registration is a powerful tool to increase transparency and registered reports allow researchers to ensure publication irrespective of the replication outcome.

The next section will summarise the replication study in Harms et al. (2018). It follows many of the recommendations on study design and analysis as outlined before. Some of the recommendations mentioned ahove are a direct result of the learnings from the study presented (e.g. issues with contacting the original authors), though.

3.3 Article III: Does it actually feel right?

This section is a summary of this thesis' third article:

- **Harms, C.**, Genau, H. A., Meschede, C., & Beauducel, A. (2018). Does it actually feel right? A replication attempt of the rounded price effect. Royal Society Open Science, 5(4), 171127. https://doi.org/10.1098/rsos.171127

The content and several quotes are taken directly from this paper, which is also included in the appendix of this thesis.

Summary: Our paper applies the recommendations mentioned in this thesis to the "rounded price effect" published by Wadhwa & Zhang (2015). After a review of the original paper using meta-analytical methods (*p*-curve and *R*-Index), we conducted a close replication study. In particular, our replication study differed primarily in the population we sampled from (a German convenience sample vs. MTurk participants in the original studies) and the materials used. Following the recommendations presented in this thesis, we analysed the results using both frequentist and Bayesian statistical methods. Despite some isolated patterns pointing in the same direction, the results led us to be sceptical about the original studies. Our conclusion summarises our results and discusses implications for applied replication research in general.

In the *Journal of Consumer Research* Wadhwa & Zhang (2015) published a series of five studies in which they investigated an effect they called the "rounded price effect": Depending on the context, customers were more likely to choose a product with a non-rounded price (e.g. 34.98 Euros) or a rounded price (e.g. 35.00 Euros), respectively. The context was either emotional (e.g. buying a personal present for someone) or cognitive (e.g. buying a product for a school project). A rounded price is said to be more fluently processed because rounded numbers appear more often in everyday life. The authors argue that rounded prices lead customers to rely more strongly on their feelings, which, in turn, leads to an increased purchase likelihood in emotional contexts. In contrast, non-rounded numbers raise customers' reliance on cognition, leading to an increased purchase likelihood for products in cognitive contexts. The first four studies in Wadhwa & Zhang (2015) investigate this interaction between context and price roundedness and found this as a robust and consistent pattern of different experimental setups and dependent variables. In their fifth study, they propose a mechanism based on the regulatory-fit theory (Cesario & Higgins, 2008; Lee, Keller, & Sternthal, 2010) for the effect: When price roundedness and context create a fit (i.e. emotional context and rounded price, or cognitive context and non-rounded price, respectively) participants "feel right," which in turn increases their purchase likelihood. In Harms et al. (2018), we aimed at replicating the findings by means of a replication study. Previously, O'Donnell & Nelson (2015) tried to replicate study three from the paper. This is why the results of O'Donnell & Nelson (2015) will also be taken into account when comparing our replication with the original study.

As the proposed effect has real-world implications for marketers, we considered a replication to be relevant, helping to validate this previous finding. We chose the fifth study as the target for our replication as it included the setting for both the original "rounded price effect" and the proposed mediation through "feeling right" about the price, thus being the most complete study in the paper. This is also in line with the selection criteria used for the "Reproducibility Project: Psychology" (Open Science Collaboration, 2015).

3.3.1 Meta-analytical Review

As proposed above (see section 3.2), we first investigated the data from the original paper using meta-analytical techniques. This allowed us to assess any evidence for publication bias or file-drawered studies. In particular, the original study reported only significant findings for the interaction of interest, context $\times$ roundedness. The R-Index (R-Index.org, 2014; Schimmack, 2012, 2014) calculates the probability of observing such a pattern in a series of studies given the power of the studies.

3.3.1.1 *p*-Curve Analysis

Based on their findings regarding *p*-hacking and researcher degrees of freedom, Simonsohn et al. (2014) noticed that many reported studies have flat or left-skewed distributions of *p*-values,

which indicate biased results. Under the null hypothesis of no effect, p-values are uniformly distributed. Conversely, if the null hypothesis is false, the distribution of p-values is right-skewed, and the skewness is dictated by the power of the test (Sellke, Bayarri, & Berger, 2001). In order to investigate for biases, Simonsohn et al. (2014) proposed to plot binned p-values smaller than .05 and test for uniform distribution. This rests on the assumption that there is an incentive to report results with $p < .05$, i.e. that "p-hacking" is used to achieve significant results and that p-hacking is stopped as soon as the criterion is reached. There is evidence that p-values just below the .05 threshold indeed are more prevalent in literature than should be expected from statistical considerations (de Winter & Dodou, 2015; Lakens, 2015b, but also 2015a; Masicampo & Lalande, 2012). Conducting p-curve analysis aims at statistically testing for suspicious patterns that might indicate the presence of bias (i.e. publication bias, file drawer, and/or p-hacking), thereby invalidating the evidential value of the data. Simonsohn et al. (2015) extended the method by introducing the "half p-curve" to also account for one-sided tests.

In order to conduct p-curve analysis, focal hypothesis tests and their p-values are collected, plotted in bins, and tested for skewness and flatness, respectively. The p-curve analysis for Wadhwa & Zhang (2015) includes 14 statistically significant results at the .05 level. Testing for right-skewness, the analysis does not show evidential value in the data ($z = -1.02$, $p = .155$ for the full p-curve; $z = -0.56$, $p = .283$ for the half curve, i.e. only for results with $p < .025$). Testing for flatness, the result is also non-significant ($z = -1.23$, $p = .109$ for the full p-curve; $z = 2.51$, $p = .994$ for the half p-curve). The analysis is therefore inconclusive, possibly because of high heterogeneity among the studies included for the analysis. See figure 3.1 for the results.

While p-curve has been used to evaluate sets of studies (Bruns & Ioannidis, 2016; Cuddy, Schultz, & Fosse, 2018; Simmons & Simonsohn, 2017), some researchers have issued a more critical view. p-curve can also be used to estimate overall power for the set of studies (if the selected test statistics all measure the exact same effect). Considering it superior to p-curve, Schimmack & Brunner (2017) proposed z-curve as an alternative way to estimate power in a sample of studies. Notably, both approaches to estimating power focus on "observed power," which is seen as problematic (see below). Morey (2017) also voiced concerns about the selection of test statistics for p-curve. Further, if p-curve is used to estimate the true effect of a set of studies, heterogeneity might lead to an overestimation of the effect (van Aert, Wicherts, & van Assen, 2016). Recently, Erdfelder & Heck (2019) have presented scenarios that conflict with the design and purpose of p-curve, advising caution in the interpretation of results.

In summary, while we cannot conclude that p-hacking has led to the pattern reported in the article, we also cannot conclude that the data is without bias. On top of that, the distribution of p-values in figure 3.1 leaves us sceptic as it also does not indicate a high powered set of studies to find the effect under investigation. This underlines, in our view, the need for independent replication of the effect.

3.3.1.2 Replicability Index (R-Index)

As a second measure to investigate the results of several significance tests reported in a series of studies, we used the "R-Index" as introduced by Schimmack (2014) and Schimmack (2012). It is intended as a measure of replicability as it accounts for the median observed power in published studies and is sensitive to the use of research degrees of freedom and, most importantly, selective reporting (Schimmack, 2012). As the method has not yet been peer-reviewed, and as there are some concerns about using "observed power" for inferences (O'Keefe, 2007; Wagenmakers et al., 2015), the R-Index should be considered cautiously and merely as an indicator.

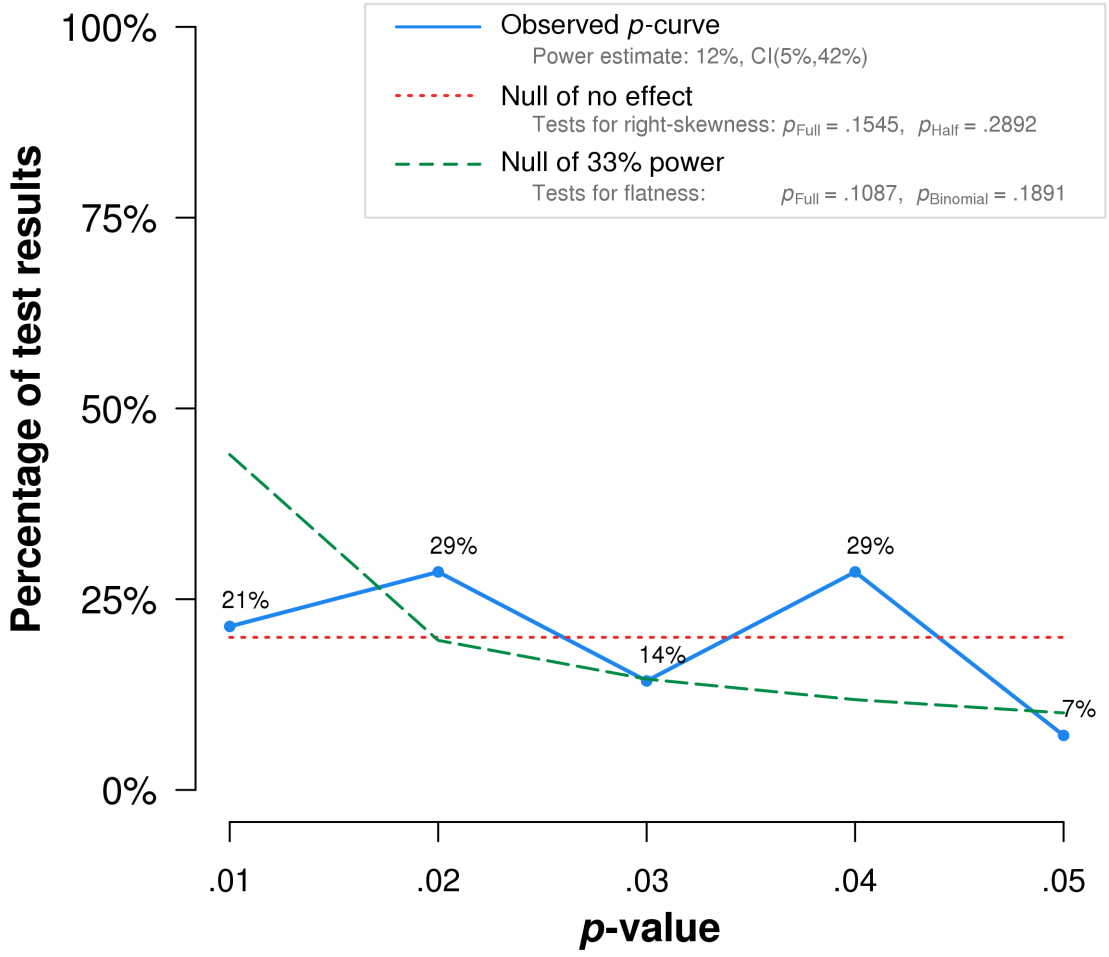


Figure 3.1: p-curve analysis from Harms et al. (2018, fig. 1, p. 4) not indicating evidential value in the data of Wadhwa & Zhang (2015). *Note:* the observed p-curve includes 14 statistically significant ($p < .05$) results, of which 8 are $p < .025$. There were 3 additional results entered but excluded from p-curve because they were $p > .05$. Disclosure table for selection of tests is available at osf.io/92t7s/.

We included the same tests as for the *p*-curve analysis and summarised the results for the five studies reported by Wadhwa & Zhang (2015) as follows (Harms et al., 2018, p. 4):

> With a proportion of significant results of 82.35% and a median observed power of 0.588, the R-Index for the paper by Wadhwa and Zhang is 0.352. Generally, the R-Index is closer to 1.0 for likelier patterns given the median observed power and smaller for unlikely numbers of significant results. The R-Index for the present case thus indicates that the proportion of 82.35% significant results (success rate) is rather unlikely given the median observed power of 0.588. Higher power or fewer significant results would have led to a higher R-Index.

Both, *p*-curve analysis and R-Index lead us to the conclusion that the results are unlikely to be unbiased. We cannot conclude that specific biases are the cause for the results, but it is known that publication bias, file drawer effect, and selective reporting are prevalent in psychological research. Overall, the results again show the need for independent replication, especially considering the real-world applications of the findings presented.

3.3.2 Methods

We aimed to replicate study five from Wadhwa & Zhang (2015) as it contained both the original "rounded price effect" and the proposed mediation through "feeling right." The study was conducted as an online study, the experiment programmed in SoSci Survey, and participants were recruited through several newsletters, posts in groups on social media websites (e.g. Facebook, Twitter), and mailings.

Based on the reported effect size for the interaction of interest ($\eta_p^2 = 0.0403$), 250 observations would have been sufficient to achieve 90% power at a level of significance of .05. As we were sceptical about the effect size, and to be in line with previous recommendations for replication studies (Asendorpf et al., 2013a; Brandt et al., 2014; Simonsohn, 2015), we aimed for 600 participants for the replication, which is near twice the sample size as in the original study. As the original study used a diverse sample of MTurk participants, we also aimed for recruiting a diverse German sample that does not comprise of only university students. After excluding participants who did not answer the priming questions diligently, our sample consists of $N = 588$ participants.

We reached out to the two original authors in order to receive feedback on the study plan and possibly receive the originally used materials and questions. For example, for the questions used for priming, only two examples were reported in the original article. Since we did not receive any feedback from the authors, we designed our own materials based on the description in the paper and in the cited works with slight adaptions to the German population, such as asking about "refugees" (a hotly debated topic at the time of the study) or "world cup champion," and translated them into German.

Before running the study, we conducted a pre-study among a sample of German psychology students in order to test the target product, the questions of the dependent variable, and whether the indicated price was perceived as reasonable. Most notably, the original study used "digital binoculars" (DB) as products, which are uncommon among the German target population. As an alternative product, we proposed a "digital instant camera" (DIC), which we expected to be more favourably rated among participants. Our pre-study revealed that both products are essentially understood, but the digital instant camera was rated better than the binoculars with respect to comprehension of the advertisement for the product (to be used in the focal study; $M_{DIC}\ (SD) = 6.50\ (2.063)$ on a 10-point scale compared to $M_{DB}\ (SD) = 5.20\ (1.969)$), comprehension of the product itself ($M_{DIC}\ (SD) = 6.41\ (2.207)$ compared to $M_{DB}\ (SD) = 4.98\ (2.005)$), and likelihood of purchase ($M_{DIC}\ (SD) = 4.86\ (2.347)$ compared to $M_{DB}\ (SD) = 2.14\ (1.229)$). Further, the willingness to pay was lower than the intended price

for both the binoculars (M_{DB} (SD) = 48.38 (51.337)) and the instant camera (M_{DIC} (SD) = 64.18 (49.095)).

The pre-study results led us to the decision to include the instant camera as a second product in addition to the product used in the original study (Harms et al., 2018, p. 6):

> Since we were aiming to conduct a close replication study, we decided not to deviate too much from the original setting and included both products, despite the poor ratings of the binocular camera. As we feared that variance would be too small for any meaningful analysis we also included the instant camera as a second product in the replication attempt. We present the binoculars first and the instant camera second for all participants—while this might introduce carry-over effects, we were still able to investigate the original effect for the original product without an influence of the second product.

We pre-registered the study with hypotheses and data analysis plan. The pre-registration is available as part of the repository at the Open Science Framework (OSF) accompanying our paper at osf.io/43u8d/. All materials, the questionnaire, the raw data, and the analysis scripts are available at osf.io/r942e/.

In line with both the pre-registration and step eight in the proposed guideline above, we first analysed the study traditionally using both frequentist statistics (classical two-way ANOVA with planned contrasts and moderated mediation analysis, Muller, Judd, & Yzerbyt, 2005) and Bayes factors (using the corresponding Bayesian variants of ANOVA (Rouder et al., 2012), *t*-test (Rouder et al., 2009), and mediation analysis (Nuijten, Wetzels, Matzke, Dolan, & Wagenmakers, 2015)). In order to evaluate whether our replication was successful or not, we compared the original study and our replication study using Simonsohn (2015)'s "small telescope" approach as well as the "Replication Bayes factor" for *t*- (Verhagen & Wagenmakers, 2014) and *F*-tests [see section 2.5.1; Harms (2018)].

While we tried to match the original study protocol as closely as possible, aiming for a direct replication, the need to design our own materials, the different population (diverse German online population versus American MTurk users) and the additional product leaves the replication attempt with some differences to the original study and should therefore be classified as a *close replication*, based on the taxonomy provided above (see section 3.1).

3.3.3 Results

First, we focus on the "rounded price effect" as is indicated by the interaction between context (invoked through priming) and price roundedness. As can be seen in figure 3.2, descriptively a similar trend emerges, with purchase likelihood being higher for the rounded price in the feeling condition. While this is the case for both products, the difference is much smaller for the digital binoculars. Notably, the average purchase likelihood is much smaller for the digital binoculars, and variance is also lower. In particular, the purchase likelihood is very similar, nearly identical, for both roundedness groups in the cognition condition.

Looking at the inferential statistics, the interaction between priming and roundedness is not significant for neither the digital binoculars ($F(1,584) = 0.608$, $p = 0.436$), nor the digital instant camera ($F(1,584) = 2.526$, $p = 0.113$). Bayes factors for the ANOVAs (Rouder et al., 2012) reveal evidence in favour of the intercept-only model compared to the full model containing main effects and the interaction, again for both the digital binoculars ($BF_{10} = 0.003 \pm 1.6\%$) and the digital instant camera ($BF_{10} = 0.039 \pm 2.2\%$). Running planned contrasts shows that the aforementioned difference for the *feeling* group for the digital instant camera is indeed significant ($t(584) = 2.322$, $p = 0.021$), but the data only show anecdotal evidence in favor of an effect ($BF_{10} = 1.653$). The results for the remaining contrasts as well as for all

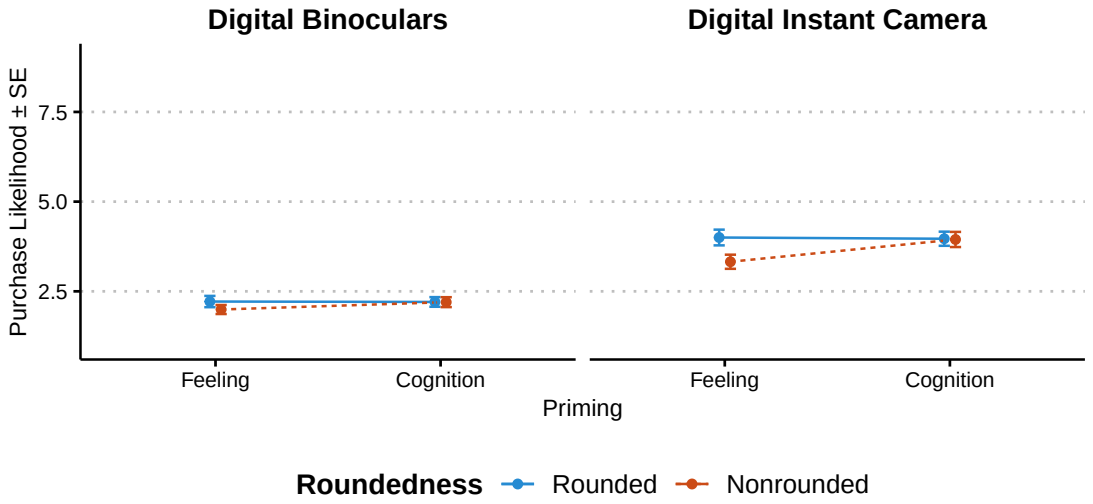
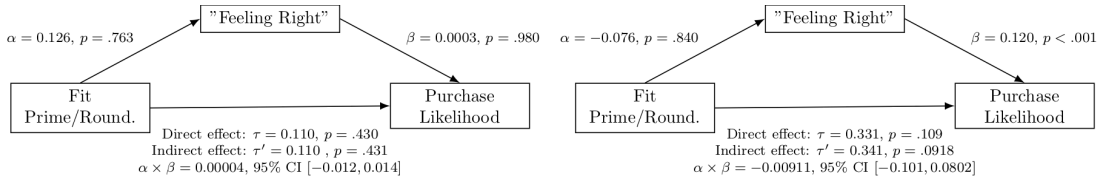


Figure 3.2: Primary results of the replication study. Purchase likelihoods for both products and different conditions. ANOVA shows no significant interaction effect for either effect with Bayes factors indicating evidence in favour of a point null result (see text for statistical results). *Note:* Figure is reproduced from Figure 2 in Harms et al. (2018).

contrasts for the digital binoculars are non-significant [see paper for detailed results; Harms et al. (2018), pp. 8-9].

Second, investigating the proposed moderated mediation through a sense of "feeling right" using Sobel's test yields non-significant results for both products (digital binoculars: $z = 0.025$, $p = .980$; digital instant camera: $z = -0.202$, $p = .840$; see also figure 3.3). Bootstrapping the indirect effect, as recommended by Preacher & Hayes (2008) yields confidence intervals which also include zero. Bayes factors for mediation (Nuijten et al., 2015) reveal strong evidence against an effect favoring the null models over the mediation models (digital binoculars: $BF_{m0} = 0.001$; digital instant camera: $BF_{m0} = 0.034$).



(a) Digital Camera Binoculars (b) Instant Camera

Figure 3.3: Results of the replication on the proposed moderated mediation. Mediation analyses did not show significant differences between direct and indirect effects. Bayes factors favoured the null models strongly over the proposed mediation models. *Note:* Figure reproduced from Figure 3 in Harms et al. (2018).

Overall, considering our replication study on its own, we find little to no evidence for either the rounded price effect or the proposed mediation. Directly comparing the original study to the replication study by converting the effect sizes to correlations r and testing the difference using a significance test for the difference of correlation coefficients, yields significant results for both the original product ($z = 2.45$, $p = .007$) and the additional product ($z = 1.971$,

$p = .024$). Our replication, therefore, suggests that there is a significant difference between the original and the replication study.[1]

It is still possible that the effect size estimated in our replication is in line with the results from the original study, considering the high uncertainty and low power associated with the sample from the original study. Simonsohn (2015) proposed to compare a replication effect size with the effect size that the original study had 33% power to detect ("small telescopes" approach). For study five from Wadhwa & Zhang (2015), this "small effect" would be $\eta^2_{(p;33\%)} = 0.012$. Figure 3.4 shows the effect sizes from all five original studies, the replication of study three by O'Donnell & Nelson (2015), and the present close replication of study five with a comparison to $\eta^2_{(p;33\%)}$. As can be seen, the replication is consistent with a small effect of 0.012. It is also notable that there is large uncertainty around the effect size estimates from the original study. All replications have a much larger sample size and much smaller effect size estimates than the original studies.

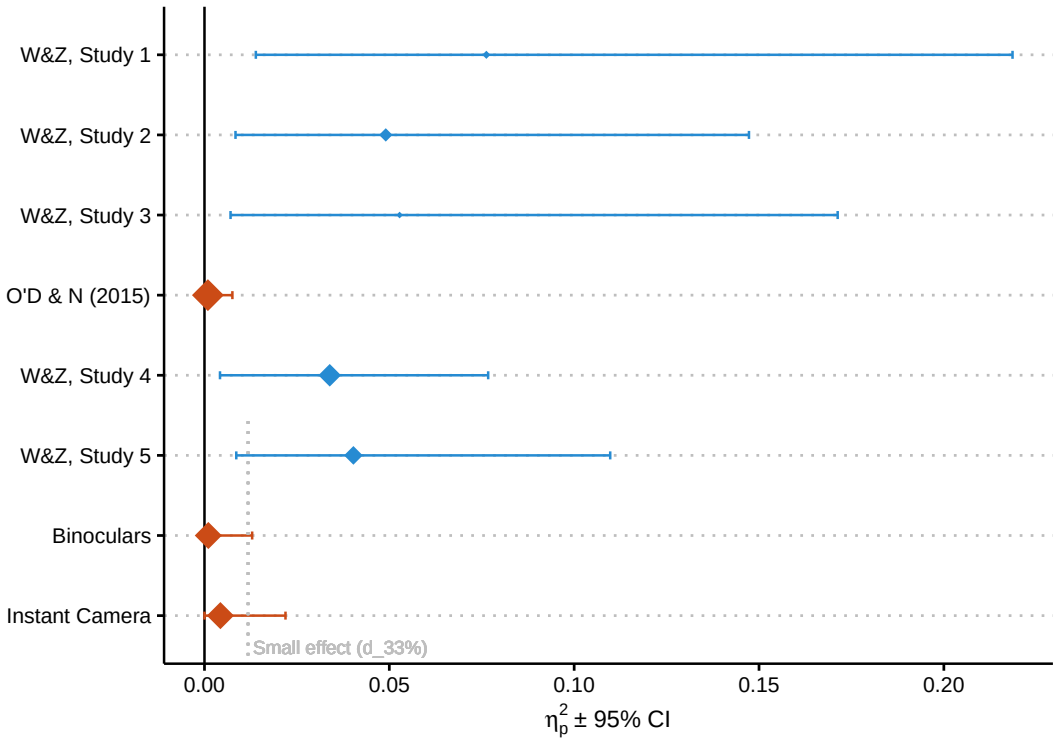


Figure 3.4: Comparison of effect size estimates for the "rounded price effect." Diamonds are point estimates for η^2_p for the interaction effect between context and roundedness with uncertainty intervals being 95% confidence intervals. Shape size is proportional to sample size. *Note:* Figure reproduced from Figure 4 in Harms et al. (2018).

Lastly, we used *replication Bayes factors* (Harms, 2018; Verhagen & Wagenmakers, 2014) in order to estimate the evidence against a successful replication (Harms et al., 2018, p. 10):

> The replication Bayes factor for the interaction term is $BF_{r0} = 0.009$ for the digital

[1] Note that this comparison was not pre-registered, not reported in Harms et al. (2018), and the replication not powered to test this difference. Error rates can therefore not be ensured for these tests and it is sensible to interpret the p-values in a Fisherian way.

camera binoculars and $BF_{r0} = 0.087$ for the instant camera, indicating strong evidence in favour of the null hypothesis for both products: That is, the model stating an effect size of $f^2 = 0$ for both studies is about 100 times and about 11 times, respectively, more likely than the model with the prior from the original experiment, under the assumption of balanced groups in both the original and the replication study. Considering the significant difference between rounded and non-rounded prices in the feeling condition for the instant camera, the replication Bayes factor is $BF_{r0} = 4.775$ and shows evidence in favour of a successful replication for this singular contrast.

3.3.4 Discussions and Limitations

In summary, our replication study did not unequivocally replicate the original findings. While some trends are in the same direction, the statistical results lead us to at least question the effect as it was stated in the original paper. In the terms proposed by LeBel et al. (2018), our replication therefore is an *inconsistent (smaller)* replication with *no signal*. Nevertheless, due to the high uncertainty in the original study, it is still consistent with a small effect of $\eta^2_{(p;33\%)} = 0.012$, which corresponds to a correlation coefficient of $r_{33\%} = 0.11$.

As with all studies, replication studies are also limited by several factors, especially when a detailed comparison with the original study is performed. Schmidt (2009) and Meehl (1990a) have argued about the difficulties of falsifying theories through experimentation and replication. Considering the experimental setup, especially the study materials as "auxiliary hypotheses" (Meehl, 1990a), our failed replication does not allow us to conclude whether the original study indeed is a "false positive," or whether we merely showed that it did not generalize to (a) other stimulus materials (items for priming, products, ...), (b) different populations and cultures, and/or (c) other combinations of products and populations. There are several more aspects of the study which might differ in detail, so that there may be many other explanations why the replication failed are possible. In practice, this leaves a lot of ground for original researchers to counter the criticism of a failed replication.[2]

While we aimed at following the guidelines proposed above, there are several shortcomings of the present replication: We contacted the original authors for their feedback and the original materials, but received those only one day after a first pre-print of the paper was posted online. Comparing the originally used materials with their description in the paper and with our study setup revealed several differences: Most notably, the items used for priming in the "feeling" condition not only contained words (as described in the paper) but also images corresponding to the words. One might argue that priming with images *and* words could lead to stronger priming, thereby increasing the average effect of the prime on the subsequent product evaluation. If raw data from the original study was available, one could include the data from all studies in a multilevel structure to better model the variability in methods to investigate generalisability, as suggested by Yarkoni (2019).

If one is to accept that our replication was indeed "close enough" in principle to the original study to be a *close* replication, the replcation's results question the originally published finding or – in a Bayesian way of reasoning – reduces our belief in the original finding. If one is to consider our study more to be a conceptual replication, since it involved different stimulus material, a different language, and a different population, our replication at least questions the generalisability of the findings made before. Both points can be made and are equally valid. At the same time, it shows that an unsuccessful close replication generates information that is relevant for the progress of science. If researchers wish to build on the findings by Wadhwa & Zhang (2015) in order to further investigate the relevance of pricing for product

[2]See some of these comments made in the openly available peer review comments to the article at the journal's website: https://royalsocietypublishing.org/doi/suppl/10.1098/rsos.171127.

marketing, our study – along with the previous replication by O'Donnell & Nelson (2015) – present first indications of boundary conditions that might be relevant for future research. In particular, product valence, willingness to pay versus prices in the study, and possibly invoking purchasing contexts through priming should be considered critically for any future study on this effect.

Chapter 4

Discussion and Outlook

Focusing on statistical practice in psychology in the context of the replication crisis, the previous chapters of this thesis aimed to give an overview of the replication crisis, potential causes, and proposed solutions. Chapter 1 gave a general overview of the emergence of the "replication crisis" and on how different publications and studies have led to the impression that the psychological discipline is in crisis. To find reasons for seemingly low rates of replications in psychology, chapter 2 took a look at the statistical practices in the field. Most importantly, frequentist and Bayesian statistics were contrasted with a focus on the different statistical questions they ask. Theoretical and epistemological foundations of different schools of thought were briefly described. Current practices in using statistical significance testing were shown to be problematic. To provide constructive perspectives, several statistical approaches were presented: Equivalence testing and Bayesian statistics in the context of null effects, and Replication Bayes factors for the evaluation of replication studies using Bayesian statistics. In chapter 3 a taxonomy of replication studies was developed based on similar, previous literature. Similarly, guidelines for conducting replication studies were proposed and a close replication study was used to exemplify the guidelines and discuss challenges in practice.

This thesis happens to be written in a time of increasing interest in meta-scientific questions, i.e. "How do we do good research?" Historically, this was primarily the domain of philosophers and historians of science. The current debate, however, spans from philosophers to applied researchers, from statisticians to clinical psychologists. The debate is more widespread than previously published and isolated criticisms and likely fuelled by the Internet and social media. Moreover, the debate on replicability in psychology is far from settled. New recommendations and initiatives are published regularly (both in journals and on academic blogs). So this work can only provide a snapshot of what has happened over the past years, taking a particular angle. This chapter summarises the major arguments from this thesis. The summary is presented along with phases of conducting a study, beginning with planning a study based on previous work or theory, touches the analysis, and finally the publication of a study in a journal article or preprint.

Science is a process of cumulative evidence synthesis. This is partly in contrast to the common practice of trusting a single paper to be corroboration or falsification of theories. While the last decade has seen an increase in the use of meta-analysis and systematic reviews, we still should move to a more holistic view on empirical results: A single study is rarely enough to judge the veracity of a theory or phenomenon. It is over the long-run, i.e. through multiple experiments, including close, direct, and conceptual replications as well as follow-up and contradicting studies, that science accumulates knowledge and allows for theory building, theory development, and corroboration or falsification. While Neyman-Pearson's approach to hypothesis testing is very much in line with this perspective (by following the strict rules

of hypothesis testing according to Neyman-Pearson we want to ensure that we do not make type I or type II errors too often in the long run), it also allows for a Bayesian way of scientific reasoning: Using incoming data, we update our beliefs about effects/theories (see also section 2.2.2). These are indeed different perspectives not only on the statistical analysis but also on the scientific mode of inference (Fidler et al., 2018).

4.1 Replication Studies

While replications have previously not been very common in published psychological research, the replication crisis certainly has increased the attention to this issue. Several initiatives have attempted to replicate bodies of empirical results in systematic replications, with the *Reproducibility Project: Psychology* (Open Science Collaboration, 2015) being the most prominent one. Munaf'o et al. (2017) and Zwaan et al. (2017) are two examples of the increasing demand for systematic replication studies even for individual published results. The silver lining and a central point of the present thesis: Replication studies in any form have crucial value to the progress of science. Even if they are not as valued in practice yet, there is a need to ensure that theories and findings are not a one-shot attempt at learning something about the human mind.

Academia offers many freedoms to researchers, but it is also an organisation with a certain incentive structure. Until recently, there have not been many incentives to perform replication research systematically and rigorously: There was little chance to publish replication if they were not used as a basis for additions to an existing theory. They were rarely seen as an essential part of the scientific progress, even when they were direct replications with a result contradicting previous research. The replication crisis has, in a first step, brought to light this lack of attention to replications. It has also increased interest in replication research. Preserving this momentum requires incentives for replications in the future. Since publications are the primary currency in an academic career, major outlets need to continue to promote replication research and offer a platform for rigorous replications. Funding agencies need also to fund research that at least starts with systematic replications before building upon previous work. Looking at the many cases where large research programmes have been under heavy attack from replications inconsistent with previous results (such as embodied cognition or behavioural priming), there is a clear societal and monetary benefit to fund replicable research and make successful replication a requirement for funding new research programmes.

A major criticism of replication research, especially since the replication crisis, has been that replications are used to discredit researchers and call into question long-standing research results. The only way to counter this criticism is by applying the same scientific and rigorous standards to replication research as to original research. Replications need to be performed systematically, based on well-known methods, and based on published recommendations on how to perform replications to be informative. They need to be scientific tools to maintain consistent and progressive research lines (Lakatos, 1970), and must not become political tools.

Recommendations: Replication studies need to become common practice in empirical psychological research. Publishers and funders will need to create incentives for performing replication research, so future research programmes are founded upon replicable, trustworthy research. Replication researchers need to be as rigorous as original researchers, so replications remain a scientific tool and not a political one.

4.2 Theory Building

Many research results in psychology are presented as isolated findings, often only loosely linked to previous research on single effects. An overarching theory or theoretical framework

is missing in many research areas. Thus, many results are explained in an ad-hoc manner (often fueled by building theories only after several different analyses are run on the same data; see *HARKing*). This makes it difficult for researchers to criticise and falsify such results – contradicting evidence from replication studies is then quickly integrated by using ad-hoc explanations such as the "hidden moderators" argument. If one considers such isolated phenomena with any explanation, loosely linked to previous research, as theory (a position one could also challenge), it is hardly testable or falsifiable. It lacks a proper definition of boundary cases and its limits of generalizability (Simons et al., 2017), inviting easy integration of any contradicting finding in the theory by quickly adding new explanations, variables, or influences to the theory, resulting in what philosopher Imre Lakatos named a "degenerate line of research" (Lakatos, 1970; see Meehl, 1990a, p. 111). This practice of theory "development" stifles a greater understanding in psychology and the development of theoretical frameworks as it happens more often in the natural sciences.

How do better theories need to look like? If one considers psychology to be quantifiable and measurable through statistical analysis, good theories should be presented as formal models. Oberauer & Lewandowsky (2019) argues, for example, that theories in psychology need to be presented as formal models by stating a set of mathematical equations or a path diagram. This not only allows for better and more falsifiable theories, but it also links the substantive theory with the statistical hypotheses, something often missing in psychological research (Meehl (1990a); cf. chapter 1.2). In similar ways, Fiedler (2017) also underlined the need for strong theory building. In his "theory-driven cumulative science," hypotheses also need to be derived from "incontestable laws and logical constraints" (Fiedler, 2017, p. 53). Such theories can make testable predictions and formulate boundary conditions, that – when met – would falsify the theory. As has been argued before, such predictions can then be tested by formal hypothesis testing, using either Neyman-Pearson's significance testing or Bayes factors. In contrast to current practices, such hypothesis testing would be more informative about the state of the theory than blindly testing point-null hypotheses that rarely inform theory apart from "something is going on." Tukey (1969, p. 86) already contrasted psychology's favour of underspecified theories to the approach in other scientific fields:

> The physical sciences have learned much by storing up amounts, not just directions. If, for example, elasticity had been confined to "When you pull on it, it gets longer 1" Hooke's law, the elastic limit, plasticity, and many other important topics could not have appeared. [...] Bear in mind a simple fact: the great majority of the useful facts that physics has learned — and recorded in numbers — are specific and detailed, not global and general.

Such an approach to theory building and testing seems to be easier to sub-fields such as cognitive psychology with a history rooted in the mathematically founded psychophysics. In other fields, it might be harder to technically specify latent constructs and quantify them in the context of a theory. Psychometrics is already well equipped for addressing this question but is rarely used in practice for the "soft areas" of psychology (in Meehl's sense). Perspectives such as evolutionary theory could provide a basis for such a framework, as Muthukrishna & Henrich (2019) or Tybur & Navarette (2017) have argued. Building such an overarching theoretical framework certainly is a challenge for psychology due to the dynamic and complex nature of the human mind (cf. Green, 2015).

A more theory-focused approach would also require more collaboration between teams of researchers: collaboration both across sub-disciplines of psychology and across geographical and cultural borders. A more collaborative way of psychological research is necessary to integrate findings, perspectives, or theories from different research teams, but also to have more culturally diverse samples right from the beginning. Most of psychology is primarily concerned with western culture and society, making replications in different countries easily dismissible on the ground of culture alone (Henrich et al., 2010; Montag, 2018). The need for

more international collaboration between research groups is another point, Tukey (1969, p. 89) already noted:

> If, as T would guess, some areas of psychology have reached the point where they need larger, more diverse bodies of data measured in common ways, there may well be no way out other than cooperation, cooperation both among thesis supervisors and among thesis workers, cooperation over large distances, whether in geography or in a subject's background.

Luckily, the Internet, social media, and international conferences make it easier to build research collaborations spanning different countries and continents today than ever before. Looking also at other natural sciences: the Large Hadron Collider at CERN would not have been possible without the investment of many different organisations and researchers from around the world. This creates not only more cultural diversity in research, but also allows to build teams of researchers from different scholarly perspectives or even invites "adversarial collaborations" between researchers holding different beliefs about effects, explanations, or theories (Kahneman & Klein, 2009; Mellers, Hertwig, & Kahneman, 2001).

The lack of theory building and testing has already invited the question whether psychology also has a "theory crisis" (Oberauer & Lewandowsky, 2019). As Fiedler (2017) repeatedly notes: We lack good, falsifiable theories in many parts of psychology. Moreover, such theories need to be more carefully translated inro defensible statistical models or be synonymous with them (Meehl, 1990a; Oberauer & Lewandowsky, 2019). Improving the theory-foundedness of psychology, in general, will also likely improve replicability: Instead of integrating contradicting results from replication studies through ad-hoc explanations, a more rigorous theory development would mean to ask whether a replication falsifies a previous theory or whther it can be integrated into the theory without changing its core (Lakatos, 1970; Meehl, 1990a).

Recommendations: Psychology needs theories with clearly stated boundary conditions that are falsifiable by replication. Ideally, such theories are formulated as formal models explicating constructs, measurements, and relationships. Building, developing, and extending theories in this way allows one to close the gap between theories and statistical methods and can improve the value of replication as a critical part of the process of theory development.

4.3 Pre-Registration and Open Science

Pre-registration has been discussed as one of the most important changes to current practices that might lead to improved replicability of research (Munaf'o et al., 2017; Nosek et al., 2018; Wagenmakers et al., 2012). Notably, pre-registration can reduce biases introduced by data-dependent analysis, post-hoc reasoning, *p*-hacking, and many questionable research practices. On top of that, registered reports have been suggested as a journal article format that incorporates peer review before data is collected by reviewing the pre-registered analysis plan (see below). Not all types of studies, however, can or should be pre-registered: Pre-registration is useful for confirmatory studies in which statistical procedures can be fixed *a priori* as they are demanded by theory or previous work. Exploratory studies in which no prediction for the outcome can be made and whose goal is to explore patterns in the data are an important part of theory development. While pre-registration is not a necessary condition for good research, it can, nonetheless, provide readers with information on why and how the study was designed and what guidelines guided the exploratory analysis. It is important to underline that pre-registrations do not prohibit exploratory analysis or research in general. It is rather a tool to properly distinguish and communicate this distinction between exploratory and confirmatory research (Wagenmakers et al., 2012).

Open Science principles have gained popularity in recent years, not only through the availability of online platforms like the Open Science Framework or GitHub. It complements pre-registration by allowing researchers to make pre-registered materials publicly available (Open Material). Reproducibility can also be improved by sharing data after publication (Open Data). More generally, providing public access to journal articles allows for communicating scientific progress more widely, especially in less privileged regions. This is related to an issue that was only briefly touched in this thesis: As long as there is an economic niche for scientific publishers to gain from "fancy theories" with a high risk of being non-replicable, there is both a scientific and an economic incentive to conduct studies that are not founded in previous theory. Removing this incentive and moving the economic gains from scientific work back in the hands of the public can help to emphasise and incentivise replicable and robust research. Open Access publishing is one way to do this, but it often requires authors to pay a fee for their submission. This makes it harder for early career researchers, researchers from developing regions, or researchers in small fields with little funding to publish their results.

Recommendations: Pre-registration is an important tool to make the empirical research process more transparent. It is not a panacea for any kind of research, but especially relevant to confirmatory studies. Exploratory research is as valuable as confirmatory research but also needs to be transparent to invite confirmatory replications and solid theory building.

4.4 Statistical Practice

The statistical practice especially in replication research was a focus of this thesis. As previous authors have repeatedly noted, there is a gap between statistical theory and its application in practice. When using significance testing, for example, the requirements and assumptions must be met. It is therefore important for researchers to understand the differences between Fisherian and Neyman-Pearsonian significance testing and using these tools in a principled way – in contrast to the "null ritual" (Gigerenzer, 2004). Whether significance levels are set at .05, .005 per default (Benjamin et al., 2018), or justified for any particular study individually (Lakens et al., 2018), is then – while being a relevant debate for practitioners – a secondary issue as long as positions can be justified statistically in a principled way.

However, in many lines of research, where theories are not yet as mature as to make a precise point or range predictions, corroboration or falsification through significance testing might not yet be a sensible goal. To understand phenomena better and to find possible boundary conditions for an effect, estimating statistical parameters can be more fruitful than trying to test a particular statistical hypothesis. Bayesian statistics, as outlined in section 2.2, provides a suitable framework for estimating parameters. Ordinary least squares or maximum likelihood estimation are similar procedures to estimate parameters but do not offer the benefits of Bayesian statistics. However, the maximum likelihood estimate is a special case of the Bayesian estimate in which the prior is flat. More general, Bayesian estimation can be understood as "maximum penalized likelihood" where the prior is a penalty function across the parameter space. Such penalty is useful when external information should be included in the statistical inference (encoded as a probability distribution in the prior) or when predictive accuracy is a concern to avoid overfitting (especially with small samples).[1] There are several methods for penalising or regularising likelihoods (Ridge regression or Elastic Net Regression to name two often used in applied settings). Bayesian statistics, on the other hand, is well suited for a wide range of complex models, is very flexible, and offers an intuitive understanding of the probabilistic results.

[1] While not a particular concern in this thesis, a consequence of the previous recommendation to use formal models to make predictions, which can be tested with empirical data, is the stronger emphasis on prediction in statistical models. A good theory might be a model that makes accurate predictions about human behaviour or responses.

As discussed before, the debate which framework is "better" has been ongoing for decades and the present thesis does not aim to settle this debate. From a practical point of view, Bayesian statistics offer an additional, flexible, and intuitive way to perform statistical analyses. In practice, however, researchers do not need to use a single approach for each and any situation. Instead, by using different methods side-by-side, a better understanding of the data can be achieved – as long as the methods are used correctly and interpretations are sensible. Besides testing through p-values or Bayes factors, statistical modelling offers an additional way to map theories to statistical terms. Statistical models, and Bayesian models, in particular, can incorporate structural knowledge about the process generating the data collection ("generative models"). Such models can then be trained on observed data and tested against past or new observations. Setting up a model requires more expertise than running a standard hypothesis test or calculating a default Bayes factor, but fundamentally improves a researcher's capabilities to make inferences from the data. Statistical modelling also has its controversies about best practices. For example, there is an ongoing debate about how to assess the fit of models (Marsh, Hau, & Wen, 2004) and how to compare models using information criteria (Dziak et al., 2018). Gelman et al. (2013) outline a general methodology for setting up probability models, evaluating them, and interpreting results in the context of a deductive Bayesian framework (see also Gelman & Shalizi, 2013).

Statistical techniques are developing continuously and the increased attention to the field of "machine learning" has further sparked the development of new methods. While the machine learning community has its problems with replicability too (Beam, Manrai, & Ghassemi, 2020; Bouckaert, 2005; Hutson, 2018), it is also an open question whether new tools offer the benefit they promise (see also Haverkamp & Beauducel, 2017 for problems with new methods in classical statistics). In essence, researchers should not apply methods "mindlessly," no matter whether it has been a common method for a long time (such as significance testing) or whether it is a new method recently introduced. With a broader variety of statistical tools, it is nevertheless crucial that these tools are used sensibly and correctly. If applied mindlessly it will lead to spurious and unreplicable results, no matter whether significance testing or statistical modelling was used. Being at the centre of this thesis, improving statistical practices by making correct use of the statistical toolbox should improve replicability by providing more robust results and allowing for a more sensible and nuanced interpretation of statistical inferences.

Recommendations: Statistical methods need to be adopted more rigorously and appropriately to the statistical questions at hand. Null-hypothesis significance testing in the Neyman-Pearson framework is a well-established tool for controlling error rates in repeated sampling. Additional statistical tools such as statistical modelling in a Bayesian statistical framework can provide more flexible approaches to investigate theories and phenomena in data. Psychologists might benefit from working closely with trained statisticians in their research, much like it is already common practice in medical research.

4.5 Publications and Peer Review

While pre-registration can easily be used for any study, irrespective of where it gets published, "registered reports" (RR) are an extension to this idea applied to the peer review process (Chambers et al., 2015; Nosek & Lakens, 2014). Since a pre-registration would usually only be included in the methods section of a manuscript it would be reviewed in combination with the outcome. This can easily introduce bias and could lead to a results-dependent interpretation of the pre-registration. When submitting a study as a registered report, the study is submitted as "stage 1 manuscript" before data is collected. This effectively is a review of a pre-registered study protocol. Reviewers provide their comments and recommendations on the protocol without consideration of actual results, thereby prohibiting any outcome-based criticisms.

After reviewing and possibly adapting the protocol according to the reviewers' suggestions, the manuscript can become "accepted in principle": If the study is carried out according to the pre-registered and accepted protocol, the study will be published irrespective of its results. In the second round of peer review, the reviewers have to check whether the study adhered to the registered protocol and that the conclusions are sound. Minor deviations are usually accepted as long as they are reported transparently.

Pre-registrations and registered reports are by no means a panacea for the replication crisis. However, they do improve the academic process by providing incentives for good scientific practice. In particular, registered reports provide certainty for both readers (higher trustworthiness of results due to RR process, pre-registered and reviewed study protocol) and authors (certainty of publishing irrespective of outcome). Whether registered reports do provide better scientific quality compared to traditional ways of publishing is a matter of ongoing meta-scientific research (Chambers & Mellor, 2018; T. E. Hardwicke & Ioannidis, 2018; Obels, Lakens, Coles, & Gottfried, 2019).

Another trend over the past years is the increase of pre-prints: papers' manuscript versions which are posted online in repositories like arXiv or PsyArXiv. While not peer-reviewed, they represent the latest research in raw form. As traditional peer review can, depending on field and journal, take several months to years until a paper is finally available, pre-prints allow researchers to share their findings and studies much earlier. It is, however, a double-edged sword: During the early phases of the COVID-19 pandemic, pre-prints were the fastest way to disseminate scientific results about the virus across the world. This also allowed badly performed studies to be presented alongside high-quality research, without any indication for non-experts what is what. Journal publication on its own, however, did not solve this problem: There have been also several journal publications that had later to be corrected or retracted (Retraction Watch, 2020). There are only a few journals who prohibit pre-prints generally. Most journals allow at least non-reviewed versions of the manuscripts to be stored online.[2]

Peer review in its current form has been criticised (a Teixeira da Silva & Dobr'anszki, 2015; Casnici, Grimaldo, Gilbert, Dondio, & Squazzoni, 2017; J. R. Gilbert, 1994; Mahoney, 1977) but remains one of the cornerstones of scientific publishing. Initiatives such as the *Peer Reviewers Openness initiative* (Morey et al., 2016) and journal policies as in the *Royal Society Open Science* have aimed to improve the review process by incentivising open science practices and making the review process itself open by posting reviewer comments. Other forms of publishing have also emerged: For example, *F1000Research* combines the notion of pre-prints and traditional review by allowing papers to be posted without review (like pre-prints) and have a post-publication open review process after uploading the paper. This allows the research to be online right after the manuscript is completed and to have transparency on reviewers' comments and how they have impacted the paper.

Pre-registration, registered reports, pre-prints, open peer review: All these attempts to improve the process of scientific publishing aim to make science more open and more transparent. Providing recipients of research with a more transparent and more honest picture of how the research was conducted and how the paper came into the journal, allows for a more critical evaluation of a study. Furthermore, by reducing publication bias and post-hoc reasoning, a more realistic picture of research programmes can be drawn. Registered reports, for example, have higher rates of null results compared to the literature in general. Published research need to present a realistic view on a scientific field, so that members of the field, as well as outsiders, can gauge the veracity of findings, and, in the words of Lakatos, detect whether a line of research is healthy or degenerate. Improving scientific publishing is another important aspect to improving replicability in both psychology and science in general.

Recommendations: Peer review has been a cornerstone of the scientific publishing system,

[2]The SHeRPa website provides a list of journal policies at http://www.sherpa.ac.uk/romeo/search.php.

but problems with this approach have been known for quite some time. The availability of pre-print platforms allows for fast dissemination of research results and invites pre- or post-publication peer review. Open and non-anonymous peer review might prove its benefits, but some caveats remain. Further meta-scientific research on peer review is necessary.

4.6 The Way Forward

As being outlined in the introduction of this thesis, the crisis has spread across all phases of the scientific process. Investigating problematic practices or incentive structures and their effects on replicability or trustworthiness are large meta-scientific research programmes. They will keep researchers busy for a long time. Furthermore, this "crisis" is not over and the discussion on how psychologists should work in future is still ongoing. Several attempts and initiatives have started to change problematic practices – either by raising awareness among students and researchers or by introducing new tools. The rapid developments and changes are fuelled by increasing speed in the dissemination and communication of scientific and meta-scientific results. It is arguably hard for researchers, working on their research, to keep up with all these developments. During the time this thesis is being reviewed and published, new initiatives and new proposals published to improve psychology as a science are likely to be published. Time, meta-scientific research, and practice will tell which of those initiatives were successful and which have actually improved replicability and trustworthiness. The history of psychology as well as debates among philosophers of science shows that the discussion resulting from the "replication crisis" will never be completely "done." As it has been shown in chapters 1 and 2, criticisms of low power, weak theory, and misuse of statistics are older than current findings on low replicability in psychology. While there are some signs to be optimistic about changes in common practice, constant re-evaluation and further research on meta-science are critical.

There is certainly enough reason to be optimistic about psychological science. On the other hand, the replication crisis has shed light on those areas of science in which improvements are required. In psychology, two of these areas are measurement and theory development. Both have been increasingly criticised in the last years and have also been named causes for the replication crisis. Some authors have already even talked about a looming "measurement crisis" (Loken & Gelman, 2017), "theory crisis" (Oberauer & Lewandowsky, 2019), or "generalisability crisis" (Yarkoni, 2019) in psychology. The discussion about these problems is already ongoing, and based on the findings so far, there is reason to be cautious and to continue questioning common practices and methods. This ongoing debate, however, is at the core of science being self-correcting and not a cause for concern.

More fundamental criticism on the nature of psychological research has been voiced in the context of the replication crisis (e.g. Wolfgang Stroebe & Strack, 2014) as well as also prior to the crises, especially by Gergen (1973), Gergen (1985), and Lee J. Cronbach (1957). Even if these authors are well-known in psychology, their perspective has not garnered widespread agreement in mainstream psychological research. While some areas of psychology have their history in philosophy with respect to the treatment of the human mind, academic psychology has mostly focused their methods on empirical research. The empirical approach has been quite successful by investigating psychological phenomena and developing psychological theories. Psychology's development towards an empirical science parallels the development of the role of science in our modern, western societies: There is a strong emphasis on *utility* of scientific research – and psychology is no exception to that. There is an increasing societal demand for psychological explanations and psychotherapeutic interventions. To warrant such interventions against misuses and charlatans, rigorous research on efficacy and benefit is required.

Our societies are becoming increasingly "data-driven," in that empirical evidence is becoming

necessary for many parts of our society. Such need for data and empirical evidence arising from a societal trend is not an argument for universal truth, though: Society's focus on data and empirical research cannot establish that an empirical approach to psychology is, in fact, the "true" or "correct" way to do psychology. Empirical psychology certainly had its successes both in fundamental and in applied research. Such successes, however, stem from diversity in competing perspectives, methods, and theories. Less empirical perspectives, like Gergen (1973)'s, are valuable and necessary to build upon. In contrast to Gergen, Schlenker (1974), for example, has argued that there is no logical or philosophical difference between social and natural sciences: Investigating humans in social environments is certainly harder than observing rocks falling due to gravitation. He argues, however, that this simplistic view of physics falls short of the challenges and the effort required to address these in the natural sciences.

No scientific field is ever "finished," neither in its subject of research nor the development of its methods. As philosophers of science have already realised, any *crisis* can help to shake up a scientific field. It also helps researchers to improve their scientific approach, leading to more and improved knowledge. During the past one and a half decade, efforts to improve methods in psychology have increased and habe introduced pre-registration, registered reports, new theories, new guidelines for theory development, and a new focus on the use of statistics in empirical studies. By reviewing and replicating past studies, a more consistent body of research could be established. The efforts are still ongoing and have sparked new interest in meta-science leading to the formation of several research groups specifically working on how we do science in psychology. One indication for being optimistic about the long-lasting impact of these reforms is the fact that a majority of researchers working on these questions are young researchers who are early in their career (early career researchers). If the academic system does not drive them out, they will be able to further develop their ideas and ambitions for rigorous science and later pass it on to their students.

Science might not be, per se, a self-correcting endeavour (W. Stroebe, Postmes, & Spears, 2012), but an intensive and ongoing debate on methods will most certainly help to spark new and innovative ideas. Not all of them will work out as expected. There will be failures: While some researchers will adopt a new method, other scientists will disagree or argue that this new method does not keep up to its promise. But the back-and-forth between different groups of researchers is what enables creativity, innovation, and what generates new ideas. This is a necessary ingredient to science. Time, and most importantly, a repeated debate is necessary. This is best illustrated by the fact that several papers on the replication crisis cite arguments and similar discussions from the 19th century as well as the 1950s, or 1980s, et cetera.

In this sense, replication and reproduction are not only a necessary scientific tool to validate empirical claims. They might even represent the way science needs to operate in general. Sampling repeatedly from a population is the key to statistical inference. Running the same or similar experiments repeatedly is the key to robust empirical findings. Repeatedly challenging theories and findings is the key to theory development. Repeatedly discussing how we should work as scientists and how science should be done is the key to improving science in general.

References

a Teixeira da Silva, J., & Dobr'anszki, J. (2015). Problems with Traditional Science Publishing and Finding a Wider Niche for Post-Publication Peer Review. *Accountability in Research*, *22*(1), 22–40. https://doi.org/10.1080/08989621.2014.899909

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *Bmj*, *311*(7003), 485–485. https://doi.org/10.1136/bmj.311.7003.485

Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature*, *567*, 305–307.

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat (p $>$ 0.05): Significance thresholds and the crisis of unreplicable research. *PeerJ*, *5*, e3544. https://doi.org/10.7717/peerj.3544

Anderson, C. J., pan Bahnik, t, Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., … Zuni, K. (2016). Response to Comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad9163

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1–12. https://doi.org/10.1037/met0000051

Angrist, J. D., & Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30. https://doi.org/10.1257/jep.24.2.3

Anvari, F., & Lakens, D. (2020). *Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest*.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013a). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013b). Replication is more than hitting the lottery twice. *European Journal of Personality*, *27*(2), 138–144.

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

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action. *Journal of Personality and Social Psychology*, *71*(2), 230–244. https://doi.org/10.1037/0022-3514.71.2.230

Bayarri, M. J., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2015). Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses. *Journal of Mathematical Psychology*, *72*, 90–103. https://doi.org/10.1016/j.jmp.2015.12.007

Beam, A. L., Manrai, A. K., & Ghassemi, M. (2020). Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA*, *323*(4), 305. https://doi.org/10.1001/jama.2019.20866

Bem, D. J. (2002). Writing the Empirical Journal Article. In J. M. Darley, M. P. Zanna, & H. L. Roediger II (Eds.), *The Compleat Academic: A Career Guide*. American Psychological Association.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influ-

ences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Bem, D., Tressoldi, P., Rabeyron, T., & Duggan, M. (2016). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Research*, *4*. https://doi.org/10.12688/f1000research.7177.2

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Berger, J. O., & Berry, D. A. (1988). Statistical Analysis and the Illusion of Objectivity. *American Scientist*, *76*(2), 159–165.

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, *66*, 115–133. https://doi.org/10.1016/j.cortex.2014.11.019

Bos, C. S. (2002). A Comparison of Marginal Likelihood Computation Methods. In *Compstat* (pp. 111–116). https://doi.org/10.1007/978-3-642-57489-4/_11

Bouckaert, R. R. (2005). Low Replicability of Machine Learning Experiments is not a Small Data Set Phenomenon. *Proceedings of the ICML-2005 Workshop on Meta-learning*, 8. Bonn, Germany.

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201–236). Academic Press.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., … van 't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*(1), 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, *84*(4), 822–848. https://doi.org/10.1037/0022-3514.84.4.822

Bruns, S. B., & Ioannidis, J. P. A. (2016). P-Curve and p-Hacking in Observational Research. *PLoS ONE*, *11*(2), e0149144. https://doi.org/10.1371/journal.pone.0149144

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munaf'o, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-018-0399-z

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*, 823. https://doi.org/10.3389/fpsyg.2014.00823

Casnici, N., Grimaldo, F., Gilbert, N., Dondio, P., & Squazzoni, F. (2017). Assessing peer review by gauging the fate of rejected manuscripts: The case of the Journal of Artificial Societies and Social Simulation. *Scientometrics*, 1–14. https://doi.org/10.1007/s11192-017-2241-1

Cawley, G. C., & Talbot, N. L. C. (2007). Preventing Over-Fitting during Model Selection via Bayesian Regularisation of the Hyper-Parameters. *Journal of Machine Learning Research*, *8*, 841–861.

Cesario, J., & Higgins, E. T. (2008). Making Message Recipients "Feel Right." *Psychological Science*, *19*(5), 415–420. https://doi.org/10.1111/j.1467-9280.2008.02102.x

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex*, *66*, A1–a2. https://doi.org/10.1016/j.cortex.2015.03.022

Chambers, C. D., & Mellor, D. T. (2018). Protocol transparency is vital for registered reports. *Nature Human Behaviour*, *2*(11), 791–792. https://doi.org/10.1038/s41562-018-0449-6

Chartier, C. R., Riegelman, A., & McCarthy, R. J. (2018). StudySwap: A Platform for Interlab Replication, Collaboration, and Resource Exchange. *Advances in Methods and Practices in Psychological Science*, *1*(4), 574–579. https://doi.org/10.1177/2515245918808767

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., & Lakens, D. (2018). The costs and benefits of replication studies. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/s0140525x18000596

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, *1*(3), 140216–140216. https://doi.org/10.1098/rsos.140216

Cronbach, Lee J. (1957). The Two Disciplines of Scientific Psychology. *American Psychologist*, *12*(11), 671–684.

Cronbach, Lee J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*(2), 116–127. https://doi.org/10.1037/h0076829

Crüwell, S., van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., … Schulte-Mecklenbeck, M. (2019). Seven Easy Steps to Open Science: An Annotated Reading List. *Zeitschrift für Psychologie*, *227*(4), 237–248. https://doi.org/10.1027/2151-2604/a000387

Cuddy, A. J. C., Schultz, S. J., & Fosse, N. E. (2018). P-Curving a More Comprehensive Body of Research on Postural Feedback Reveals Clear Evidential Value for Power-Posing Effects: Reply to Simmons and Simonsohn (2017). *Psychological Science*, *29*(4), 656–666. https://doi.org/10.1177/0956797617746749

Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Cumming, G., & Finch, S. (2001). A Primer on the Understanding, Use, and Calculation of Confidence Intervals that are Based on Central and Noncentral Distributions. *Educational and Psychological Measurement*, *61*(4), 532–574. https://doi.org/10.1177/0013164401614002

de Groot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioral sciences.* The Hague: Mouton.

de Winter, J. C., & Dodou, D. (2015). A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ*, *3*, e733. https://doi.org/10.7717/peerj.733

Dienes, Z. (2008). *Understanding Psychology as a Science: An introduction to scientific and statistical inference.* Palgrave Macmillan.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral Priming: It's all in the Mind, but Whose Mind? *PLoS ONE*, *7*(1), e29081. https://doi.org/10.1371/journal.pone.0029081

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2018). *Sensitivity and Specificity of Information Criteria* [Preprint]. https://doi.org/10.1101/449751

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. https://doi.org/10.1037/h0044139

Elderton, W. P. (1902). Tables For Testing The Goodness of Fit of Theory to Observation.

Biometrika, *1*(2), 155–163.

Elms, A. C. (1975). The Crisis of Confidence in Social Psychology. *American Psychologist*, 967–976.

Erdfelder, E., & Heck, D. W. (2019). Detecting Evidential Value and p-Hacking With the p-Curve Tool: A Word of Caution. *Zeitschrift für Psychologie*, *227*(4), 249–260. https://doi.org/10.1027/2151-2604/a000383

Evanschitzky, H., Baumgarth, C., Hubbard, R., & Armstrong, J. S. (2007). Replication research's disturbing trend. *Journal of Business Research*, *60*(4), 411–415. https://doi.org/10.1016/j.jbusres.2006.12.003

Fahrmeir, L., Kneib, T., & Konrath, S. (2010). Bayesian regularisation in structured additive regression: A unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, *20*(2), 203–219. https://doi.org/10.1007/s11222-009-9158-3

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

Faust, D., & Meehl, P. E. (2002). Using Meta-Scientific Studies to Clarify or Resolve Questions in the Philosophy and History of Science. *Philosophy of Science*, *69*(S3), S185–s196. https://doi.org/10.1086/341845

Feest, U. (2018). Why Replication is Overrated. *PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association*. Seattle.

Fidler, F., Thorn, F. S., Barnett, A., Kambouris, S., & Kruger, A. (2018). The Epistemic Importance of Establishing the Absence of an Effect. *Advances in Methods and Practices in Psychological Science*, *1*(2), 237–244. https://doi.org/10.1177/2515245918770407

Fiedler, K. (2017). What Constitutes Strong Psychological Science? The (Neglected) Role of Diagnosticity and A Priori Theorizing. *Perspectives on Psychological Science*, *12*(1), 46–61. https://doi.org/10.1177/1745691616654458

Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Field, S. M., Hoekstra, R., Bringmann, L., & Van Ravenzwaaij, D. (2019). When and Why to Replicate: As Easy as 1, 2, 3? *Collabra: Psychology*, *5*(1), 46. https://doi.org/10.1525/collabra.218

Fisher, R. A. (1950). *Statistical methods for research workers* (11th ed., rev.). Edinburgh: Oliver & Boyd.

Fisher, R. A. (1955). Statistical Methods and Scientific Induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, *17*(1), 69–78.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd.

Gamerman, D., & Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (Third Edit). Boca Raton, FL: CRC Press.

Gelman, A., & Hennig, C. (2015). *Beyond subjective and objective in statistics*. Retrieved from https://arxiv.org/abs/1508.05453

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis*.

Gelman, A., & Robert, C. P. (2013). "Not Only Defended But Also Applied": The Perceived Absurdity of Bayesian Inference. *The American Statistician*, *67*(1), 1–5. https://doi.org/10.1080/00031305.2013.760987

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British*

Journal of Mathematical and Statistical Psychology, *66*(1), 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Gelman, A., Simpson, D., & Betancourt, M. (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, *19*(10), 555. https://doi.org/10.3390/e19100555

Gelman, A., & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006x152649

Gelman, A., & Yao, Y. (2020). *Holes in Bayesian Statistics* (p. 11).

Gergen, K. J. (1973). Social Psychology as history. *Journal of Personality and Social Psychology*, *26*(2), 309–320. https://doi.org/10.1037/h0034436

Gergen, K. J. (1985). The Social Constructionist Movement in Modern Psychology. *American Psychologist*, *40*(3), 266–275.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016a). *A Response to the Reply to our Technical Comment on "Estimating the Reproducibility of Psychological Science"*. Harvard University; University of Virginia.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016b). Comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad7243

Gilbert, J. R. (1994). Is There Gender Bias in JAMA's Peer Review Process? *JAMA: The Journal of the American Medical Association*, *272*(2), 139. https://doi.org/10.1001/jama.1994.03520020065018

Glöckner, A., Fiedler, S., & Renkewitz, F. (2018). Belastbare und effiziente Wissenschaft. *Psychologische Rundschau*, *69*(1), 22–36. https://doi.org/10.1026/0033-3042/a000384

Gomatam, R. V. (2012). How do classical and quantum probabilities differ? *AIP Conference Proceedings*, *1424*(1), 105–110. https://doi.org/10.1063/1.3688958

Gorard, S. (2010). All evidence is equal: The flaw in statistical reasoning. *Oxford Review of Education*, *36*(1), 63–77. https://doi.org/10.1080/03054980903518928

Green, C. D. (2015). Why psychology isn't unified, and probably never will be. *Review of General Psychology*, *19*(3), 207–214. https://doi.org/10.1037/gpr0000051

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Greenwald, A. G. (1975). Consequences of Prejudice Against the Null Hypothesis. *Psychological Bulletin*, *82*(1).

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2019). Informed Bayesian t-Tests. *The American Statistician*.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., … Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Gu, X., Hoijtink, H., & Mulder, J. (2016). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology*, *72*, 130–143. https://doi.org/10.1016/j.jmp.2015.09.001

Haig, B. D. (2005). An Abductive Theory of Scientific Method. *Psychological Methods*, *10*(4), 371–388. https://doi.org/10.1037/1082-989x.10.4.371

Han, C., & Carlin, B. P. (2001). Markov Chain Monte Carlo Methods for Computing Bayes Factors. *Journal of the American Statistical Association*, *96*(455), 1122–1132. https://doi.org/10.1198/016214501753208780

Hand, D. J. (1994). Deconstructing Statistical Questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*(3), 317–356. https://doi.org/10.2307/2983526

Hardwicke, T. (2016). *A pre-registration primer*.

Hardwicke, T. E., & Ioannidis, J. P. A. (2018). Mapping the universe of registered reports. *Nature Human Behaviour*, *2*(11), 793–796. https://doi.org/10.1038/s41562-018-0444-y

Harms, C. (2015). *Statistische Überprüfung der Qualität empirischer Daten mit Hilfe von Simulationsverfahren* (Master-{{Thesis}}). University of Bonn.

Harms, C. (2018). A Bayes Factor for Replications of ANOVA Results. *The American Statistician*. https://doi.org/10.1080/00031305.2018.1518787

Harms, C., Genau, H. A., Meschede, C., & Beauducel, A. (2018). Does it actually feel right? A replication attempt of the rounded price effect. *Royal Society Open Science*, *5*(4), 171127. https://doi.org/10.1098/rsos.171127

Harms, C., Jackel, L., & Montag, C. (2017). Reliability and completion speed in online questionnaires under consideration of personality. *Personality and Individual Differences*, *111*, 281–290. https://doi.org/10.1016/j.paid.2017.02.015

Harms, C., & Lakens, D. (2018). Making Null Effects Informative: Statistical Techniques and Inferential Frameworks. *Journal of Clinical and Translational Research*, *3*(S2), 7. https://doi.org/10.18053/jctres.03.2017S2.007

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd Ed.). New York: Springer.

Haven, E., & Khrennikov, A. (2016). Statistical and subjective interpretations of probability in quantum-like models of cognition and decision making. *Journal of Mathematical Psychology*. https://doi.org/10.1016/j.jmp.2016.02.005

Haverkamp, N., & Beauducel, A. (2017). Violation of the Sphericity Assumption and Its Effect on Type-I Error Rates in Repeated Measures ANOVA and Multi-Level Linear Models (MLM). *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01841

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/s0140525x0999152x

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, *14*(4), 382–417. https://doi.org/10.1214/ss/1009212519

Hoffman, B. M., Papas, R. K., Chatkoff, D. K., & Kerns, R. D. (2007). Meta-analysis of psychological interventions for chronic low back pain. *Health Psychology*, *26*(1), 1–9. https://doi.org/10.1037/0278-6133.26.1.1

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science*, *359*(6377), 725–726. https://doi.org/10.1126/science.359.6377.725

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

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Isager, P. M. (2018). *What to Replicate? Justifications of study choice from 85 replication studies.* (p. 7). Technical University Eindhoven.

Isager, P. M. (2019, March). *Quantifying Replication Value: A formula-based approach to study selection in replication research.* 17. https://doi.org/10.23668/psycharchives.2392

Iso-Ahola, S. E. (2017). Reproducibility in Psychological Science: When Do Psychological Phenomena Exist? *Frontiers in Psychology*, *8*(June), 1–16. https://doi.org/10.3389/fpsyg.2017.00879

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science* (L. G. Bretthorst, Ed.). Cambridge: Cambridge University Press.

Jeffreys, H. (1961). *Theory of Probability* (3. ed.). Oxford: Clarendon Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree.

American Psychologist, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kass, R. E. (2011). Statistical Inference: The Big Picture. *Statistical Science*, *26*(1), 1–9. https://doi.org/10.1214/10-sts337

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203/_4

Klein, Richard A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahn'ik, Štěp'an, Bernstein, M. J., ... Nosek, B. A. (2014). Investigating Variation in Replicability. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, Richard A., Vianello, M., Hasselman, F., Adams, B. G., & Adams, R. B. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.17605/osf.io/s5vdy

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-016-1221-4

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLoS ONE*, *9*(9), e105825. https://doi.org/10.1371/journal.pone.0105825

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave, *Criticism and the growth of knowledge* (pp. 91–195). Cambridge University Press.

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

Lakens, D. (2015a). Comment: What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *Quarterly Journal of Experimental Psychology*, *68*(4), 829–832. https://doi.org/10.1080/17470218.2014.982664

Lakens, D. (2015b). On the challenges of drawing conclusions from p-values just below 0.05. *PeerJ*, *3*, e1142. https://doi.org/10.7717/peerj.1142

Lakens, D. (2017). Equivalence Tests. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., ... Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-018-0311-x

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

Leamer, E. E. (1983). Let's Take the Con Out of Econometrics. *The American Economic Review*, *73*(1), 31–43.

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A Unified Framework to Quantify the Credibility of Scientific Findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2018). *A Guide to Evaluate Replications: A Comment on Zwaan et al. (2017)*. https://doi.org/10.17605/osf.io/paxyn

Lecoutre, B. (1999). Two useful distributions for Bayesian predictive procedures under normal

models. *Journal of Statistical Planning and Inference*, *79*(1), 93–105. https://doi.org/10.1016/s0378-3758(98)00231-6

Lee, A. Y., Keller, P. A., & Sternthal, B. (2010). Value from Regulatory Construal Fit: The Persuasive Impact of Fit between Consumer Goals and Message Concreteness. *Journal of Consumer Research*, *36*(5), 735–747. https://doi.org/10.1086/605591

Liang, F., Paulo, R., Molina, G., Clyde, M., & Berger, J. O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, *103*(481), 410–423. https://doi.org/10.1198/016214507000001337

Liao, J. G., Midya, V., & Berg, A. (2019). *Connecting Bayes factor and the Region of Practical Equivalence (ROPE) Procedure for testing interval null hypothesis*. Retrieved from https://arxiv.org/abs/1903.03153

Liao, J. G., Midya, V., & Berg, A. (2020). Connecting and Contrasting the Bayes Factor and a Modified ROPE Procedure for Testing Interval Null Hypotheses. *The American Statistician*, 1–19. https://doi.org/10.1080/00031305.2019.1701550

Lindley, D. V. (1957). A Statistical Paradox. *Biometrika*, *44*(1-2), 187–192. https://doi.org/10.1093/biomet/44.1-2.187

Lipton, P. (2014). *Inference to the Best Explanation* (2nd ed.). London: Routledge.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2017). *Replication Bayes Factors from Evidence Updating* (pp. 1–19). https://doi.org/10.17605/osf.io/u8m2s

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3, Pt.1), 151–159. https://doi.org/10.1037/h0026141

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, *55*(2), 95–107. https://doi.org/10.1037/h0056029

Magnusdottir, B. B., Faiola, E., Harms, C., Sigurdsson, E., Ettinger, U., & Haraldsson, H. M. (2019). Cognitive Measures and Performance on the Antisaccade Eye Movement Task. *Journal of Cognition*, *2*(1). https://doi.org/10.5334/joc.52

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*(2), 161–175. https://doi.org/10.1007/bf01173636

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Marsman, M., Schönbrodt, F. D., Morey, R. D., 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). https://doi.org/10.1098/rsos.160426

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279. https://doi.org/10.1080/17470218.2012.711335

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989x.9.2.147

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

Mayo, D. G., & Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. In *Optimality* (Vol. 49, pp. 77–97). https://doi.org/10.1214/074921706000000400

Mayo, D. G., & Morey, R. D. (2017). *A poor prognosis for the diagnostic screening critique of statistical tests*. https://doi.org/10.17605/osf.io/ps38b

Mayo, D. G., & Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction. *The British Journal for the Philosophy of Science*, *57*(2), 323–357. https://doi.org/10.1093/bjps/axl003

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (1st Edition). Chapman and Hall/CRC.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*(2), 103–115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006x.46.4.806

Meehl, P. E. (1990a). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102/_1

Meehl, P. E. (1990b). Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*, *66*(1), 195. https://doi.org/10.2466/pr0.66.1.195-244

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do Frequency Representations Eliminate Conjunction Effects? An Exercise in Adversarial Collaboration. *Psychological Science*, *12*(4), 269–275. https://doi.org/10.1111/1467-9280.00350

Meng, X. L., & Wing, H. W. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*, 831–860.

Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference*, *26*(2), 231–245. https://doi.org/10.1016/j.foodqual.2012.05.003

Montag, C. (2018). Cross-cultural research projects as an effective solution for the replication crisis in psychology and psychiatry. *Asian Journal of Psychiatry*, *38*, 31–32. https://doi.org/10.1016/j.ajp.2018.10.003

Morey, R. D. (2017). The p-curve is not what you think it is. Retrieved from http://richarddmorey.org/content/Psynom17/pcurve//#/

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., … Zwaan, R. A. (2016). The Peer Reviewers' Openness Initiative: Incentivising Open Research Practices through Peer Review. *The Royal Society Open Science*, *3*. https://doi.org/10.1098/rsos.150547

Morey, R. D., Wagenmakers, E.-J., & Rouder, J. N. (2016). Calibrated Bayes Factors Should Not Be Used: A Reply to Hoijtink, van Kooten, and Hulsker. *Multivariate Behavioral Research*, *51*(1), 11–19. https://doi.org/10.1080/00273171.2015.1052710

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515. https://doi.org/10.1177/2515245918797607

Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology*, *72*, 104–115. https://doi.org/10.1016/j.jmp.2014.09.004

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, *89*(6), 852–863. https://doi.org/10.1037/0022-3514.89.6.852

Munaf'o, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Neuliep, J. W., & Crandall, R. (1990). Editorial Bias Against Replication Research. *Journal of Social Behavior and Personality*, *5*(4), 85–90.

Neuliep, J. W., & Crandall, R. (1993). Reviewer Bias Against Replication Research. *Journal of Social Behavior and Personality*, *8*(6), 21–29.

Neyman, J., & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *231*(694-706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. https://doi.org/10.1037/1082-989x.5.2.241

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). Registered Reports: A Method to Increase the Credibility of Published Results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Nuijten, M. B. (2012). *A Default Bayesian Hypothesis Test for Mediation* (PhD thesis). University of Amsterdam.

Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., & Wagenmakers, E.-J. (2015). A default Bayesian hypothesis test for mediation. *Behavior Research Methods*, *47*(1), 85–97. https://doi.org/10.3758/s13428-014-0470-2

O'Donnell, M., & Nelson, L. D. (2015). *Wadhwa & Zhang, 2015, Study 3 Replication*.

O'Keefe, D. J. (2007). Brief Report: Post Hoc Power, Observed Power, A Priori Power, Retrospective Power, Prospective Power, Achieved Power: Sorting Out Appropriate Uses of Statistical Power Analyses. *Communication Methods and Measures*, *1*(4), 291–299. https://doi.org/10.1080/19312450701641375

Obels, P., Lakens, D., Coles, N. A., & Gottfried, J. (2019). *Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology*.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

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

Orben, A., & Lakens, D. (2020). Crud (Re)Defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238–247. https://doi.org/10.1177/2515245920917961

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science*, *7*(6), 531–536. https://doi.org/10.1177/1745691612463401

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science. *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*, *6*(Mar), 1–11. https://doi.org/10.3389/fpsyg.2015.00223

Popper, K. R. (1935). *Logik der Forschung*. https://doi.org/10.1007/978-3-7091-4177-9

Popper, K. R. (2002). *The Logic of Scientific Discovery* (2nd ed.). London: Routledge.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891. https://doi.org/10.3758/brm.40.3.879

R Core Team. (2019). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, *92*(437), 179–191. https://doi.org/10.1080/01621459.1997.10473615

Renkewitz, F., & Heene, M. (2019). The Replication Crisis and Open Science in Psychology: Methodological Challenges and Developments. *Zeitschrift für Psychologie*, *227*(4), 233–236. https://doi.org/10.1027/2151-2604/a000389

Retraction Watch. (2020). Retracted coronavirus (COVID-19) papers [Blog].

R-Index.org. (2014). *R-Index 1.0*.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. https://doi.org/10.1037/0033-2909.113.3.553

Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research*, *47*(6), 877–903. https://doi.org/10.1080/00273171.2012.734737

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/pbr.16.2.225

Savalei, V., & Dunn, E. (2015). Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology*, *6*(MAR), 10–13. https://doi.org/10.3389/fpsyg.2015.00245

Schad, D. J., Betancourt, M., & Vasishth, S. (2019). *Toward a principled Bayesian workflow in cognitive science*. Retrieved from https://arxiv.org/abs/1904.12765

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00813

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551–566. https://doi.org/10.1037/a0029487

Schimmack, U. (2014). *Quantifying Statistical Research Integrity: The Replicability-Index* (pp. 1–31). University of Toronto Mississauga.

Schimmack, U., & Brunner, J. (2017). *Z-Curve: A Method for the Estimating Replicability Based on Test Statistics in Original Studies*. University of Toronto Mississauga.

Schlenker, B. R. (1974). Social psychology and science. *Journal of Personality and Social Psychology*, *29*(1), 1–15. https://doi.org/10.1037/h0035668

Schmidt, S. (2009). Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Review of General Psychology*, *13*(2), 90–100. https://doi.org/10.1037/a0015108

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/met0000061

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316. https://doi.org/10.1037/0033-2909.105.2.309

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p Values for Testing Precise Null Hypotheses. *The American Statistician*, *55*(1), 62–71. https://doi.org/10.1198/000313001300339950

Shaw, J. (2017). Was Feyerabend an anarchist? The structure(s) of 'anything goes.' *Studies in History and Philosophy of Science Part A*, *64*, 11–21. https://doi.org/10.1016/j.shpsa.2017.06.002

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., & Simonsohn, U. (2017). Power Posing: P-Curving the Evidence. *Psychological Science*, 95679761665856. https://doi.org/10.1177/0956797616658563

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An Introduction to Registered Replication Reports at Perspectives on Psychological Science. *Perspectives on Psychological Science*, *9*(5), 552–555. https://doi.org/10.1177/1745691614543974

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Simonsohn, U. (2013). Just Post It: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone. *Psychological Science*, *24*(10), 1875–1888. https://doi.org/10.1177/0956797613480366

Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Sinharay, S., & Stern, H. S. (2002). On the Sensitivity of Bayes Factors to the Prior Distributions. *The American Statistician*, *56*(3), 196–201. https://doi.org/10.1198/000313002137

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Smith, N. C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist*, *25*(10), 970–975. https://doi.org/10.1037/h0029774

Sprenger, J. (2015). *The Objectivity of Subjective Bayesian Inference* (pp. 1–22).

Srinivas, M. D. (1975). Foundations of a quantum probability theory. *Journal of Mathematical Physics*, *16*(8), 1672–1685. https://doi.org/10.1063/1.522736

Steiger, J. H. (2004). Beyond the F Test: Effect Size Confidence Intervals and Tests of Close Fit in the Analysis of Variance and Contrast Analysis. *Psychological Methods*, *9*(2), 164–182. https://doi.org/10.1037/1082-989x.9.2.164

Sterling, Theodore D. (1959). Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.1080/01621459.1959.10501497

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *The American Statistician*, *49*(1), 108–112. https://doi.org/10.1080/00031305.1995.10476125

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific Misconduct and the Myth of Self-Correction in Science. *Perspectives on Psychological Science*, *7*(6), 670–688. https://doi.org/10.1177/1745691612460687

Stroebe, Wolfgang, & Strack, F. (2014). The Alleged Crisis and the Illusion of Exact Replication. *Perspectives on Psychological Science*, *9*(1), 59–71. https://doi.org/10.1177/1745691613514450

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics - Theory and Methods*, *25*(7), 1595–1610. https://doi.org/10.1080/03610929608831787

Tsang, E. W. K., & Kwan, K.-M. (1999). Replication and Theory Development in Organizational Science: A Critical Realist Perspective. *Academy of Management Review*, *24*(4), 759–780. https://doi.org/10.5465/amr.1999.2553252

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*(2), 83–91. https://doi.org/10.1037/h0027108

Tybur, J. M., & Navarette, C. D. (2017). Interrupting Bias in Psychological Science: Evolutionary Psychology as a Guide. In J. T. Crawford & L. Jussim (Eds.), *Politics of Social Psychology*. Psychology Press.

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting Meta-Analyses Based on p Values: Reservations and Recommendations for Applying p-Uniform and p-Curve. *Perspectives on Psychological Science*, *11*(5), 713–729. https://doi.org/10.1177/1745691616650874

Van Horn, K. S. (2003). Constructing a logic of plausible inference: A guide to Cox's theorem. *International Journal of Approximate Reasoning*, *34*(1), 3–24. https://doi.org/10.1016/s0888-613x(03)00051-3

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498. https://doi.org/10.1016/j.jmp.2010.07.003

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. https://doi.org/10.1037/a0036731

Wadhwa, M., & Zhang, K. (2015). This Number Just Feels Right: The Impact of Roundedness of Price Numbers on Product Evaluations. *Journal of Consumer Research*, *41*(5), 1172–1185. https://doi.org/10.1086/678484

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/bf03194105

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. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian Benefits for the Pragmatic Researcher. *Current Directions in Psychological Science*, *25*(3), 169–176. https://doi.org/10.1177/0963721416643289

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., … Morey, R. D. (2015). A power fallacy. *Behavior Research Methods*, *47*(4), 913–917. https://doi.org/10.3758/s13428-014-0517-4

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. https://doi.org/10.1037/a0022790

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wandel, S., Juni, P., Tendal, B., Nuesch, E., Villiger, P. M., Welton, N. J., … Trelle, S. (2010). Effects of glucosamine, chondroitin, or placebo in patients with osteoarthritis of hip or knee: Network meta-analysis. *Bmj*, *341*(sep16 2), c4675–c4675. https://doi.org/10.1136/bmj.c4675

Wang, D., Zhang, W., & Bakhai, A. (2004). Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. *Statistics in Medicine*, *23*(22), 3451–3467. https://doi.org/10.1002/sim.1930

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p -Values: Context, Process,

and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064. https://doi.org/10.3758/s13423-012-0295-x

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Yarkoni, T. (2019). *The Generalizability Crisis*.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making Replication Mainstream. *Behavioral and Brain Sciences*, 1–50. https://doi.org/10.1017/s0140525x17001972

Appendix A

Reproduction of Original Publications

The reproduction of original articles is omitted from the published version of this thesis. All articles are available through the journal websites with articles I and III being available through Open Access.

A.1 Article I: Harms & Lakens (2018)

Original publication *Article I*:

Harms, C., & Lakens, D. (2018). Making Null Effects Informative: Statistical Techniques and Inferential Frameworks. Journal of Clinical and Translational Research, 3(S2), 7. https://doi.org/10.18053/jctres.03.2017S2.007

A.2 Article II: Harms (2018)

Original publication *Article II*:

Harms, C. (2018). A Bayes Factor for Replications of ANOVA Results. The American Statistician. https://doi.org/10.1080/00031305.2018.1518787

A.3 Article III: Harms et al. (2018)

Original publication *Article III*:

Harms, C., Genau, H. A., Meschede, C., & Beauducel, A. (2018). Does it actually feel right? A replication attempt of the rounded price effect. Royal Society Open Science, 5(4), 171127. https://doi.org/10.1098/rsos.171127