PhD DISSERTATIONS OF
NORTHWESTERN ITALIAN PHILOSOPHY CONSORTIUM – FINO

Methods and Philosophy: Doing Science

Noah N'Djaye Nikolai van Dongen

(2021)

Examination Committee/Commissione di esame:

Supervisor: Prof. J. M. Sprenger
Co-supervisor: Dr. L. van Grootel
Examiner: Prof. E. J. Wagenmakers
Prof. J. W Romeijn

UNIVERSITÀ DEGLI STUDI DI GENOVA
UNIVERSITÀ DEGLI STUDI DI TORINO
UPO UNIVERSITÀ DEL PIEMONTE ORIENTALE
UNIVERSITÀ DEGLI STUDI DI PAVIA

CONTENTS

PREFACE

Kintsugi is the Japanese art of repairing broken pottery. The areas of breakage are mended with lacquer, which is dusted or mixed with powdered silver, gold, or platinum. The philosophy is that these breakages and repairs are elements of the object, its history, and beauty, which should not be disguised. Flaws and imperfections are embraced; they are part and parcel of the object's aesthetic. Delicate cups and bowls, glazed with an elegant but modest patina of deep browns or blues, beget an asymmetric pattern of something precious. This pattern that emerges appears chaotic, unstructured, and unpredictable, though is still the product of a deterministic process.

I find Kintsugi a pleasing metaphor for science. That is why I chose such a mended cup as a cover image for this dissertation. Please forgive me for taking some artistic liberties and making some simplifications and generalizations to tell this story. Science is constantly under pressure from all sides. Science and how we perceive the world through it is battered and bruised in various places when exposed to new discoveries, more precise measurement from more advanced instruments, or alternative theories that provide better fitting explanations, but also through cultural changes, political demands, financial constraints, and ethical considerations. To continue the metaphor and without technical jargon, an actual break can occur when slowly building stress (e.g., increasing concerns about certain methods) becomes too much to bear or it comes into contact with a solid object at speed (e.g., a new theory). We, scientists and philosophers, continually break science (or parts of it), not entirely by accident, and put it back together again. These breaks and their repairs —adoption of new theories and changes in methods, culture, and perspectives— remain visible to the interested perceiver.

This is where the metaphor ends. In Kintsugi, one refits the chards into the cup

it once was. Its aesthetic appeal might have been enhanced and a spider web of metal line now dissects the cup's surface, it still has the same shape and utility that it had before. In science, none of us can predict what it and our worldview will look like after such a break-and-repair act. Please forgive me for being cliché and give the too much used examples of upheavals in science. Our perspective on biology, animals, and the position of human irrevocably changed through Darwin's theory of natural selection. Our perspective on space and time changed with Einstein. Our perspective of chance and determinism changed with advent of quantum mechanics. After each repair, the world was a more interesting and explicable place. At least, that is what I think.

Now, psychological science also has an example that could become typical or even cliché. In the second decade of the twenty-first century the pressure that had been building for years around the use of certain research methods was pushed beyond the breaking point. Large-scale replication attempts caused the shattering of pedestals of eminent scientists and the disintegration of edificial research. A crisis ensued from which we are still picking up the pieces. Healing and reconstruction is still ongoing and it is yet unclear what psychological science will look like when all the chards are fitted and lacquered together. I expect that it will be glorious.

I started my PhD in 2016, when the replication crisis was in full swing. I had a background in fine arts and arts and culture studies. I knew some philosophy of science, but I could not be considered educated. I had some experience with empirical research and statistics, both Bayesian and frequentist. In comparison to most philosophers, I had a lot of experience with empirical research and statistics. Like most (junior) academics, I suffered from (a mild case of) imposter syndrome, which I addressed by avid reading and education. I also tried to sidestep it by focusing more on what I did know (something about) – methods and statistics – than on what I did not – philosophy of science. This was not too bad, because I like to apply and test things. The end result, I want(ed) my work to be connected to what researchers do and (might) need. The end result is a dissertation that mixes philosophy of science with doing science.

Some might say that it lacks philosophical depth and nuance. To which I would counter that it is at least connected to what scientists currently do and made in close collaboration with them. In my experience, most of present day philosophy of science is practiced in isolation of science and living scientists. There are some exceptions of course, but in general journals and conferences are kept separate and streams of communications are better described as the dripping of a leaky faucet. To

illustrate the situation, in 2019 I was at a philosophy conference on biases in science. A topic of which one would expect scientists themselves to know most about and who could also best inform what kind of solutions would be feasible. However, not a single scientist was present. The fault does not only lie with philosophy. It seems that both scientists and philosophers want to be in contact with each other, but nobody really knows how to do it. Like high school teenagers at their first dance, the philosophers stand on one side of the decorated gymnasium hall and the scientists on the other. They all want to dance with each other, almost nobody knows how to do it, and only a brave few manage it. So, I don't mind giving up a bit of philosophical nuance if I can accomplish having my work more closely connected to that which I am to philosophize about.

To conclude the beginning of this dissertation, I want to express here how uncertain I am. I often feel overwhelmed by the vastness of nature and the intricacies of the tools we have developed to understand her. I am confused by what I know and humbled by what I do not. There is so much to know; yet I know little and understand less. Nonetheless, I am curious and enjoy doing science and figuring stuff out. This dissertation is the product of a four-year journey coping with doubt and uncertainty, of puzzling and learning, and of great joy in the things I did figure out.

1

INTRODUCTION

1.1. Opening Words

This dissertation was written during an interesting time to be either a philosopher of science or a scientist. Many famous results of empirical research various fields of science do not hold up in replication attempts under rigorous conditions (e.g., medicine, economics, psychology; Begley & Ellis, 2012; Camerer et al., 2016; Open Science Collaboration, 2015; Prinz, Schlange, & Asadullah, 2011) and the largest science experiment, the Large Hadron Collider, has produced negative results with the exception of a single discovery that was predicted decades ago (discovery of the Higgs boson; CERN, 2013). There are many calls of crises of various sorts —reproducibility crisis, theory crisis, generalizability crisis, to name a few— any more than a few scientists turn towards philosophy for answers and aid (e.g., Lakens, 2019a).

Simultaneously, there has been a growing trend of philosophers doing empirical experiments, instead of thought experiments. The results of these studies have added fuel to the fire of some philosophical debates (e.g., the stability of philosophical intuitions; Joshua, 2019; Machery & Stich, 2019) and the value, quality, and credibility of these experiments are themselves the topics of fierce discussions. As a result, philosophy and science are moving towards each other and interaction is increasing. It could be the enduring reunion of long lost friends or a brief encounter of old acquaintances. However this will develop, at the moment much is happening at this intersection.

We are at a point where we should take stock and evaluate not just what we do, but how we do it: evaluate the practice and tools of our trade. We should assess the quality and usefulness of the tool we use and the skill required to adequately use them. Broadly speaking, this dissertation is about doing science and doing it well, or at least slightly better than completely terrible. It is an attempt at a modest contribution to both the practice of science from the perspective of philosophy and the practice of (experimental) philosophy from the perspective of research methodology. Below, I provide an outline of the current situation in which the individual chapters are situated.

1.2. Philosophers Doing Science

The beginning of the twenty first century saw the rise of a new academic field, experimental philosophy (Sytsma & Buckwalter, 2016). This field aims to supplement analytic philosophy's armchair approach with empirical research. Practitioners in this field use data-driven experimental methods employed by social science to tackle

questions studied by philosophers. At first, experimental philosophy was focused on the empirical investigation of philosophically relevant intuitions (e.g., semantic intuitions of reference; Machery, Mallon, Nichols, & Stich, 2004).

Philosophers argue for (against) theories on concepts such as knowledge or causation, in part by reflect on scenarios, hypothetical and actual, to question whether they instantiate the target concept or category. The mental responses to such a scenario are called 'intuitions', because the subject who is considering the scenario generally has the immediate inclination to give a particular response without being consciously aware of preceding reasoning. The philosopher's intuition concerning such scenarios are assumed to be universal and thus acceptable as evidence for (against) the theory if the response aligns with (contradicts) the scenario that is purported to instantiate the theory (Stich & Tobia, 2016).[1]

However, the validity of such metaphysical investigations, of for instance the nature of knowledge, was put into question by studies showing that intuitions can be, for instance, culturally diverse (e.g., Machery et al., 2004), dependent on personality (e.g., Feltz & Cokely, 2009), or whether or not one is a philosopher (e.g., Vaesen, Peterson, & Van Bezooijen, 2013). In other words, experimental philosophy investigated the modulating factors of philosophical intuitions and their underlying psychological mechanisms and showed the lack of its universality (Knobe et al., 2012; Knobe & Nichols, 2013; Machery, 2017).

Experimental philosophy did not stay limited to the testing of intuitions. In time, experimental philosophy broadened its scope and is currently dedicated to empirically test any key premises of philosophical arguments, which include people's intuitions, but also attitudes, emotional responses, behaviors, and perceptions (Knobe, 2007, 2016). Currently, several hundreds of papers have been published that contain at least one empirical study (242 papers published between 2003 and 2016; Cova et al., 2018) and many more on discussing an interpreting their results.

Experimental philosophy is not only a flourishing field, it also appears to be in good health. In a large-scale replication project, 70% of the results of a representative sample of 40 studies were replicated under controlled conditions (Cova et al., 2018). This is considerably better than psychological science, where only 36% of a similar replication attempt of 100 studies obtained statistically significant results (Open Science Collaboration, 2015), and medical science, where for instance results of only 11% of 56 landmark cancer studies could be replicated (Begley & Ellis, 2012). In addition, relatively few errors (intentional or unintentional) are made in reporting of

[1] Note that it is not categorically accepted that this is actually what philosophers do (e.g., Cappelen, 2012; Deutsch, 2015).

statistics (Colombo, Duev, Nuijten, & Sprenger, 2018) and the literature generally lacks publication bias and shows evidential value (Stuart, Colaço, & Machery, 2019).[2]

A noteworthy episode in the history of experimental philosophy is the study by Machery, Mallon, Nichols, and Stich (2004). The aim of this experiment was to test people's intuitions about the meaning of proper names. Concretely, whether Western people and East Asian people differed in their intuitions about names referring to people through *association of descriptive properties* (descriptivist theory of reference; Frege, 1892; Russell, 1905; Searle, 1958) or its *causal history* derived from an original act of baptism (causal-historical theories of reference; Kripke, 1980). Broadly speaking, participants were asked to read four vignettes about a description associated with the wrong person or an imaginary person and answer a question for each one to which person the name refers. Based on their results, Machery et al. (2004) concluded that Western and East Asian people differ in their semantic intuitions concerning proper names. This conclusion reverberated through analytic philosophy, because, as noted above, intuitions were considered to be invariant and universal and thus acceptable as premises in philosophical arguments. This paper sparked debates on the methodology of analytic philosophy (e.g., Cappelen, 2012; Deutsch, 2015; Machery, 2017) and inspired many to replicate and elaborate on this study (e.g., Beebe & Undercoffer, 2016; Izumi, Kasaki, Zhou, & Oda, 2018; Kasaki, 2017; Lam, 2010; Sytsma & Livengood, 2011; Sytsma, Livengood, Sato, & Oguchi, 2015).

However, curious data hides behind the sweeping conclusion of the Machery et al. (2004) study. For only two of the four vignettes, the Westerners and East Asians differed in their answers to a statistically significant degree. Machery et al. (2004) argued that this was due to the other vignettes being of a different type. This might be the reason, though it does not explain why the Western sample was, as a group, close to ambivalent about reference; 17 for the first vignette and 18 for the second vignette choose the causal-historical answer, while 14 and 13 choose descriptivist answers. The East Asian sample leaned more towards descriptivist intuitions (about 70% of 41 participants), but far from unanimously, even if one would take into account a moderate amount of noise in the measurement. This curiosity can either be the result of a lack of shared semantic intuitions even within cultures or there are methodological issues of reliability or validity.[3] The former seems unlikely

[2]It needs to be noted that the method used to detect publication bias and evidential value in the literature has its flaws (e.g., miss publication bias under certain conditions, Bishop & Thompson, 2016; Carter, Schönbrodt, Gervais, & Hilgard, 2019)

[3]In quantitative empirical research, reliability refers to the amount of noise in the measurement and validity refers to how good it is at actually measuring what it is supposed to measure.

because of the relative lack of errors in communications when referring to people. The later seems most plausible and has been addressed in several studies that followed Machery et al. (2004). However, these studies focused on the language and phrasing of the vignettes, questions, and answer options, which were adapted based on philosophical 'arm chair' considerations with the intention to increase reliability (e.g., Sytsma & Livengood, 2011; Sytsma et al., 2015). Consequently, the researchers neglected empirical evaluation of the design (e.g., test inter- and intra-subject reliability) and neglected to investigate whether they were actually testing semantic intuitions with the vignettes and answer options.

The Machery et al. (2004) study is not an exception, but typical case of how experimental philosophy has thus far approached experimental research. Empirical evaluation of study design in terms of reliability, validity, and other methodological qualities is generally lacking. In summary, there is still more that can be learned from social science methodology when it comes to the supplementing of classical analytic philosophy with empirical research.

1.3. Scientists doing Philosophy

In 2011, two events shook the social sciences to its core. One was the report on data fabrication by Dutch professor and dean of the Social and Behavioral Sciences Faculty at Tilburg University, Diederik Stapel. His work (over 50 publications) had past all checks that research results go through before they get published, it had been lauded by his peers, and had been published in the most prestigious journals. He had not used any fancy simulation software nor did he had to use much subterfuge or sophisticated trickery to get his data past students, collaborators, and reviewers. He always took responsibility for the data collection, typed in the data manually on his laptop, provided the clean data to his collaborators and students, and just was evasive when asked about the data's provenance and his need to do the data collection. His fraud only came to light when PhD students and an assistant professor, suspicious about his behavior and authenticity of the data, started investigating the structure of the, as it turned out, fabricated data (Stapel, 2012). The second major event was 'Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect', an article published by distinguished professor Daryl Bem in the prestigious social psychology journal *Journal of Personalty and Social Psychology*. In this paper, Ben reported nine experiments, involving over 1,000 participants, eight of which showing evidence in favor of some *precognitive* ability (i.e., the future having some effect on the participants responses). The devastating element was

that the results are in stark contradiction with well established laws of physics (e.g., speed of light and second law thermodynamics), though these results supporting precognition followed from ubiquitously used and accepted methods of statistical analysis of quantitative data.[4]

These events triggered an reevaluation of scientific methodology and the veracity of the published literature. In psychology, a hundred landmark studies were selected for replication. For several years, researchers around the globe conducted experiments under rigorous conditions that resembled the original hundred studies a closely as possible. In 2015, the results were published, which painted a disenchanting picture. Of the original studies they had selected, 97% had reported statistically significant results (indication as evidence for the research hypothesis). In stark contrast, only 36% of the replication studies resulted in statistical significance (Open Science Collaboration, 2015). Not just psychology, but life science, neuroscience, and behaviroal economics seem to suffer a similar fate (e.g., Aguinis, Cascio, & Ramani, 2017; Begley & Ellis, 2012; Button et al., 2013; Camerer et al., 2016, 2018; Ioannidis, Munafo, Fusar-Poli, Nosek, & David, 2014; Poldrack et al., 2017; Prinz et al., 2011). Many other fields might share this fate, though this remains unclear until similar rigorous replication studies are undertaken. What we do know is that many scientists from fields such as biology, chemistry, physics, and earth sciences have reported that they have experienced failure in trying to reproduce their own results or that of their colleagues (Baker, 2016). Of course, failed replications by themselves can be considered unproblematic (Amrhein, Trafimow, & Greenland, 2019): researchers are fallible and no study comes with methodological guaranties of truth conduciveness, false-positives and false-negatives are unavoidable. However, it clearly is a problem when tentative results are published as established facts while failed replications of those studies usually are difficult, almost impossible to publish.

Simultaneous with the replication projects, researchers started investigating practices that could produce this inflated amount of false-positive results[5] that (individually) passed then current academic standards. In 2011, Simmons, Nelson, and Simonsohn (2011) showed that under certain conditions one could attain a twelve-fold increase in the probability (from 5% to 60.7%) for obtaining positive results for the obviously false hypothesis that listing to particular music reduces your age (i.e., raises your year of birth). These conditions were not esoteric, difficult,

[4]Many prior events brought problems to light, but it is generally accepted that these two opened the flood gates to the current movement of scientific reform.

[5]Conventional practice of social science is to fix the false positive rate of statistical inference at 5%. This mean that, in the long run and if assumption of the method are not violated, at most 5% of results support an hypothesis that is actually false.

or costly to enact. The high false-positive result probability was obtained by A) switching to another outcome variable (from age of participant to age of father); B) sampling 10 more observations when statistical significance was not reached; C) adding a control variable (gender of the participant); and D) dropping and experimental condition. These and other *questionable research practices* are (or at least were) ubiquitous among social scientists (John, Loewenstein, & Prelec, 2012) and were promoted in standard textbooks by eminent scientists (e.g., Mutz, 2011; Sansone, Morf, & Panter, 2003).

In recent years, the focus on replicability has intensified (according to Google Scholar: 815 publication in psychology on "replication crisis" in 2018 with respect to 10 in 2011) and many proposals have been made and implemented to combat the low replication rate. Prominent among these proposal are the call for increased replication effort (e.g., Zwaan, Etz, Lucas, & Donnellan, 2018), transparency through data and method sharing (e.g., Aczel et al., 2020; Farnham et al., 2017; Frankenhuis & Nettle, 2018; Munafò, 2016), and the specification of study design and data analysis prior to carrying out the research (.e., preregistration; Nosek, Ebersole, DeHaven, & Mellor, 2018). The implementation of these proposal are expected to reduce the canonization of 'false facts' (Nissen, Magidson, Gross, & Bergstrom, 2016), because they should fail to replicate and error or questionable research practices can be read from the open data and/or preregistered analysis plan. Furthermore, many have called into question and are attempting to change the incentive structure (e.g., M. A. Edwards & Roy, 2017; Lindner, Torralba, & Khan, 2018) of rewarding (e.g., keynote, hiring, tenure, etc.) based on the number of publications in 'eminent' journals, which favor positive, novel, and eye-catching results over methodological rigor (Ioannidis et al., 2014; Munafo, Clark, & Flint, 2004; Nissen et al., 2016).

According to many (e.g., Aczel, Palfi, & Szaszi, 2017; Johnson, 2019), another main culprit of the replication crisis, is the statistical procedure of the *null hypothesis significance test*[6] (NHST) in which a null hypothesis is rejected in favor of the hypothesis of interest if a particular statistic has crossed a conventional threshold (i.e., '$p < 0.05$'). This has led to proposals such as lowering the conventional threshold (Benjamin et al., 2018), requiring justification for the use of particular threshold levels (Lakens, Adolfi, et al., 2018), or refrain from using such thresholds all together (McShane, Gal, Gelman, Robert, & Tackett, 2019). The problems with NHST have been know for decades (e.g., Berkson, 1942) and discussions about it have waxed and waned in the intervening years (e.g., Bakan, 1966; J. Cohen, 1994). The repli-

[6]See Chapter 3 for in-depth explanation of the method.

cation crisis has added new fuel to this fire (e.g., the special issue of *The American Statistician* "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$"; volume 73 (supp1), 2019). Some have proposed abandoning hypothesis tests in favor of estimation, while staying within the frequentist school of statistics (e.g., Coulson, Healey, Fidler, & Cumming, 2010; Cumming, 2013a, 2013b, 2014; Cumming, Fidler, Kalinowski, & Lai, 2012). Other have strongly opposed this solution through arguing in favor of the necessity of hypothesis tests in science (Morey, Rouder, Verhagen, & Wagenmakers, 2014). They and others have been promoting Bayesian statistics in general and the Bayesian hypothesis tests in particular as the alternative to NHST and frequentist statistics, which should contribute to remedying the replication crisis (e.g. Wagenmakers et al., 2017).[7]

In reaction to this literature on the lack of replicability, other scientists have been arguing that something else than the lack of replicability is the actual problem that needs to be addressed. To convey a similar level or urgency these problems are also labelled 'crisis'. The lack of replicability is a by-product of deeper rooted problems that should be considered the actual crisis. *Measurement crisis:* a general lack of strong and verified connection between what the intended target of a measurement is and how it is measured (i.e., validity), high variability, and /or sensitivity to irrelevant circumstances ensures that results are spurious and will most likely not replicate when measurements are repeated (Flake & Fried, 2019). *Generalizability crisis*: the general tendency in social science to draw broad theoretic conclusions from statistical results from single studies creates an inferential gap that cannot be addressed by replicating these single studies. The lack of taking relevant factors into account across which results might vary creates room for spurious results and low replication rates (Yarkoni, 2019). *Theory crisis*: the general absence of strong theories in social science that clearly demarcate what should and should not be observed has allowed fantastic results to be published (e.g., Bem, 2011). In light of such a strong theory, it is obvious that these results do not replicate on subsequent attempts, but also that they would not have been published unapologetically in the first place (Oberauer & Lewandowsky, 2019).

In this time of crisis, social scientists have turned to philosophy of science for guidance. Almost a century ago, physicists turned to philosophy when the establishment of quantum mechanics shattered their world view of classical mechanics

[7]It should be noted that this is but the latest chapter of an ongoing conflict about the merits and defects of frequentist and Bayesian statistics that has been going on since the inception of these schools of statistics and which shows no sign of abating (For the interested reader, see: Mayo, 2018; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008; Wagenmakers et al., 2018). I have saved the few comments I have about this conflict for the conclusion of this dissertation.

(e.g., Howard, 1997). This time the social scientists are turning to philosophy of science. Problems in results and practice have come to light that cannot be easily dismissed or ignored. Potential causes of these problems are identified, proposed, and debated on grounds of rhetoric and arguments. To make sense of this novel state of affairs that is void of empirical information, scientists wanting to understand and improve the situation, have sought aid in classical text of philosophy of science. For instance, scientists have reasoned on how one's perspective on the existence of reality informs scientific practice (Lakens, 2019a); perspectives such as 'instrumentalism', 'constructive empiricism', 'entity realism', and 'scientific realism' (Ladyman, 2002).[8] Also, scientists have turned to the abductive method of inference to the best explanation (Borsboom, van der Maas, Dalege, Kievit, & Haig, 2020; Haig, 2014). In another instance, the conceptual analysis method was adopted to investigate (the value of) preregistration (Lakens, 2019b). In this case, the author reaches back to Popper (2002[1959]), Lakatos (1980), and van Fraassen (1980) to argue for the value of preregistration in terms of the severity of the test and the trustworthiness of the results. Furthermore, scientists have argued for increased replication attempts to foster Lakatosian (1970) progressive research programs (Zwaan et al., 2018). Finally, old debates between deduction (see Lakens, 2020) and induction (see Yarkoni, 2020) are resurfacing as well.

These events indicate that the social sciences can learn much from present philosophy of science in their attempt to understand and improve their field. They research back to classical literature and many are still stuck on a naive view on Popper's falisificationism (Lemoine, 2019).[9] For most researchers, contact with philosophy of science ended with a single course as an undergraduate and have never seen or even heard of a living philosopher of science. In brief, improving researchers' familiarity with current philosophy of science will most likely aid them in traversing this period of upheaval.

1.4. Doing Science and Doing it Well

If the last decades have taught us anything, it is that the methods we employ in science and philosophy are flawed, that we can (eventually) identify these flaws, and that we can improve our methods by addressing them. Feyerabend might have been right when he proclaimed that "anything goes" was the only methodological rule for distinguishing good from bad scientific practice that would not wrongly

[8] Ladyman (2002) was the source used as reference for these perspectives.
[9] Although, physics is suffering from a similar problem (Hossenfelder, 2017).

exclude some elements. In advance, there is no certainty about methodological quality. Only when the proverbial rubber hits the road are weaknesses, errors, and deficiencies detected. Falsificationism is employed to improve our methods and science progresses in tandem with its development of methodology (Wootton, 2015). Like Neurath's ship (1973), we are repairing and improving science's methodological vessel, manned by imperfect and bias-prone humans, as it traverses the perilous waters of the unknown. In less poetic terms, we need to be vigilant for mistakes and foster an environment that cultivates such vigilance.

Experimental philosophy uncovered psychological biases in philosophical reasoning. For instance, several studies provided insight into how seemingly irrelevant elements (e.g., the order of stimuli) influenced philosophers' conclusions on various thought experiments (e.g., Schwitzgebel & Cushman, 2012). These results open the doorway to the possibility that principles, intuitions, assumptions, and/or axioms on which analytical analysis in philosophy are based might be influenced, unwanted and unexpected, by the predispositions of the philosopher. Our tendency to conformity (e.g., Ioannidis & Nicholson, 2012) might be the reason this has escaped our attention until it was flagged by an external arbiter (e.g., empirical tests). Just like regular human beings, scientists and philosophers suffer from psychological biases (e.g., Alger, 2019; Pohl, 2004).[10] In addition, we are prone to chase noise due to an overly effective pattern recognition system (e.g., seeing faces on toast and rock formations on mars) and we are by nature bad statistician (e.g., Tversky & Kahneman, 1971). This is exacerbated by our leaning towards positive results (i.e., publication bias; Marks-anglin & Chen, 2020), reporting them as more favorable than they actually are (i.e., spin bias; Boutron, Dutton, Ravaud, & Altman, 2010), and citing only favorable papers with positive results (i.e., citation bias Duyx, Urlings, Swaen, Bouter, & Zeegers, 2017). Finally, philosophers and scientists make many choices over the course of a project that culminates in a manuscript that is to be published (e.g., Gelman & Loken, 2013; Wicherts et al., 2016). These choices can significantly affect the end-result reported in the manuscript, but generally remain hidden from the reader. More importantly, our peer review system is ill equipped in flagging instances where such hidden subjectivity is (intentionally) abused or has undesired consequences (Heathers, 2020). In short, (unintended) mistakes, flaws, and errors abound that have the nasty habit of evading detection.

One way to combat these problems is to increase transparency. Mistakes and

[10]Prominent biases are *confirmation bias*, the tendency to be uncritical about that which confirms one's beliefs and ignore or be overly critical of that which contradicts it, and *hindsight bias*, the tendency to interpret events that have already taken place to be more predictable than they actually were.

problematic choices can be identified when, for empirical research, data and analysis code is made (publicly) available (Vazire, 2017) or, for formal modeling or mathematical proofs, a progress journal is made available (Lee et al., 2019). Th ease with which the robustness of results and conclusions are evaluated can be improved by increasing transparency concerning methodology and uncertainty (Aczel et al., 2020). Empirical research has the additional option of preregistration. By specifying methods and hypotheses upfront, much of the deleterious effects of psychological biases and hidden subjectivity can be prevented (Lakens, 2019b). A special case is the *registered report*, were studies, before they are conducted, are registered and accepted for publication on their methodological and theoretical merits alone (Chambers, Dienes, McIntosh, Rotshtein, & Willmes, 2015). In brief, increasing openness and transparency will boost the ease with which we detect mistakes, flaws, and errors and the easier this becomes the more vigilant we might become concerning our own errors before they are detected by others.

Another way to combat these problems is to investigate and improve methodological quality. In the most basic sense, this can be accomplished by the development and use of tool for error detection in reporting of statistical results (e.g., N. J. L. Brown & Heathers, 2017; Heathers, Anaya, van der Zee, & Brown, 2018; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016) and checking their validity (e.g., Nuijten, van Assen, Hartgerink, Epskamp, & Wicherts, 2017). Additional contributions can be made by quality assessment of measurement and methods (e.g., Downs & Black, 1998a; Reitsma et al., 2009) and evaluation of validity (e.g., Hussey & Hughes, 2020). Next, studies can and are being conducted on how error-rates can be inflated or results can be biased (e.g., Schimmack, 2020; Simmons et al., 2011) and tools can and have been developed for detection such inflation and bias (e.g., Bartoš & Schimmack, 2020; Simonsohn, Nelson, & Simmons, 2014). Furthermore, systematic schemes are proposed for a process of direct replication and extension[11] studies to test hypotheses and their auxiliary assumptions (e.g., Tunç & Tunç, 2020). Finally, methods are being proposed to systematically develop and improve scientific theories in tandem with empirical research which can then inform each other on their qualities and deficiencies (e.g., Borsboom et al., 2020; Oberauer & Lewandowsky, 2019).

From a birds-eye view, these developments in the improvement of scientific practice appear to embody properties of *scientific objectivity* (Reiss & Sprenger, 2014) and *severity* (Mayo, 2018; Popper, 2002[1959]). Specifically, they concern the identification

[11] Extension or conceptual replication studies, are studies where methodological changes are made with respect to the original study in terms of design, measurement, and population. The purpose is to test the extend to which the interpretation of the results can be generalized.

and prevention of biasing influences (i.e., the effects of non-epistemic values) and increasing the strength with which scientific hypotheses are tested (Chapters 5 and 6 for further details on these concepts). It is my opinion that philosophy's role in these developments is or at least could be two-fold. On the one hand, philosophy is the *beneficiary* of these developments. Experimental philosophy benefits directly from these developments, because of its methodological improvement. Philosophy proper can benefit indirectly, because the assumptions, intuitions, and axioms on which analyses and arguments are based are at least partially empirical and can thus benefit from improvements in their empirical examination. On the other hand, philosophy can be a *contributor* to these developments. Many if not all developments are based on speculations, principles, assumptions, and arguments that lie within the purview of philosophy. It can and already has contributed to some extent (e.g., Muthukrishna & Henrich, 2019; Romero, 2018; Tunç & Tunç, 2020) to founding, guiding, and improving these developments. It is my hope that philosophy will expand and continue to fulfil this role and that this dissertation makes a small and meaningful contribution to this end.

1.5. Aims, Outline, and Methods

The purpose of this dissertation can be summarized as small set of modest attempts to contribute to improving scientific practice. Each of these attempts was geared towards either increasing understanding of a particular problem or making a contribution to how science can be practiced. The general focus was on philosophical nuance while remaining methodologically practicable. The next five chapters, are both methodologically and philosophically diverse. The first three (Chapters 2 through 4) are more empirical in nature and are focused on understanding and evaluating how science is practiced. The last two (Chapters 5 and 6) are focused on the improvement of scientific practice by providing tools for the improvement of empirical research with a strong philosophical foundation.

For the next chapter, my colleagues and I conducted a systematic review on the semantic intuition literature (Section 1.2) of experimental philosophy (Chapter 2). Specifically, we carried out a meta-analysis and methodological quality appraisal of all pertinent studies executed between 2004 and 2018. The purpose of this review was to assess the overall evidential value of the literature and to bring methodological strengths and weaknesses into the limelight.

Chapter 3 is a report on a systematic review on literature on the null hypothesis significance test (NHST). NHST has been standard practice in the social sciences for

decades even though in has been criticized as a deficient method for just as long (e.g., Berkson, 1942). In spite of this criticism and the recently arrived replication crisis, NHST continues to be practiced. Criticism on the method has been haphazardly published as essays without providing a clear overview of its qualities and deficiencies. To promote understanding, my colleague and I conducted a systematic review of this literature. The map this provides is expected to prove useful in the understanding and improvement of practice of statistical methods.

Chapter 4 is a report of an empirical study on how statisticians conduct analyses and evaluate each others work. In order to improve understanding of scientific practice, my colleagues and I observed applied statisticians in their analysis and interpretation of two simple data sets. These analyses were followed by a round-table discussion on their analyses, results, and statistical methods in general. These results provide insight into expert practice and exposes their similarities and disagreements.

For Chapter 5, my colleague and I investigated the concept of scientific objectivity. Specifically, we intended to develop a notion of objectivity that could be put into practice my empirical researchers. To this end, on the one hand, we evaluated the current philosophical literature on the concept of objectivity. On the other hand, we scrutinized the methodological literature on practice that could endanger objectivity. As a result, we propose a philosophically nuanced and methodologically informed conception of objectivity that can be used with relative ease.

In Chapter 6, my colleagues and I assessed to concept of a *severe test* from a Bayesian perspective. Popper (2002[1959]) never defined his notion of a severe tests that a theory should survive in order to be corroborated. However, it could be considered a valuable concepts, as several methodologists explicitly or implicitly adopted it (e.g., Meehl, 1978; Roberts & Pashler, 2000). Mayo (1996, 2018) incorporated the concept in her approach to frequentist statistics. However, some have considered the result wanting, too fare removed from Popper's (2002[1959]) original notion, and difficult to practice. Thus, we attempted to provide a Bayesian interpretation of the severity concept that can be practiced with ease by empirical researchers and captures the relevant properties as described in Popper's (2002[1959]) philosophy.

To conclude, in the next chapters, philosophy, methodology, and empirical research are employed in assisting improvement in scientific practice. This improvement is to be attained via contribution to understanding and increasing protection against scientific error and problematic practice. The results of these endeavours are summarized and integrated in the final chapter.

2

SEMANTIC INTUITIONS

Abstract:
The finding that intuitions about the reference of proper names vary cross-culturally (Machery, Mallon, Nichols, & Stich, 2004) was one of the early milestones in experimental philosophy. Many follow-up studies investigated the scope and magnitude of such cross-cultural effects, but our paper provides the first systematic meta-analysis of studies replicating Machery et al. (2004). In the light of our results, we assess the existence and significance of cross-cultural effects for intuitions about the reference of proper names.

This chapter is adapted from van Dongen, N.N.N., Colombo, M., Romero, C.F., & Sprenger, J.M. (2020). Intuitions about the reference of proper names: A meta-analysis, 1-30. *Review of Philosophy and Psychology*.

2.1. Introduction

Most people who have heard of the Italian mathematician Giuseppe Peano credit him with inventing the standard axioms of arithmetic. This is all they associate with the name "Peano". Yet, the axioms were invented by the German mathematician Richard Dedekind, and Peano published a simplified version only afterwards. If people identify Peano only by the description "the inventor of the standard axioms of arithmetic", to whom are they referring when they use the name "Peano"? To Peano or to Dedekind? And more generally, what kind of meaning must people associate to a proper name like "Peano" in order to be competent users of that name?

In philosophy of language, there are two main classes of theories about the meaning of proper names: descriptivist theories and causal-historical theories. According to descriptivist theories (Frege, 1892; Russell, 1905; Searle, 1958), proper names have definite descriptions as their meaning. The idea is that a proper name can refer to a person only via the *descriptive properties that users of the name associate with it*. Thus, people who identify Peano only by the description "the inventor of the standard axioms of arithmetic" would actually refer to Dedekind when they use the name "Peano". After all, it is Dedekind who uniquely satisfies that description.

According to causal-historical theories (Kripke, 1980), proper names do not imply any descriptive property of the individuals to which they refer. Proper names refer directly to their bearers without being essentially associated with any descriptive properties of an individual. People processing a proper name like "Peano" certainly rely on some mental representations of descriptive properties, but these representations play no role in determining the meaning of "Peano". Instead, what is crucial for determining the meaning of a proper name is its *causal history*. All uses of the name that causally derive from an original act of baptism refer to the individual originally baptized with that name. Even if people falsely associate the description "the inventor of the standard axioms of arithmetic" with Peano, the proper name "Peano" actually refers to Peano. This is because of a relevant causal chain between the original act of baptism for Giuseppe Peano, and people's usage of the name "Peano".

One influential argument for why proper names cannot be semantically equivalent to definite descriptions is that the referent of a proper name "is stipulated to be a single object whether we are speaking of the actual world or a counterfactual situation" (Kripke, 1980, p. 21). In contrast, definite descriptions can refer to different individuals in counterfactual situations.

Kripke illustrates this argument with an example analogous to the Peano case

above (Kripke, 1980, p. 83). Suppose that the only description you associate with the name "Gödel" is "the mathematician who proved the incompleteness of arithmetic". Suppose that you discover that a certain Mr. Schmidt rather than Gödel actually proved the incompleteness of arithmetic. If the name "Gödel" is semantically equivalent to the definite description "the mathematician who proved the incompleteness of arithmetic", then you are committed to the conclusion that the name "Gödel" actually refers to Mr. Schmidt. But, of course, this conclusion is false, which suggests that descriptivist theories of proper names are false.

In a landmark study, Machery, Mallon, Nichols, and Stich (2004) used this Gödel case, and three other similar vignettes, to explore the question of whether judgments about the referents of proper names are cross-culturally robust. They found that Westerners tend to have intuitions in line with Kripke's causal-historical theory, while East Asians' intuitions tend to agree with the descriptivist theory. On the basis of this finding, Machery et al. reasoned that if people's judgments about the meaning of a proper name are systematically influenced by demographic variables like their culture, then Kripke's and other Anglophone speakers' semantic judgments cannot constitute reliable evidence for a theory of proper names.

Several follow-up studies have extended and probed the original result, and embedded it into a broader methodological discourse about the use of intuitions about particular cases as evidence for philosophical theories (e.g., Cova et al., 2018; Lam, 2010; Machery, 2017; Machery, Olivola, & LeBlanc, 2009; Sytsma & Livengood, 2011). Machery et al.'s effect has been found to show significant variation both within and across studies, but no convincing explanation has been offered for when and why we should expect cross-cultural variation in semantic intuitions. It might be that scenarios like the Peano or Gödel cases above are not the right tools for eliciting people's semantic intuitions. Or, it might be that for making semantic judgments, people have access to multiple cognitive strategies, which include both causal-historical and descriptivist factors, and shift between them depending on one's ease to take the perspective of the speaker, background knowledge, audience, and purpose of communication (Genone & Lombrozo, 2012).

Philosophers are not the only ones interested in proper names. Proper names have also been studied by several other disciplines, including linguistics, psychology, neuroscience, and anthropology.

Linguists agree that proper names are mainly used to identify individuals uniquely. However, the referential and connotative use of a proper name depends on the communicative intentions the speaker wants to convey to the hearer, the

knowledge frame shared by the interlocutors, and their discourse-relative perspective (Dancygier, 2009; Dancygier & Vandelanotte, 2017; Marmaridou, 2000).

Cognitive psychologists focus on the mental representations of the meaning of proper names, and in particular on differences between the cognitive processing of proper names and common nouns (Sophia & Marmaridou, 1989; Valentine, Brennen, & Brédart, 1996). Much of this research provides evidence that proper names and common nouns are associated with different mental representations, and that they are processed differently. It has been found that, generally, people from different cultures judge more quickly if a word is a proper name as opposed to a common noun (Müller, 2010), and that they retrieve the meaning of a proper name from memory with more difficulty compared to common nouns (Proverbio, Mariani, Zani, & Adorni, 2009; Wang, Verdonschot, & Yang, 2016). Additional evidence for these processing differences between proper names and common nouns comes from neuropsychological studies, which have shown a double dissociation between retrieval of proper names and retrieval of common nouns (Semenza, 2006; Yen, 2006).

Finally, for anthropologists, the meaning of proper names depends on principles governing naming practices across cultures (Bright, 2003; vom Bruck & Bodenhorn, 2006). Proper names have been found to fulfil two main cultural functions: to identify their bearers differentiating them from other individuals, and to classify them in terms of their parental, economic, ethnic or geographical group. These two functions can trade off, since the more a proper name differentiates its bearer from other individuals, the less social information it carries. Focusing on these trade-offs, Alford (1988) examined the use of proper names in sixty cultures around the world to better understand the social and communicative situations where one function is more prevalent than the other.

Given this rich and varied literature, it is noteworthy that the philosophical literature studying cross-cultural variation of proper names has overlooked the connection to studies in other disciplines about how people use, memorize, recall, and mentally represent the meaning of proper names.[1] At the very least, studies from other disciplines, which employ various methods to study proper names, would provide experimental philosophers with independent evidence to probe the robustness of putative cross-cultural effects.

Locating experimental philosophers' work in the wider scientific context high-

[1]This is the more surprising because the original study and hypotheses tested by (Machery et al., 2004) were studies in cultural psychology (Nisbett, Peng, Choi, & Norenzayan, 2001). These studies suggested that different cultural groups exhibit different styles of thought: while "holistic thought" would be prevalent among East Asians, "analytic thought" would be predominant among Westerners.

lights the general importance of the questions we set out to address with our meta-analysis in this paper. For example, to what extent are the demographic effects observed by experimental philosophers large and interpretable? Do they depend on superficial features of the experimental material and design experimental philosophers have been using? Or do they indicate variation in the mechanisms and representations, which different communities of speakers would recruit to make sense of proper names? Answering these questions is not only important for philosophers; it will also clarify whether researchers from different disciplines are warranted to rely on the demographic effect from Machery et al. 2004, and on simple vignettes like the Gödel case for studying proper names.

A relatively large and interpretable meta-analytic effect will be convincing reason that the cross-cultural effect is not an artefact of particular vignettes. The finding that a descriptivist theory of proper names predicts East Asians' intuitions, while a causal-historical theory predicts Westerners' intuitions, will motivate new hypotheses that cognitive psychologists and neuroscientists could test. For example, as noted above, psychologists and neuropsychologists have proposed that the retrieval of proper names is particularly difficult compared to common nouns. One reason for this hypothesis is that proper names would be detached from the semantic network representing descriptive or biographical information (e.g., Semenza, 2006). But, if East Asian speakers have descriptivist intuitions, then their retrieval of proper names should be easier compared to the retrieval of proper names exhibited by Western speakers, whose mental representations of proper names would be detached from descriptive information.

A small, noisy, and hard-to-interpret meta-analytic effect will give us reason to either call into question the idea that demographic factors play a substantial role in semantic intuitions, or doubt that the prevalent vignettes are reliable tools to elicit such intuitions. Either way, this finding should motivate experimental philosophers interested in the semantics of proper names to pay closer attention to relevant anthropological evidence about different functions of proper names, as well as to the experimental designs cognitive psychologists and linguists have been using to study proper names.

The central elements of our meta-analysis are the questions of whether Westerners' and East Asians' intuitions about the reference of person names are in line with either the causal-historical or the descriptivist theory (Hypotheses 1a and 2a), and how the two populations compare in this respect (Hypothesis 3a).[2] We evaluate

[2]Compare the following quote from the original paper (Machery et al., 2004, p. B5): "W[esterner]s would be more likely to respond in accordance with causal-historical accounts of reference, while

these hypotheses on the basis of a set of empirical studies, which rely on probes similar to the Gödel and Peano cases we have described above. To answer these questions, we test three hypotheses:

Hypothesis 1a The majority of Westerners have causal-historical intuitions.

Hypothesis 2a The majority of East Asians have descriptivist intuitions.

Hypothesis 3a Westerners are more likely than East Asians to have causal-historical intuitions.

There are two types of vignettes in the literature we aggregate: first, the *Gödel probe*, which is structurally identical to our introductory Peano/Dedekind example, and the *Jonah probe*, where the properties associated with a proper name gradually shift over time. In the Gödel probe both the causal-historical and the descriptivist theory single out a specific referent of the proper name, while in the Jonah probe the proper name does not refer at all according to the descriptivist account. Since both probes follow the same experimental design (exposure to vignette, binary response, two theories with different predictions, etc.), the results can be aggregated to test hypotheses 1a–3a. Moreover, since the Gödel probe has particular significance in the philosophical literature (going back to Kripke, 1980), and since the original study found a notable effect of cross-cultural variation only for Gödel probes, but not for Jonah probes (Machery et al., 2004), we also test the following three hypotheses:

Hypothesis 1b The majority of Westerners have causal-historical intuitions in the Gödel probes.

Hypothesis 2b The majority of East Asians have descriptivist intuitions in Gödel probes.

Hypothesis 3b Westerners are more likely than East Asians to have causal-historical intuitions in Gödel probes.

One final note before we describe our meta-analysis in detail. Different studies in the literature originated from Machery et al. 2004's work differ in the samples of Westerners and East Asians. Machery et al. (2004) tested participants in the United States and Hong Kong. Follow-up studies use for example Dutch participants (Cova et al., 2018) or French participants (Machery et al., 2009). We note these differences in detail in Section 2.3.1. Nonetheless, we formulated the hypotheses in a general form as a comparison between Westerners and East Asians for two reasons. First, we intended to study the effect as reported in Machery et al. 2004 and hence we followed their design. Second, we did not have theoretical reasons to exclude e.g., Dutch or French from the group of Westerners or e.g., Japanese from the group of East Asians.

E[ast] A[sian]s would be more likely to respond in accordance with descriptivist accounts of reference."

2.2. Methods

2.2.1. Data sources and searches

We conducted a comprehensive literature search of studies on cross-cultural semantic intuitions. To find replications of the original experiment, we started with the Google Scholar list of all papers citing Machery et al. 2004 and checked whether they contained experimental data. Our search was aided by a list of known replications that we obtained via e-mail from Édouard Machery (the first author of the original study). We used this search strategy because any replication of Machery et al. 2004 would, in virtue of its aims and scope, refer to the original paper.

2.2.2. Study selection

Studies were eligible if they were published or publicly available in English, and used a design sufficiently similar to the original study. Specifically, eligible studies had to contain results from experiments featuring East Asian and/or Western participants with one or more binary-choice probe. The binary-choice answer to such probes had to correspond to the descriptivist and the causal-historical theory of reference, respectively.

2.2.3. Data extraction

The eligible studies were classified by two teams of authors (NvD/MC and FR/JS). For each study, team members independently extracted the name of the first author, the year of publication, and the data of the probe responses. Per study and probe, they independently extracted data on the type of the probe (e.g., Gödel, Jonah, etc.), the number of Western and East Asian participants, total sample size, the number of causal-historical responses (per subgroup and in total), and deviations from the original design, such as language of the probe, phrasing of the question, and phrasing of the answers. Disagreements were resolved by consensus.[3]

2.2.4. Data synthesis and analysis

We carried out a confirmatory (Section 2.3.2) and an exploratory analysis (Section 2.3.3), as well as a quality appraisal (Section 2.3.4 and Appendix A.3). A confirmatory analysis assessed the overall meta-analytic evidence and its generalisability for the main hypotheses tested in Machery et al. 2004 (Hypotheses 1a–3a) and the meta-analytic evidence for the particular Gödel probes (Hypotheses 1b–3b) for which Machery et al. 2004 reported positive results. Specifically, for Hypotheses 1b-3b, only

[3]The full data spreadsheet is available online at `https://osf.io/et86f/`.

Gödel probes were used, which did not deviate from the design of the original study (i.e., direct replications). The analyses were conducted on the level of the individual probes within a study, because not all studies had the same (number of) probes. We used a multilevel random effects (RE) model to capture the hierarchical structure (i.e., probes within studies) of the data and the inter-study dependencies between probes via the 'metafor' package in R (Viechtbauer, 2010).

In total, we performed six confirmatory meta-analyses, one per hypothesis. Specifically, we calculated summary proportion estimates of Causal/Historical response for single cultural group responses (Hypotheses 1a/b, 2a/b) with a restricted maximum-likelihood heterogeneity estimator[4] to model the random effects. For the probes that compare Westerners to East Asians (Hypotheses 3a and 3b), we calculated summary Relative Risk ratios (RR), that is, the quotient of the proportions of Causal/Historical responses in both groups: If A and B are the two groups and p_A and p_B are the proportions of event E in the groups (in our case: Causal/Historical responses), then the relative risk ratio is defined as $\mathrm{RR}_{\mathrm{E}}(A, B) = p_A/p_B$.[5] The extent of heterogeneity between probes was assessed by the I^2 measure (indicating the percentage of total variance due to between-probe variance; Higgins, Thompson, Deeks, & Altman, 2003; Ioannidis, Patsopoulos, & Evangelou, 2007). In addition to the 95% confidence intervals (CIs) for the unknown parameters (i.e., the proportion of causal-historical responses and the RR between populations), we calculated the 95% prediction intervals (PI). Unlike CIs, which are based on compatibility of the observed data with an unknown parameter value, PIs predict the distribution of future data points by taking into account inter-study variability. Therefore, they are better suited to express a plausible range of values for the next conducted study, and to predict whether effects are likely to replicate (Higgins, Thompson, & Spiegelhalter, 2009; Riley, Higgins, & Deeks, 2011).

The exploratory analysis focused on the observed variance in effect-size between the various probes; that is, it aimed to identify factors that would explain why studies often deliver heterogeneous results. Specifically, we repeated the tests performed for Hypotheses 1a–3a with a meta-regression analysis where deviations from the original design were included as moderator variables. We ignored the hierarchical structure of the data because we were interested in explaining variance by the general design factors.For, taking account of hierarchical structures allows one to assess the

[4]This estimator is approximately unbiased and efficient(Raudenbush, 2009; Viechtbauer, 2005), and is the software package's default (Viechtbauer, 2010).

[5]Relative Risk is a well-probed measures of association between categorical data that is easy to interpret and using a different effect-size measure like Odds Ratio does not change the inference (Higgins & Green, 2011).

associations between level, but ignoring this structure has a negligible effect on the main effect estimates (O'Mara, 2008).

2.3. Results

2.3.1. Studies and data

Via Google Scholar, we identified 482 records published in 2017 or earlier that cite Machery et al. 2004. Using the search criteria described in Section 2.2.1, we ended up with 15 potential studies. Four studies were added in addition to our initial search results, resulting in a total of 19 studies.[6]

Eventually, from this set of 19 studies, 13 were included and 6 were excluded. Reasons for exclusion were the following: [1] missing data, [2] culturally mixed samples, and [3] structure of the data (i.e., data from non-binary questions, compare Section 2.2.2). See Appendix A.1 for the list of excluded papers and their reason for exclusion. In addition, the Indian sample from Machery 2009 was excluded, because they are considered to be South Asian instead of East Asian. Combined, these studies tested 61 probes on 4691 participants, who produced a total of 8959 binary responses. Of these probes, 35 tested both Western and East Asian samples [median sample size: 181]; 15 probes tested Western samples [median sample size: 60]; and 11 probes tested East Asian samples [median sample size: 211]. Table 2.1 provides an overview of the number of probes per study and their characteristics. Most Western samples were from the United States of America. Exceptions were 'Machery, 2009' (France), 'Cova, 2018' (The Netherlands), and 'Colombo, n.p.' (The Netherlands, USA, England, Germany, and Italy). In general, the East-Asian samples were from Hong Kong. Exceptions were 'Machery, 2009' (Mongolia), 'Sytsma, 2015' (Japan), 'Kazaki, 2017' (Japan), 'Izumi, 2018' (Japan), and 'Colombo, n.p.' (Hong Kong and China). We identified ten factors on which probes could deviate from the original design (Appendix A.2), from the language in which probes were presented to the phrasing of the question that participants were asked.[7]

In particular, the studies in our data set did not always use the same phrasing for eliciting a judgment on causal-historical versus descriptivist intuitions. Most studies followed the original design by Machery et al. (2004) and asked participants the following question: Who is the person that the vignette character John is "talking

[6] In personal communication, Édouard Machery suggested to add the original results reported in Machery, Sytsma, and Deutsch 2015 and a forthcoming by Izumi et al. (2018). In addition, we included the results of our replication attempt reported in Cova et al. 2018 and an online replication attempt carried out via Qualtrics, whose details are also available online at `https://osf.io/et86f/`.

[7] The complete list of the design deviation factors and their values is given in the data extraction plan, which is available online in the additional materials file.

Table 2.1: Overview of the Studies and Their Characteristics.

First author	Year	Study Type	Study Design	Number of probes	Sample Size	Design Deviations
Machery	2004	Comparison	Within-subjects	4 (Gödel, Jonah)	72	No (original study)
Lam	2009	Comparison	Within-subjects	3 (Shakespeare, Julius)	69	Yes (2 deviations)
Machery	2009	Comparison	Between-subjects	2 (Gödel, Shakespeare)	144	Yes (2 deviations)
Machery	2010	Asians only	Single probe	1 (Gödel)	19	Yes (1 deviation)
		Comparison	Single probe	1 (Gödel)	153	Yes (1 deviation)
Sytsma	2011	Westerners only	Within-subjects	3 (Gödel)	35	Yes (1 deviation)
		Westerners only	Between-subjects	3 (Gödel)	223	Yes (1 deviation)
Beebe	2015	Westerners only	Between-subjects	6 (Satyricon)	304	Yes (1 deviation)
		Westerners only	Between-subjects	3 (Satyricon)	180	Yes (2 deviations
		Comparison	Within-subjects	4 (Satyricon)	177	Yes (1 deviation)
Machery	2015	Asians only	Single probe	1 (Gödel)	65	Yes (2 deviations)
		Comparison	Within-subjects	2 (Gödel)	129	Yes (2 deviations)
Sytsma	2015	Comparison	Between-subjects	3 (Gödel)	621	Yes (2 deviations)
Beebe	2016	Comparison	Within-subjects	4 (Gödel, Jonah)	207	Yes (1 deviation)
	2016	Comparison	Within-subjects	4 (Gödel, Jonah)	406	Yes (1 deviation)
Kazaki	2017	Asians only	Within-subjects	2 (Gödel)	231	Yes (3 deviations)
Izumi	2018	Asians only	Between-subjects	7 (Gödel)	1283	Yes (2 deviations)
Cova	2018	Comparison	Within-subjects	4 (Gödel, Jonah)	181	No (only sample deviation)
Colombo	n.p.	Comparison	Within-subjects	4 (Gödel, Jonah)	192	No (only sample deviation)

about" when using the proper name "Peano"? By contrast, Sytsma and Livengood (2011) suggest rephrasing the question in the following way: Who would you *take* is the person that John is talking about when using the proper name "Peano"? (this modification is known as the "clarified narrator's perspective", see also Section 2.3.3), and Machery et al. (2009) asked for the *truth value* of sentences such as "Peano discovered the incompleteness of mathematics". Similarly, answers to the binary question were sometimes phrased as descriptions (e.g., in the original study) and sometimes as bare noun phrases without further explanations (e.g., in Lam, 2010). For the purpose of the meta-analysis, we treated all these dependent variables as answering the same question, independently of the exact phrasing. However, for testing Hypothesis 1b-3b, we focused on direct replications of the original Gödel probe and excluded variations of the design such as Sytsma and Livengood's "clarified narrator's perspective". These different ways of analyzing the data allow us to answer two different questions: first, the question of the replicability and internal validity of the original effect observed by Machery et al., and second, the generality and external validity of the observed effect to a wider hypothesis about the reference of proper names.

2.3.2. Meta-Analytic Estimates

The results per hypothesis are displayed as forest plots in Figures 2.1 and 2.2 and summarized below.[8] The analyses provide confirmatory evidence for Hypotheses 1a,

[8]These plots show the distribution of effect-sizes (location, confidence intervals, weight in the meta-analysis) of the probes that were used for the analysis, their summary meta-analytic effect-size (diamond at the bottom), and its confidence interval (width of the diamond). The forest plots for Hypothesis 1a and 2a can be found in the online supplementary files. Please note that the plots show log-transformed effect-sizes. The untransformed effect-sizes are reported in the main text.

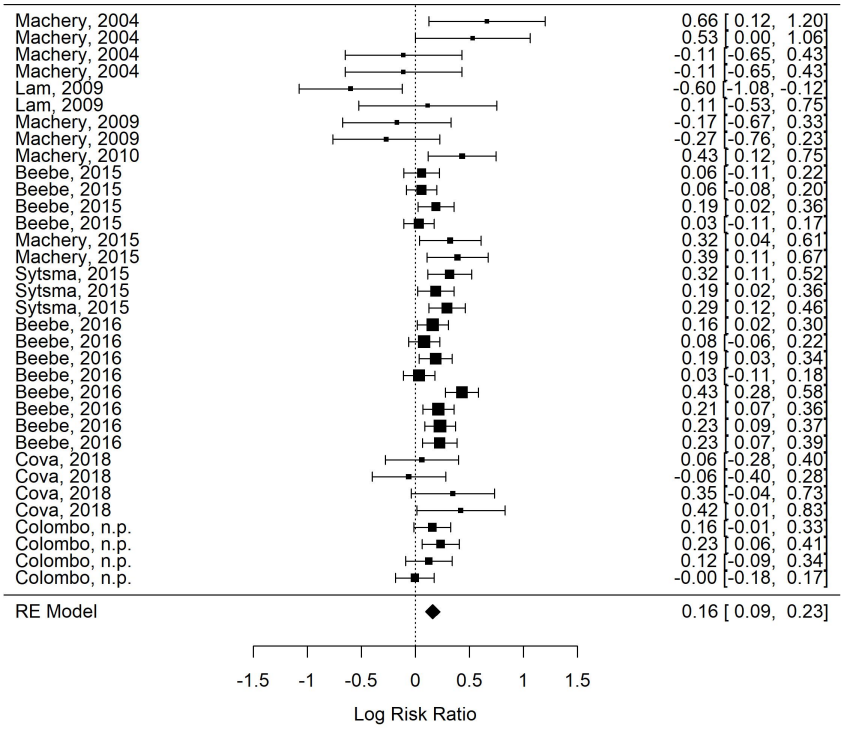


Figure 2.1: Forest plot for Hypotheses 3a, showing distribution of effect sizes and confidence intervals.

2b, 3a and 3b in the sense that the confidence interval for the unknown parameter consists only of values in the expected direction, and excludes the hypothesis of zero effect—that is, a relative risk ratio of RR = 1 for hypotheses 3a and 3b, and a 50% proportion of Causal/Historical responses for hypotheses 1a through 2b. Forest plots for all six hypotheses are given in Figure 2.1 and 2.2. Below is a precise statement of our results.

Hypothesis 1a: The summary proportion of Causal-Historical probe responses in Westerners was 0.58 [95% CI: 0.52–0.62] with large observed heterogeneity [I^2 = 82.10%] and a prediction interval that included zero-effect [95% PI: 0.35–0.77].

Hypothesis 1b: The summary proportion of Causal/Historical responses to Gödel probes in Westerners (in the original experimental design) was 0.55 [95% CI: 0.50–0.59] with moderate observed heterogeneity [I^2 = 21.89%, 95% PI:

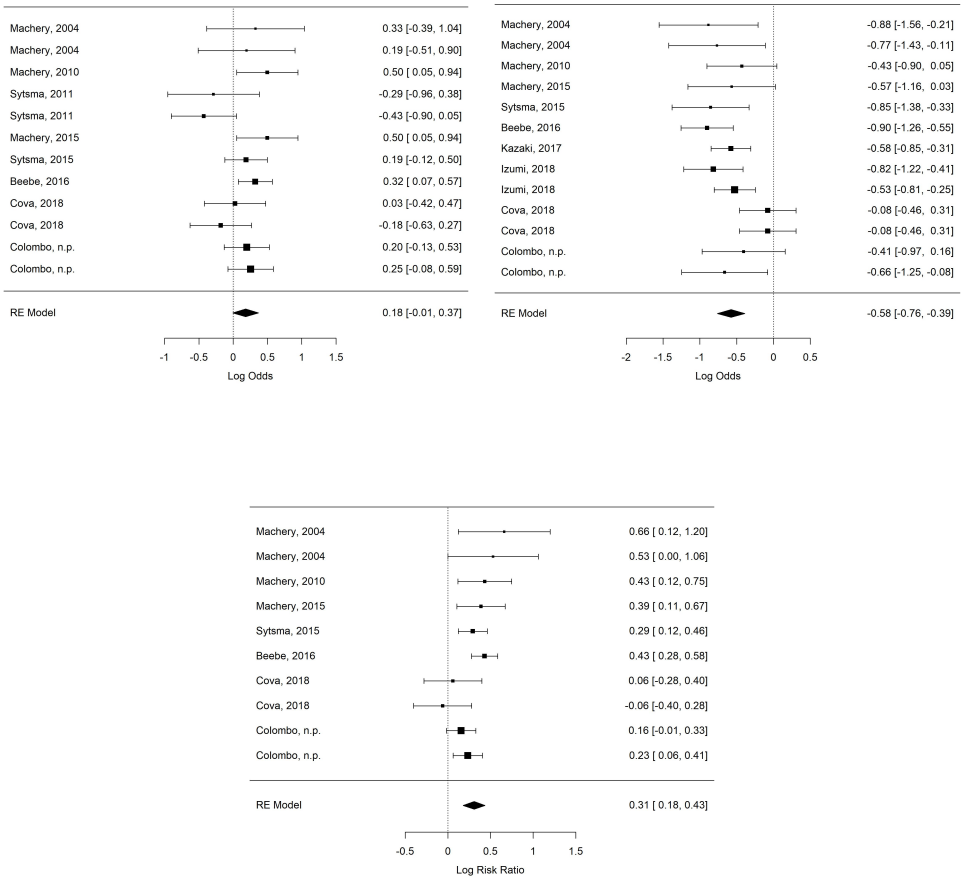


Figure 2.2: Forest plot for Hypotheses 1b, 2b and 3b, respectively, showing distribution of effect sizes and confidence intervals.

0.35–0.77].

Hypothesis 2a: The summary proportion of Causal-Historical probe responses in East Asians was 0.50 [95% CI: 0.41–0.59] with large observed heterogeneity [I^2 = 87.50%] and prediction interval [95% PI: 0.19–0.80].

Hypothesis 2b: The summary proportion of Causal-Historical to Gödel probes in East Asians (in the original experimental design) was 0.36 [95% CI: 0.32–0.41] with large observed heterogeneity [I^2 = 52.59%], but a prediction interval that excluded the zero-effect value [95% PI: 0.26–0.48].

Hypothesis 3a: The summary RR of Causal-Historical probe responses in Westerners versus East Asians was 1.18 [95% CI: 1.09–1.27] with moderate-to-large observed heterogeneity [I^2 = 47.42%] and a prediction interval that included zero-effect [95% PI: 0.96–1.44].

Hypothesis 3b: The summary RR of Causal-Historical responses to Gödel probes in Westerners versus East Asians (in the original experimental design) was 1.36 [95% CI: 1.20–1.54] with moderate observed heterogeneity [I^2 = 30.08%] and a prediction interval that excluded the zero-effect value [95% PI: 1.03–1.80].

In addition, we wanted to make sure that these results did not depend on how we modelled the structure of the data. Most studies included in the meta-analysis contain several probes. Some studies used one sample of participants to test several probes (i.e., repeated measure design), while other studies used separate samples of participants to test each probe individually (i.e., between-subject design). Our analyses take into account the hierarchical structure of probes nested in studies, but they neglect the potential dependencies between probes in studies with repeated measures design.

To take account of these differences, we ran two additional sets of analyses. For the first analyses, we used a flat data structure (i.e., a regular RE meta-analysis). For the second analyses, we also used a regular RE meta-analysis, but included the average effect size and standard error of probe clusters with shared repeated measure design. For instance, if a study contained one sample of (Western) participants (and/or one sample of Asian participants) that responded to four probes (e.g., the two Gödel and two Jonah probes of the original study by Machery et al. 2004), then only the average of their response was included in the analyses. The results of these analyses were similar to those of the analyses with the hierarchical data structure, thereby warranting the same conclusion (code and results of the these analyses can be found in the additional materials file).

2.3.3. Results of exploratory analyses

The width of the confidence and prediction intervals in the hypotheses tests (Section 2.3.2) point to high observed inter-study variance of participants' responses. By means of a meta-analytic regression, we explored whether this variance could be explained by specific deviations in study design, such as language of the participants (Appendix A.2 for a list of these deviations and their descriptions). Concretely, we fitted three generalized meta-analytic regressions with the Knapp and Hartung method[9] (Knapp & Hartung, 2003; Viechtbauer, 2010) for the data used to test hypotheses 1a, 2a, and 3a with the identified design deviation factors as predictors and log-transformed RR (hypothesis 3a data) or Odds Ratio (hypotheses 1a and 2a data) as outcome. We fixed the sampling variance to zero, because we were only considering variance between probe outcomes. This choice is acceptable, since we are not estimating an overall effect-size weighted by the precision of each datum in the analysis. However, we did include the sample size in the model as a predictor.

For all three analyses, much of the variance was explained by the model.[10] About 47% of the variance in the probe responses by Westerners could be accounted for by deviations in study design [$R^2 = 0.47$, $F(18,30) = 3.34$, $p = 0.002$]. For probe responses by East Asians, about 64% of the variance could be accounted for [$R^2 = 0.64$, $F(22,23) = 4.60$, $p = 0.0003$]. In the comparison in probe responses between Westerners and East Asians, about 28% of the variance could be accounted for by deviations in study design [$R^2 = 0.28$, $F(18,15) = 1.70$, $p = 0.15$].[11] Although the relatively large number of predictors in relation to the size of the data set made the results unsuited for drawing strong conclusions with certainty, a noteworthy result was that none of the design deviations, which include participant nationality and different types of probes, made a statistically significant contribution ($p < 0.05$) to all of the models.[12]

In addition, we separately analyzed studies that used an alternative formulation of the binary-response question in the Gödel probes: the "clarified narrator's perspec-

[9]This method was used because with this method the individual coefficients are based on the t-distribution and the omnibus test statistic uses an F-distribution. In other words, this methods allows for interpretation of the result similar to a regular linear regression.

[10]Because of the fixed sampling-variance, the analysis is identical to a generalized linear model with R^2 equal to adjusted-R^2.

[11]The models do not have the same number of predictors, because of group specific variables (e.g., language of participant group) and redundant dummy predictors were left out.

[12]Specifically, for the analysis with Western participants only, significant ($p < 0.05$) predictors were 'Julius 2' probe, 'Satyricon 2' probe, 'Satyricon 3', the 'good moral valance' and 'bad moral valance' manipulations, and the 'clarified narrator's perspective' question formulation. For the East Asians only analysis, significant ($p < 0.05$) predictors were the probes 'Jonah 1', 'Jonah 2', 'Julius 1', 'Julius 2' and 'Shakespeare 1', the 'good moral valance' manipulation, and the 'sono-demonstrative' answer formulation. For the analysis that compared Western and East Asian responses, significant ($p < 0.05$) predictors were the 'Gödel 2' probe and the 'Shakespeare 1' probe. For a complete overview can be found under '2. Exploratory analyses results' in the additional materials file.

tive" (e.g., Sytsma & Livengood, 2011). Based on existing literature, this formulation was expected to be less ambiguous in eliciting participants' semantic intuitions. We found six such probes in our data set, of which three compared the causal-historical and descriptivist responses between Westerners and East Asians; one tested East Asians only; and two tested Westerners only. To test the summary effect-size of these data, the same kind of analyses were used as for testing H1b-H3b (Section 2.2.4). The results are summarized in Table 2.2. The proportion of causal-historical intuitions is higher for both Westerners and East Asians in the clarified narrator's perspective than in the set of all Gödel probes. In particular, for Westerners, the effect—as measured by the proportion of causal-historical responses—is more pronounced in the "clarified narrator's perspective" while it is slightly smaller for East Asians and the comparison of East Asians and Westerners. These comparisons are relative to the outcomes of the tests of H1b–H3b in Section 2.3.2.

Table 2.2: Results of the analysis for probes with the clarified narrator's perspective proposed by Sytsma and Livengood (2011).

Analysis	Effect Type	Effect Size	95% CI	*P* value	I^2	95% PI
Westerners	proportion	0.66	0.60–0.72	$p < 0.0001$	3.27%	0.56–0.75
East Asians	proportion	0.39	0.28–0.52	$p < 0.10$	68.10%	0.20–0.62
Comparison	RR	1.31	1.17–1.47	$p < 0.0001$	0.0%	1.17–1.47

Finally, we analyzed the difference in responses to non-Gödel probes. This analysis is a complement to the test of hypothesis 3b, a comparison in responses between Westerners and East Asians on all the probes except the Gödel probes. Cultural difference is absent [RR = 0.08, 95% CI =-0.04, 0.20], but this is not particularly surprising: the analysis consisted overwhelmingly of Jonah probes which did not show a significant difference between the samples in the original study by Machery et al. (2004).

2.3.4. Evaluation of the study's results and methods

Evaluation of the results of the included papers indicate absence of small-study effects and presence of evidential value, but high inter-study variance. The p-curves are right-skewed (see Figure 2.3), which could be considered indicative of the set of studies containing evidential value (Simonsohn et al., 2014).[13] Figure 2.4 shows the funnel plots (Sterne, Sutton, Ioannidis, et al., 2011) for Hypothesis 1a-3a and

[13]It should be noted that the reliability of p-curves is negatively related with the heterogeneity between the included results, which is quite large in the case of our meta-analyses. Also, p-curves can fail to show p-hacking and publication bias in certain scenarios where it is actually present (Bishop & Thompson, 2016; Carter et al., 2019).

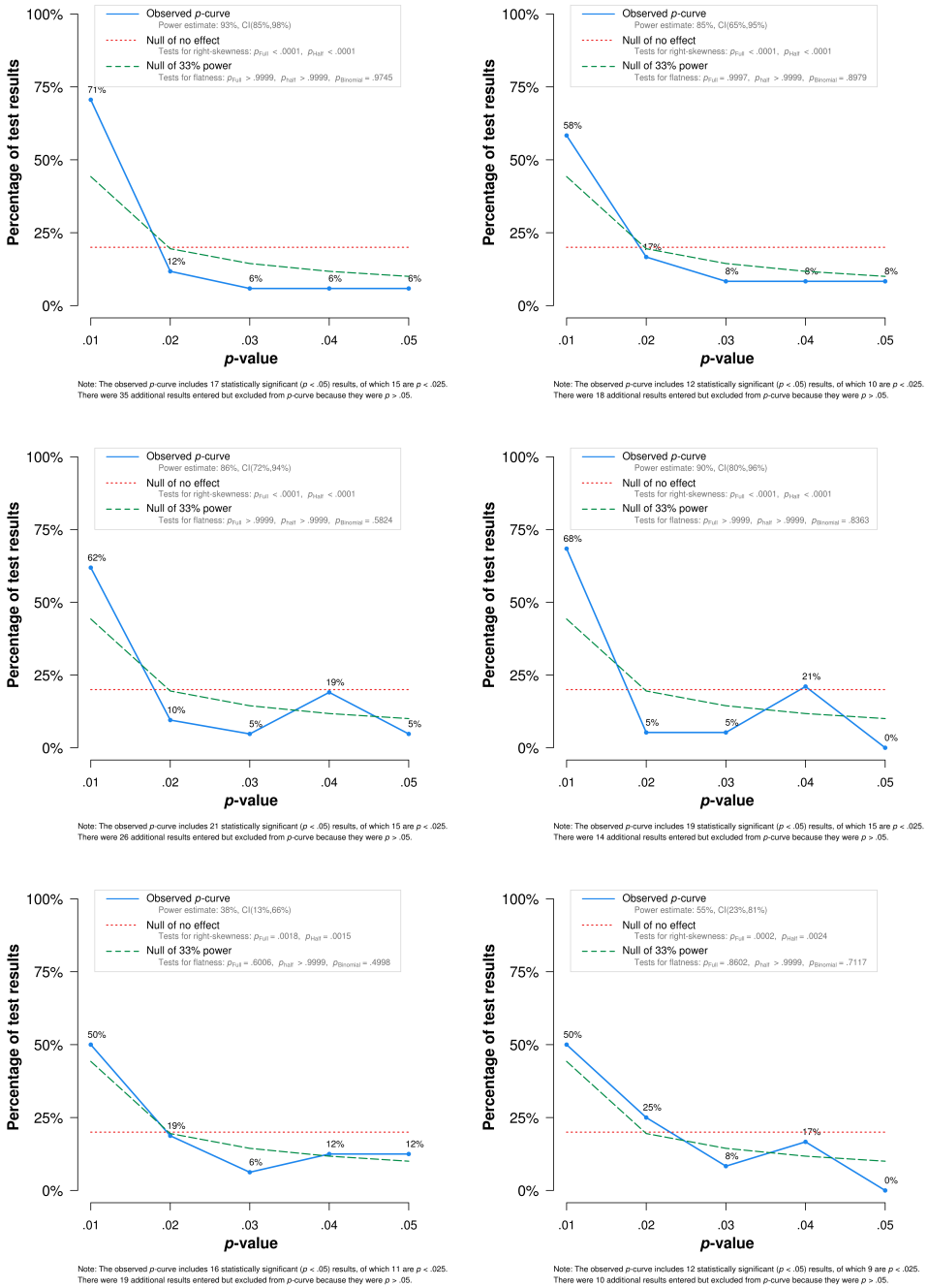


Figure 2.3: Left column: *p*-curve plots for Hypotheses 1a, 2a and 3a. Right column: *p*-curve plots for Hypotheses 1b, 2b and 3b.

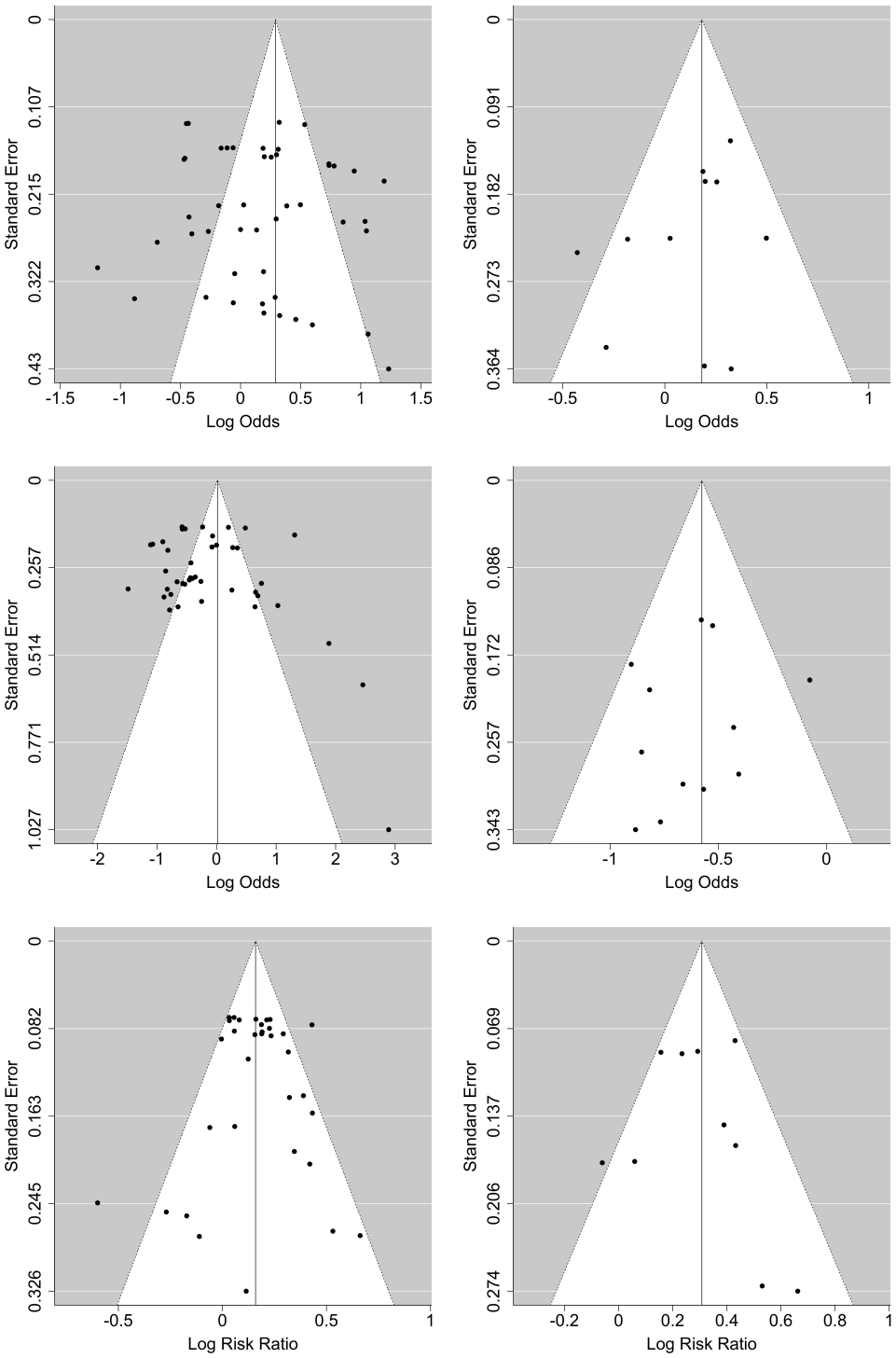


Figure 2.4: Left column: funnel plots for Hypotheses 1a, 2a and 3a. Right column: funnel plots for Hypotheses 1b, 2b and 3b. x-axis: observed effect size; y-axis: precision of study as measured by standard error.

1b-3b. Small to large studies show symmetrical distributions around the summary effect-sizes, but for Hypothesis 1a and 2a, inter-study variance is high even among studies with high precision (i.e., large sample size, low standard error). This indicates that inter-study variance does not diminish with increasing precision as it should. No such effect is observed when cross-cultural variations are directly studied, that is, in Hypothesis 3a/b. For a quality appraisal of the methods employed in the experiments, see Appendix A.3.

2.4. Discussion

All in all, our meta-analysis supports the hypothesis that cross-cultural factors affect semantic intuitions about proper names (Machery et al., 2004). For four out of the six tested hypotheses, the meta-analytic confidence interval for the summary effect size does not include the zero effect value. Neither do specific analysis tools aimed at detecting publication bias or QRPs (e.g., funnel plots, *p*-curves) provide evidence of systematic suppression of negative results. The wide prediction intervals point to high inter-study variability of the data, which cannot be consistently explained by the studies' differences in experimental design.

Our study cannot test the overall scope of cross-cultural factors in eliciting intuitions about the reference of proper names, since the meta-analysis is restricted to a very specific set of probes. In addition, some aspects of the meta-analysis are limited by the methodological design of the studies and the small number of studies with respect to the differences in design between the studies. The use of a few vignettes with binary responses per study with unknown dependencies forced us to analyze the data on the level of individual probes (vignettes) and may have contributed to the large heterogeneity. However, three findings of general interest stand out.

First, there is a notable difference between the confidence intervals (CIs) for the unknown parameter and the prediction intervals (PIs) for the next observation: only two of the six PIs do not contain zero effect. The clearest difference between the CIs and the PIs is perhaps visible in the meta-analysis of Hypothesis 1a, 1b and 2a: the 95% PIs are too wide to make a theoretically meaningful prediction. The reason is that PIs take into account how the variability in the data transfers to the expected value of future data points. In a case where individual studies scatter over a wide range of points, like in the case of Hypothesis 1a, 1b, and 2a, the PIs will be considerably wider than the CIs. The same remarks apply, though with a less pronounced effect, to the other three hypotheses. For these reasons, it is not

surprising that a recent replication attempt, included in Cova et al. 2018, has not reproduced the original effect: although there is, on average, a cross-cultural effect in the predicted direction, the inter-study variance is too high to reliably predict a result in the vicinity of the meta-analytic mean. The funnel plots of Figure 2.4 confirm this diagnosis: since variation in the observed proportions of causal-historical responses is extremely high, there seems to be a lot of "noise" in the data.

Second, the clarified narrator's perspective, suggested by Sytsma and Livengood (2011), seems to push causal-historical intuitions: it increases the tendency of participants to give causal-historical responses across the board. At the same time, it reduces the cross-cultural difference between Westerners and East Asians. However, due to the small sample size for this type of probe, our finding here should be treated with caution.

Third and last, our meta-analytic regression found statistically significant dependencies of the results on variations in experimental design. However, we could not identify predictor variables that uniformly explained inter-study variance for Westerners, East Asians and the comparison of both populations. We could not decide whether this lack of consistency between the analyses is due to overfitting, or because relevant methodological deviations from the original study were not reported in the papers, and thus function as hidden moderators. On the other hand, there are interesting dependencies on the particular probes. For example, for the population of East Asians, the tendency to give descriptivist answers is almost exclusively driven by Gödel-type probes (meta-analytic estimate of 64% as opposed to 50% overall). This is supported by the null result of the exploratory analysis of all non-Gödel probes. These results show a high sensitivity to the particular instrument for eliciting semantic intuitions, without presence of a theoretically convincing explanation (the effect is absent for Westerners). More generally, the large amount of heterogeneity and lack of uniform explanation may (partly) be due to methodological deficiencies (Appendix A.3). We therefore suggest that future research on cross-cultural differences in semantic intuitions extends the range of probes in experiments, in order to avoid that substantial philosophical and psychological conclusions depend (too) much on the specifics of a particular probe (see also Devitt & Porot, 2018).

Given these qualifications, the take-home message of our meta-analysis can be summed up as follows: there is cross-cultural variation in intuitions about the reference of proper names, but there is a high unexplained inter-study variance, compromising predictive validity; and it is not easy to disentangle this cross-cultural effect from the (random) effect of a particular study.

In relation to existing findings from other disciplines, our results are consistent with anthropologists' and linguists' work showing that there are many features of naming practices that differ across cultures and contexts of discourse (vom Bruck & Bodenhorn, 2006). However, underlying these differences, the primary pragmatic functions of person names of referring directly to a unique individual and also marking social connections, may be stable (Alford, 1988; Sophia & Marmaridou, 1989). Also stable may be the kinds of cognitive processes involved in understanding proper names, which probably recruit a number of structured representations (or frames) associated with the name (Dancygier, 2009; Valentine et al., 1996). Given the high degree of variation, both within and across experimental groups, that our meta-analysis uncovered, it is plausible that the workings of these kinds of processes are modulated by both causal and descriptive factors in different linguistic and social contexts (Genone & Lombrozo, 2012).

Due to the dependency on particular probes, we recommend that researchers interested in understanding the mechanisms underlying the processing of person names move beyond the contrast between descriptivists and causal-historical theories, and use a broader and more diverse set of vignettes in different languages and contexts. To this end, methodologies similar to those employed in cognitive psychology and neuroscience can be useful as they are designed to uncover the mechanisms and cognitive representations underlying people's semantic intuitions. For readers who are primarily interested in questions about philosophical methodology, our results confirm that philosophical scenarios are less reliable instruments for eliciting semantic intuitions than current philosophical practice seems to presume.

3

NHST: A SYSTEMATIC REVIEW

Abstract:
For decades, waxing and waning, there has been an ongoing debate on the values and problems of the ubiquitously used null hypothesis significance test (NHST). With the start of the replication crisis, this debate has flared-up once again, especially in the psychology and psychological methods literature. Arguing for or against the NHST method usually takes place in essay and opinion pieces that cover some, but not all the qualities and problems of the method. The NHST literature landscape is vast, a clear overview is lacking, and participants in the debate seem to be talking past one another. To contribute to a resolution, we conducted a systematic review on essay literature concerning NHST published in psychology and psychological methods journals between 2011 and 2018. We extracted all arguments in defense of (20) and against (70) NHST, and we extracted the solutions (33) that were proposed to remedy (some of) the perceived problems of NHST. Unfiltered, these 123 items form a landscape that is prohibitively difficult to keep in one's sights. Our contribution to the resolution of the NHST debate is twofold. 1) We performed a thematic synthesis of the arguments and solutions that carves the landscape in a framework of three zones: mild, moderate, and critical. This reduction summarizes groups of arguments and solutions, thus offering a manageable overview of NHST's qualities, problems, and solutions. 2) We provide the data on the arguments and solutions as a resource for those who will carry-on the debate and/or study the use of NHST.

This chapter is adapted from van Dongen, N.N.N., van Grootel, L. (under review).Overview on the null hypothesis significance test: A systematic review on essay literature on its problems and solutions. *Psychological Methods.*

3.1. Introduction

In 2015, the social sciences, psychology in particular, were dealt a devastating blow. The Open Science Collaboration (2015) replicated 100 landmark studies in the field of psychology, from which less than half yielded sufficiently similar result. Replication attempts in other academic fields showed comparable results (Begley & Ellis, 2012; Camerer et al., 2016, 2018; Prinz et al., 2011). In addition, Simmons, Nelson, and Simonsohn (2011) had shown how surprisingly easy it can be to increase the probability of obtaining positive results for a(n obviously) false hypothesis, using research practices that were considered conventional. According to many, a major culprit of this state of crisis was the Null Hypothesis Significance Test procedure for statistical inference, or at least the misuse of the method (e.g., Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011).

The Null Hypothesis Significance Test (NHST) is the procedure that is ubiquitously used in the social sciences for the analysis of quantitative data. Roughly summarizing NHST, the data that are produced by a study are compared to a null hypothesis of no-effect (e.g., no difference between two groups on the investigated measure). The null hypothesis is rejected if the observed data are sufficiently improbable according to this null hypothesis. Specifically, the null hypothesis is rejected if, under the assumption that the null hypothesis is true, the probability of observing the obtained data or data more extremely deviating from the null hypothesis is below a certain threshold (i.e., 5% probability; $p < .05$). In such a case, the null hypothesis is generally rejected in favor of an unspecified alternative hypothesis (e.g., there is a difference between the two groups). Otherwise, the null hypothesis is retained, though one is not supposed to consider this as evidence in its favor.

This seemingly simple procedure is associated with two major problems: 1) is easily misused and 2) it possesses (hidden) deficiencies. There is abundant empirical evidence (e.g., see Engman, 2013) of scientists unwittingly misusing the method and misinterpreting its results (e.g., interpreting the probability of the data under $\mathcal{H}_0$ as the probability of $\mathcal{H}_0$ given the data). With regard to the second problem, deficiencies and flaws of NHST that are independent of misuse have been brought into the limelight (e.g., ease of rejecting $\mathcal{H}_0$ with a large enough sample; Wagenmakers, 2007). These issues, in combination with publication bias (Sutton, 2009) and a publish-or-perish incentive structure are expected to be at the root of the low replicability rate of social science research (e.g., Szucs & Ioannidis, 2017).

As a response to the first mentioned problem, NHST's misuse, several proposals have been published. People have called for and are starting to implement more

transparency (e.g., open data and methods; Freese & King, 2018; Munafò, 2016) and design and analysis specification prior to data collection (i.e., preregistration; Lindsay, Simons, & Lilienfeld, 2016; van 't Veer & Giner–Sorolla, 2016). For instance, registered reports have, at the moment of writing, been adopted by more than 250 journals (Center for Open Science, 2020). Roughly speaking, a registered report is a study design and analysis plan that is submitted for peer review at a journal (that accepts registered reports) prior to the data collection and accepted on its methodological merits and general theoretical relevance alone. Early meta-scientific results on the effectiveness of registered reports are promising (Scheel, Schijen, & Lakens, 2020).

In response to the second mentioned problem, NHST's deficiencies, researchers and statisticians have called for reform in statistical methodology. Some have proposed that, if we were to retain NHST, we should at least lower the threshold for statistical significance (Benjamin et al., 2018), which would reduce the problem of inflated false-positive rates to some extent. This view is opposed by those that think that there should not be a default threshold and that the critical value should be set and justified by the study design, the research question, etc. (Lakens, Adolfi, et al., 2018). Then there are others who want to abandon the statistical significance threshold all together (McShane et al., 2019). In this case, the advice is to refrain from hypothesis tests and use, for instance, only estimation methods instead (e.g., Cumming, 2014). This perspective is, in turn, opposed by those who see the merits of hypothesis testing and binary decision-making in science (e.g., Morey et al., 2014). Instead on estimation, they propose a Bayesian version of hypothesis testing that is supposed to be less prone to error and misuse (e.g., Wagenmakers et al., 2017). All these groups agree that the current practice involving NHST is detrimental to science. However, there is a worrisome absence of opinions and suggestions of these groups converging on a semblance of consensus about how these problems should be addressed.

The literature on the problems of the significance test goes back to the 1940s (Berkson, 1942) and a steadily increasing stream has been published on the topic. The current attempt at methodological reformation is also not the first time that scientists have tried to change the NHST system. For instance, around the 1990s there was a serious collective attempt to change the way in which scientists analyze data and report results (e.g., J. Cohen, 1994; Gigerenzer, 2004; Meehl, 1990b, 1992). Meehl (1990b, p. 139) put it succinctly when he wrote:

> Let me say as loudly and clearly as possible that what we critics of weak signifi-

> cance testing are advocating is not some sort of minor statistical refinement (e.g., one-tailed or two-tailed test? unbiased or maximum likelihood statistics? pooling higher order uninterpretable and marginal interactions into the residual?). It is not a reform of significance testing as currently practiced in soft psychology. We are making a more heretical point than any of these: We are attacking the whole tradition of null-hypothesis refutation as a way of appraising theories. The argument is intended to be revolutionary, not reformist.

It seems that all these events did however not inspire the intended methodological revolutions and that they have not left many traces in academia's collective memory. This could be because papers about (how to solve) NHST's problems are focused only a small number of issues and are contradicted in other papers by authors who disagree, instead of forming a coherent structure that provides a clear plan of attack. For instance, J. Cohen (1994) claimed that we, scientists, are interested in the probability that the hypothesis is true and should thus use Bayesian statistics. This was then contradicted in responses to Cohen's paper (e.g., Baril & Cannon, 1995; Frick, 1995; McGraw, 1995; Parker, 1995). In turn, Cohen tempered his claim to some extent in a rejoinder (J. Cohen, 1995).

As it appears, many authors agree on the fact that there is problem to be addressed concerning NHST, but most differ in the deficiencies and misuses of the method they address. The papers seem to talk past each other, each vying for attention, and in doing so, inhibiting improvement of the practice of statistical inference. A clear perspective on NHST is absent and without such a perspective —a map of NHST's qualities, problems, its solutions— it will be neigh impossible to plot the best course of action away from the replication crisis and towards methodological health. To this end, a framework is required in which the different perspectives on NHST can be placed side-by-side, compared, and evaluated. Such a framework will allow us to assess the veracity of each argument for or against NHST and assess the potential consequences of proposed methodological reform prior to implementation. With the systematic review reported in this paper, we aim to provide the foundations of such framework of NHST and its practice.[1]

3.2. Methods

We reviewed opinion literature describing the misconceptions and limitations of the Null Hypothesis Significance Test (NHST). We synthesized essays and opinion pieces, because we are interested in the arguments researchers use to make their

[1] Our data, notes, and research materials are available at `https://osf.io/ayf56/`.

Table 3.1: ENTREQ items adhered to and where they can be found in the text.

No.	Item	Location
1	Aim	Section 3.2.1
2	Synthesis methodology	Section 3.2.4
3	Approach to searching	Section 3.2.2
4	Inclusion criteria	Section 3.2.2, Appendix B.1
5	Data sources	Section 3.2.2, Appendix B.1
6	Electronic search strategy	Section 3.2.2, Appendix B.1
7	Study screening methods	Section 3.2.3, Appendix B.1
8	Study characteristics	Section 3.3.1
9	Study selection results	Section 3.2.2
14	Data extraction	Section 3.2.3
15	Software	Section 3.2.4
16	Number of reviewers	Section 3.2.3, Section 3.2.4
17	Coding	Section 3.2.4
19	Derivation of themes	Section 3.2.4
20	Quotations	Section 3.3.3
21	Synthesis output	Section 3.3.3

point for or against NHST and the solutions they provide. To our knowledge, at the moment of writing, there are no specific guidelines for the synthesis of opinion literature. Since our review question is mainly exploratory and open, we made use of existing guidelines for reviews of qualitative and mixed methods studies. To ensure transparency, extension, and reproducibility, we adhere to the ENTREQ transparency statement for qualitative research (Tong, Flemming, McInnes, Oliver, & Craig, 2012). To aid this endeavor, Table 3.1. provides the list of applicable ENTREQ items and in which sections the information is given that pertains to that item. Five out of 21 ENTREQ items were not addressed. One item (ENTREQ: 18.) is concerned with study comparison, which is not applicable to our review. Four items (ENTREQ: 10. - 13.) are concerned with the quality appraisal of the data, which is absent from our review. We considered the checklist for text and opinion tools from the Joanna Briggs Institute (2017), but decided against using it. We preferred to focus on the quality of the individual arguments that make up the opinion that is to be evaluated, which is something this tool does not provide. Using the tool despite this deficiency could have led to the exclusion of papers based on their overall opinion, whereas articles with 'illogical' opinions could still contain arguments with interesting information for our analysis (e.g., C. Carroll, Booth, & Lloyd-Jones, 2012). The rest of this section contains the formulation of the review question, the search strategy, the selection of studies, the data extraction, and the data analysis plan.

3.2.1. Review question

We used the acronym SPIDER (Cooke, Smith, & Booth, 2012) to aid us in the focus on a review question. The SPIDER acronym allows the reviewer to specify the sample, phenomena of interest, design, evaluation, and research type. We specified the components of the SPIDER acronym as follows.

Sample: Opinion statements supported by arguments for or against the Null Hypothesis Significance Test method and proposed solutions to problems, which in turn are based on evidence in the form of literature references, examples, mathematical proofs, and/or data simulations.

Phenomena of Interest: *deficiencies* of NHST, *misconceptions* about NHST, *defenses* of NHST, *solutions* to problems pertaining to NHST, and overall *conclusions* about NHST. *NHST deficiencies:* Intrinsic properties of the method that are considered undesirable, limiting or pernicious (e.g., *p*-value is not an absolute measure of evidence, but relative to the sample size and effect-size of the population). *NHST misconceptions:* Misconceptions about the results of the Null Hypothesis Significance Test are specific misinterpretations of the results produced by the NHST method that are linked to (elements of) the method (e.g., interpreting a *p*-value as the probability of the null hypothesis on the observed data instead of the probability of the data on the assumption that the null hypothesis is true). *NHST defenses:* Properties of NHST that speak to its qualities or counterarguments to the misconceptions and deficiencies. *Solutions to NHST:* Solutions that are purported to mitigate or remedy the mentioned problem. Examples of such solutions are: a) alternative methods of statistical inference that supposed to suffer less from limitations and potential for misconception than the NHST procedure; b) safeguards or policy changes that mitigate or remedy the purported problems; c) changes to the NHST method mitigate or remedy the purported problems. *Conclusions:* The particular conclusions about, perspectives on, or positions to NHST that the writers have, which follow from / are built on the arguments they use in their paper.

Design: Written discussions / opinions on the phenomena of interest using arguments supported by evidence.

Evaluation: Identification and description of perspectives and opinions on the NHST.

Research type: Essays and opinion pieces (i.e., non-empirical scientific publications) in the psychological science and the accompanying methodological literature published between 2011 and 2018. This boils down to the following two general review questions:

What are the deficiencies, misconceptions, and defenses of NHST?
What are the proposed solutions to problems pertaining to NHST?

3.2.2. Search strategy

Although this is a novel type of systematic review, we tried to adhere as much as possible to reporting standards already in existence. We adopted the STARTLITE reporting standard for literature searches (Booth, 2006). We explicate the items that make up the STARLITE mnemonic:

Sampling strategy: We used purposive sampling. Specifically, we restricted our search to papers published in psychology or psychological methods literature published between 2011 and 2018. We choose this restriction because 2011 is recognized as the year the replication crisis started and psychology is the field in which it takes place or, at least, is most engaged with investigating and remedying the situation.

Type of study: As indicated in Section 3.2.1, we limited this review to essays and opinion pieces. We choose these types of papers because this is where the pros, cons, and solutions of NHST are presented.

Approaches: We used electronic subject searches. No other approaches were used, because we expected that all published papers on the subject are reachable through electronic searches on online databases.

Range of years: We included papers published from 1 January 2011 up to 31 December 2018. We selected this starting date because Bem's study that triggered the replication crisis was published in January 2011 and had received some media coverage at the end of 2010 (e.g., Lehrer, 2010). We performed the search in January 2019.

Limits: The functional limits that were applied concern English language and publication in peer reviewed scientific journals. The literature was limited to peer reviewed journal publications because an adequate quality assessment instrument is lacking for essay literature and peer review offers some quality control.

Inclusion and exclusion: Papers were included if they were about the Null Hypothesis Significance Test inference procedure (i.e., the correct *Type of study*); either wholesale or parts of the procedure. Specifically, papers were included if they were about the concepts '*p*-value', 'statistical significance', 'significance tests', or 'null hypothesis'. Papers were excluded if they were about empirical research instead of the opinion or arguments of the author(s); if the concepts were used in the paper, but the of the papers were not about these concepts (e.g., "the results were statistically significant at the $p < .05$ level"); or if the papers were about another statistical proce-

dure that hare some of the concepts (e.g., the null hypothesis in Bayesian hypothesis test). Figure 3.1 depicts more specific information about the reasons for exclusion of papers.

Terms used: We used the following search terms: 'null hypothesis', 'hypothesis test', 'statistical significance', 'significance test', 'p-value', and 'p value.' We used these terms, because these are the concepts most commonly associated with NHST. All search terms with syntax and operators are provided in Appendix B.1. Electronic sources: We used two databases for the literature search, PsycInfo (via EBSCOhost) and Web of Science Core (via EndNote 8). We limited our search to these two databases, because they cover the width of psychological literature. The schematic representation of the selection process is presented in Figure 3.1. The initial search yielded 6221 potential paper and the final selection consisted of 42 papers.

3.2.3. Data extraction

We extracted descriptive information on the articles and descriptive information concerning the authors' opinion on NHST. The descriptive information on the article consisted of the name of the (first) author, year of publication and journal, and whether or not the article was part of a special issue. Furthermore, we extracted the *purpose* of the article if explicitly stated by the author(s) in the abstract or introduction; the *readability* for non-expert (i.e., absence of advanced mathematics and assumed technical know-how in the argumentation); the *applicability of solutions* (i.e., solutions do not require costly statistical software or mathematical experience); and the *conclusion* if explicitly stated at in the last section of the article.

Descriptive information concerning the authors' opinion on NHST consists of arguments and solutions related to NHST was extracted from all parts of the papers. Arguments and solutions were identified in terms of explicit claims pertaining to the phenomena of interest (as specified in Section 3.2.1). Specifically, a description in one or more sentences on their position (or the views/position of others they endorse) with respect to, for instance a particular deficiency of NHST. An example of data extraction of and argument from one of the papers is as follows. On page 1, Trafimow (2013) states: "p-values generally are inaccurate estimators of probabilities of null hypotheses." This claim indicates NHST's deficiency that its outcome measure, the *p*-value, does not provide an accurate estimation of the probability of the null hypothesis. It was therefore interpreted as a deficiency of NHST.

Evidence used to support these claims was also extracted. Empirical research, examples, other literature, simulations, and/or mathematical proofs were extracted

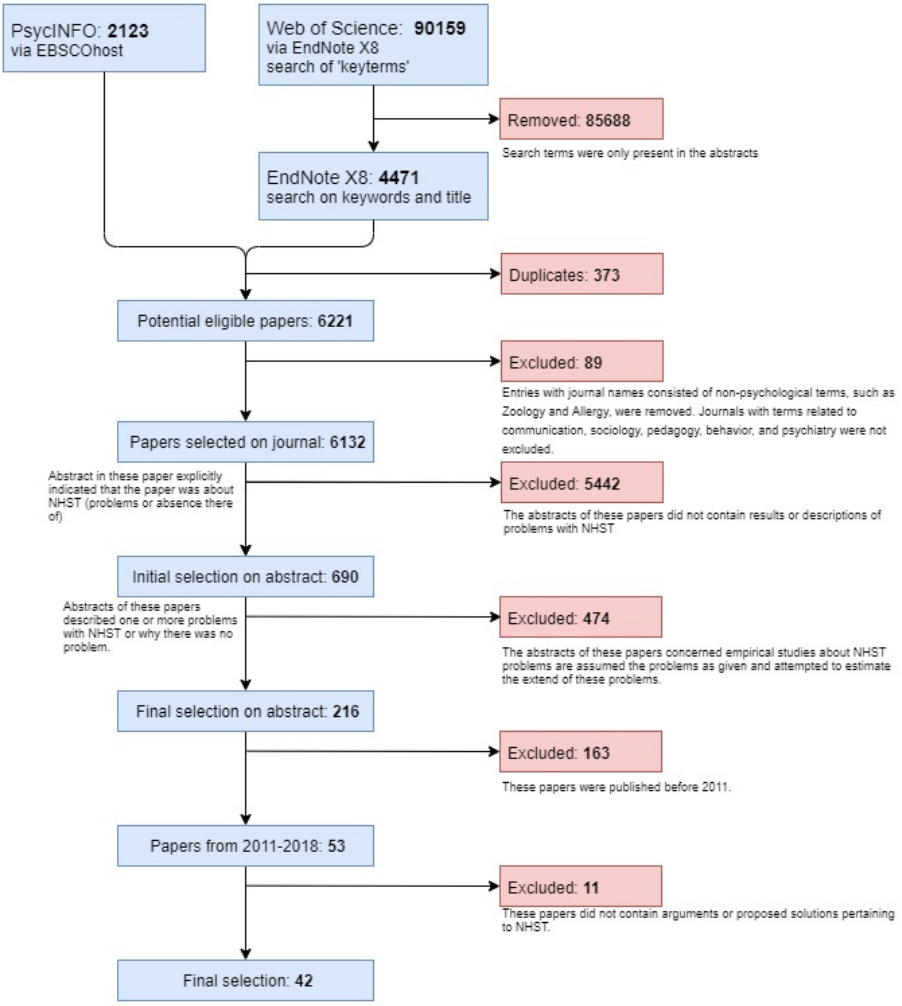


Figure 3.1: Selection process of essay and opinion pieces pertaining to the qualities and problem of NHST.

as evidence if they were referenced for that claim. For example, on page 13, Cumming (2014) makes the claim: "[w]hatever the N, a [p-value] gives only extremely vague information about replication (Cumming, 2008)." In this case, "(Cumming, 2018)" is used as support for this claim. We interpreted this information as evidence of the type reference to other literature. The extraction form, provided as online supplementary material (`https://osf.io/8cjm7/`), gives further details concerning what was extracted the indicators that were for their classification.

3.2.4. Data analysis

The purpose of our systematic review is to provide an overview of the qualities, deficiencies, and misconceptions of, and solutions to NHST. The collection of extracted arguments, arguments, proposed solutions, conclusions, and their supporting evidence provide this overview, but it is too vast interpret without some kind of focus or filter. To this end, we conducted and initial qualitative thematic synthesis (Thomas & Harden, 2008) in which we identified clusters of arguments and solutions described in the articles.

Thematic synthesis

The thematic synthesis contains four steps: identification of claims in the papers, development of *descriptive themes*, the development of analytical themes within the phenomena of interest, development of overarching analytical clusters. These four steps were as follows. Firstly, we identified pertinent text fragments (i.e., the claims pertaining to arguments, solutions, or conclusions) from all sections of the articles.

Secondly, we developed descriptive themes (Thomas & Harden, 2008) by clustering fragments into groups using the software program EPPI-reviewer 4.11.4.0 (Thomas, Graziosi, Brunton, O'Driscoll, & Bond, 2020). Fragments were grouped in descriptive themes based on interpreted overlap in content, meaning, or intention.[2] These descriptive themes were the individual arguments concerning NHST's deficiencies, misconceptions, and defenses, solutions, and conclusions.

Thirdly, from the descriptive themes, analytical themes were developed (Thomas & Harden, 2008). Analytical themes go beyond the information from the primary studies in order to address our review question directly. During this step, the analysis was done separately for each phenomena of interest as specified in the SPIDER research question: Deficiencies, misconceptions, NHST defense arguments, solutions, and conclusions (Section 3.2.1. The descriptive themes were first grouped in analytical subthemes, based on similarities in their semantic content. These analytical subthemes were then grouped in overarching analytical themes within each phenomenon of interest (see Figure 3.2. for a graphical representation of this process).

Fourthly, overarching analytical clusters[3] were developed from the analytical themes identified within the phenomena of interest. These overarching clusters form

[2]Note that this is a slight deviation from Thomas and Harden (2008). We adopted this deviation to make the method amendable to the unorthodox nature of our systematic review.

[3]We adopted this term to add an additional layer in the hierarchy that provide some conceptual distinction.

semantically coherent zones within the landscape of NHST, containing its qualities, problems and the solution to these problems.

Syntactically coherent zones in in the NHST landscape could be a counterpart to these semantic zones that are identified by the overarching analytical clusters. For instance, how arguments, solutions, conclusions, and their supporting pieces of evidence are connected could be represented by a network, which allows analysis of how they are connected within and between the papers. Unfortunately, a proper treatment of such a network analysis is beyond the scope of this paper. A first stab at this is presented in Appendix B.2 and the code and data are available at `https://osf.io/ayf56/files/`.

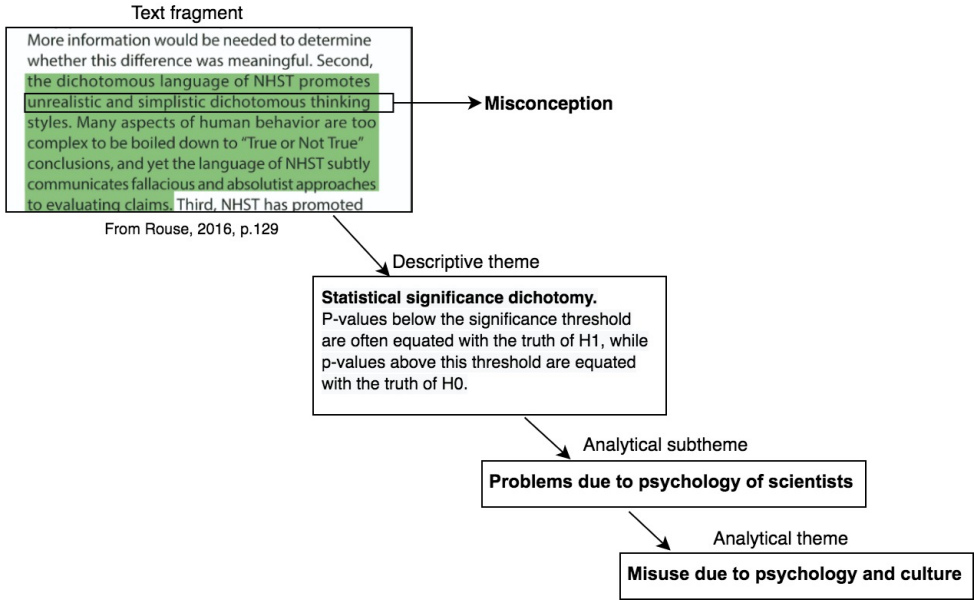


Figure 3.2: Example of the analysis hierarchy within a phenomenon of interest. A text fragment is first labeled with a descriptive theme. These descriptive themes are then clustered in analytical subthemes based on their similarity in semantic content. These were then clustered into higher order analytical themes within each phenomenon of interest.

3.3. Results

3.3.1. Descriptive results

A total of 42 articles were included in the analysis. Of these, 34 were regular articles, two editorials and six commentaries. Eleven articles were part of a special issue; of which eight were published in "Perspectives on the Use of Null Hypothesis Statistical Testing (NHST)" a three-part special issue edited by Educational and Psychological

Measurement. Many papers (12 out of 42) were published in 2017 and only a few papers were published in 2012, 2014, and 2018 (respectively, 1, 3, and 2 out of 42). All but one of the papers were considered readable. The exception was a paper published in *Statistica Sinica* (Martin & Liu, 2014) and required the understanding of mathematical notation, the stated mathematical theorems and their proofs. Out of 42 articles, 29 offered one or more potential solutions to the problems pertaining to NHST. In only three out of these 29 cases, no relevant information was given on how to implement the suggested solution(s). A complete overview of the paper characteristics can be found is presented in Table 3.2.

Table 3.2: Overview of the papers and their characteristics. The details on the special issues and the subjects of the commentaries can be found in the online data set. Readability and usefulness of solutions are not shown, because it was 'Yes' in all cases but one. 'S.issue' stands for 'special issue.'

First author	Year	Journal	Context	S.issue	Solutions
Gelman	2013	Journal of Mathematical Psychology	Commentary	Yes	No
Rouse	2016	Psi Chi Journal of Psychological Research	Editorial	No	Yes
Gelman	2018	Personality and Social Psychology Bulletin	Article	No	Yes
Dienes	2018	Psychonomic Bulletin and Review	Article	Yes	Yes
Campitelli	2017	Educational and Psychological Measurement	Article	Yes	Yes
Haig	2017	Educational and Psychological Measurement	Article	Yes	Yes
Haggstrom	2017	Educational and Psychological Measurement	Article	Yes	No
Schneider	2015	Scientometrics	Article	No	Yes
Wilcox	2017	Educational and Psychological Measurement	Article	Yes	Yes
Trafimow	2016	Basic and Applied Social Psychology	Editorial	No	No
Chen	2017	NeuroImage	Article	No	Yes

Continued on next page

Table 3.2: Continued from previous page

First author	Year	Journal	Context	S.issue	Solutions
Trafimow	2013	Frontiers in Psychology	Commentary	No	No
Marsman	2017	Educational and Psychological	Article	Yes	No
Lu	2015	Shanghai Archives of Psychiatry	Article	No	Yes
Hupé	2015	Frontiers in Neuroscience	Article	No	Yes
Laber	2017	Journal of the American Statistical Association	Commentary	No	No
Rao	2016	Journal of Modern Applied Statistical Methods	Article	No	Yes
Goddard	2015	METODE Science Studies Journal	Article	No	Yes
Cumming	2014	Psychological Science	Article	No	Yes
Cumming	2013	Australian Psychologist	Article	No	Yes
Klugkist	2011	International Journal of Behavioral Development	Article	No	Yes
Garcia-Perez	2017	Educational and Psychological Measurement	Article	Yes	Yes
van Helden	2016	Nature Methods	Commentary	No	No
Perezgonzalez	2014	Theory & Psychology	Article	No	Yes
Citrone	2011	Bulletin of Clinical Psychopharmacology	Article	No	Yes
Konijn	2015	Communication Methods and Measures	Article	Yes	Yes
Hasley	2015	Nature Methods	Article	No	Yes
Chang	2017	Educational and Psychological Measurement	Article	Yes	No
LeBel	2011	Review of General Psychology	Article	No	Yes
Miller	2017	Educational and Psychological Measurement	Article	Yes	Yes
Lambdin	2012	Theory & Psychology	Article	No	No
Hofmann	2011	Psychotherapy	Commentary	No	No
Perezgonzalez	2015	Frontiers in Psychology	Article	No	Yes

Continued on next page

Table 3.2: Continued from previous page

First author	Year	Journal	Context	S.issue	Solutions
Szucs	2017	Frontiers in Human Neuroscience	Article	No	Yes
Savalei	2015	Frontiers in Psychology	Article	No	Yes
Bradley	2016	Psychological Reports	Article	No	No
Baird	2016	Journal of Modern Applied Statistical Methods	Article	No	Yes
Garamszegi	2017	Trends in Neuroscience and Education	Commentary	No	Yes
Martin	2014	Statistica Sinica	Article	No	Yes
Hubbard	2011	Journal of Applied Statistics	Article	No	No
Johanson	2011	Scandinavian Journal of Psychology	Article	No	Yes
Engman	2013	Quality & Quantity: International Journal of Methodology	Article	No	No

3.3.2. Arguments, solutions, conclusions, and evidence

We identified a total of 90 arguments, 33 proposed solutions and 14 conclusions across the 42 articles that were included. Concerning the arguments, we identified 39 deficiencies of NHST, 31 arguments about how NHST is misinterpreted and misused, and 20 on the defense of NHST. These 137 items provide the unfiltered answers to the two review questions: they describe the NHST landscape. A rough overview of the distribution of arguments and solutions across the papers is presented in Table 3.3. The complete lists of these arguments, solutions, and conclusions, their descriptions, and from which papers they originate can be found in the online supplementary materials (`https://osf.io/86tnz/`). As is shown in Table 3.3, the range in the number of items per paper is large (i.e., from 1 to 30 per paper). One paper contained as many as 11 deficiencies. Another paper contained as many as 14 misconceptions. The maximum number of NHST defenses was 8. The highest number of solutions was 11. Not a single paper had the highest count on more than a single category.

Table 3.3: Overview of the number of arguments and solutions per paper. See Appendix A for details on which argument or solution is present in which paper and in how many papers each item is present.

Author	Year	Deficiency	Misconception	Defense	Solution	Total
Gelman	2013	3	0	0	0	3
Rouse	2016	0	3	4	0	7
Gelman	2018	1	2	0	5	8
Dienes	2018	3	0	0	1	4
Campitelli	2017	0	0	0	1	1
Haig	2017	1	3	0	2	6
Haggstrom	2017	0	8	0	0	8
Schneider	2015	11	8	0	10	29
Wilcox	2017	0	0	0	1	1
Trafimow	2016	0	3	0	0	3
Chen	2017	2	3	2	1	8
Trafimow	2013	4	0	0	0	4
Marsman	2017	2	0	0	0	2
Lu	2015	1	1	0	1	3
Hupé	2015	6	1	0	3	10
Laber	2017	0	0	1	0	1
Rao	2016	2	0	0	1	3
Goddard	2015	1	0	0	1	2
Cumming	2014	2	0	0	3	5
Cumming	2013	2	0	0	2	4
Klugkist	2011	2	1	0	1	4
Garcia-Perez	2017	0	0	8	0	8
van Helden	2016	0	0	1	0	1
Perezgonzalez	2014	2	0	0	1	3
Citrone	2011	1	0	0	1	2
Konijn	2015	2	3	0	1	6
Hasley	2015	4	0	0	1	5
Chang	2017	1	0	3	0	4
LeBel	2011	5	0	0	4	9
Miller	2017	1	0	4	1	6
Lambdin	2012	5	14	0	1	20
Hofmann	2011	2	0	0	0	2
Perezgonzalez	2015	0	3	0	2	5
Szucs	2017	7	12	0	11	30
Savalei	2015	0	0	3	3	6

Continued on next page

Table 3.3: Continued from previous page

Author	Year	Deficiency	Misconception	Defense	Solution	Total
Bradley	2016	3	1	0	1	5
Baird	2016	4	1	0	4	9
Garamszegi	2017	7	2	0	1	10
Martin	2014	1	1	0	1	3
Hubbard	2011	1	1	0	0	2
Johanson	2011	6	0	0	1	7
Engman	2013	3	5	0	0	8

In total, we identified 508 unique pieces of evidence that were referenced in the 42 papers to support the 123 arguments and solutions. 15 were published before 1950; 139 were published between 1950 and 1999; 181 were published between 2000 and 2010; and 173 were published between 2011 and 2017 (see Figure 3.3 for the distribution of publication dates). The body of evidence is diverse. For instance, some references that were used as evidence for the arguments in the papers came from journals such as *Journal of Wildlife Management*, *The Journal of Experimental Education*, *Philosophy of science*, *Political Analysis*, *Biological Psychiatry*, *Seminars in Hematology*, *Ecology*, and *Biology of Blood and Marrow Transplantation* (See the online supplements for further details on the extracted evidence, `https://osf.io/bktfx/`).

3.3.3. Thematic synthesis results

The opinions on NHST as presented in the literature were analyzed using thematic synthesis. We generated three overarching analytical clusters, which can be identified as three zones in the (problems and solutions concerning) NHST landscape: 1) the mild problems zone 2) the moderate problems zone and 3) the critical problems zone. For all three zones, we have provided the identifying element(s) and we specify the associated analytical themes of arguments, solutions and conclusions, and provide quotes from the papers to support our synthesis. The terms in italic represent the analytical (sub)themes generated during the analysis (Section 3.2.4). See Figures 3.4, 3.5, and 3.6 for an overview of the mild, moderate, and critical zone of the NHST landscape respectively. The online available data (`https://osf.io/bktfx/`) provides a complete breakdown of the zones from their analytic themes and subthemes to descriptive themes (i.e., arguments, solutions, conclusions).

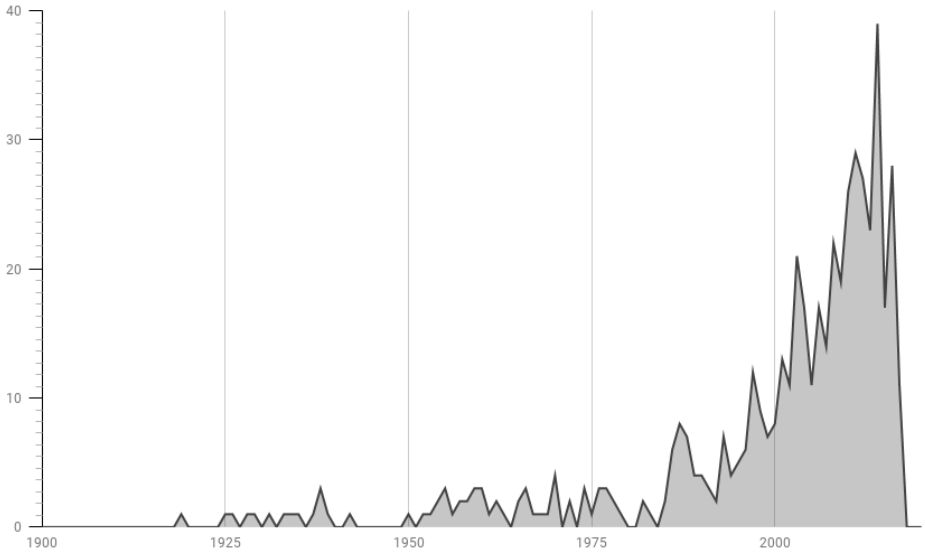


Figure 3.3: Distribution of publications dates of the pieces of evidence extracted from the papers. These pieces of evidence where referenced in the papers to support the particular claims that were extracted as the descriptive themes.

The mild problems zone

The mild problem zone of NHST (see Figure 3.4) can be labelled as: NHST itself is not the problem. The problems with statistical inference, such as inflated false-positive rates and biased results, have to do with misconceptions and faulty practice due to the psychology of the researchers and the research culture, not with the method that is being used. Specifically, these problems result from *psychological heuristics that lead to misunderstanding*, *the tendency to dichotomous reasoning*, and *bad research practices that are considered acceptable by the community*. This means that the problems have not so much to do with the method per se, but with peripheral aspects, like the language of NHST in relation human reasoning and behavior. For instance, Rouse (2016, p.129) writes

> [...] the dichotomous language of NHST promotes unrealistic and simplistic dichotomous thinking styles. Many aspects of human behavior are too complex to be boiled down to "True or Not True" conclusions, and yet the language of NHST subtly communicates fallacious and absolutist approaches to evaluating claims.

Within this zone, solutions are geared towards leaving the method as it is, but

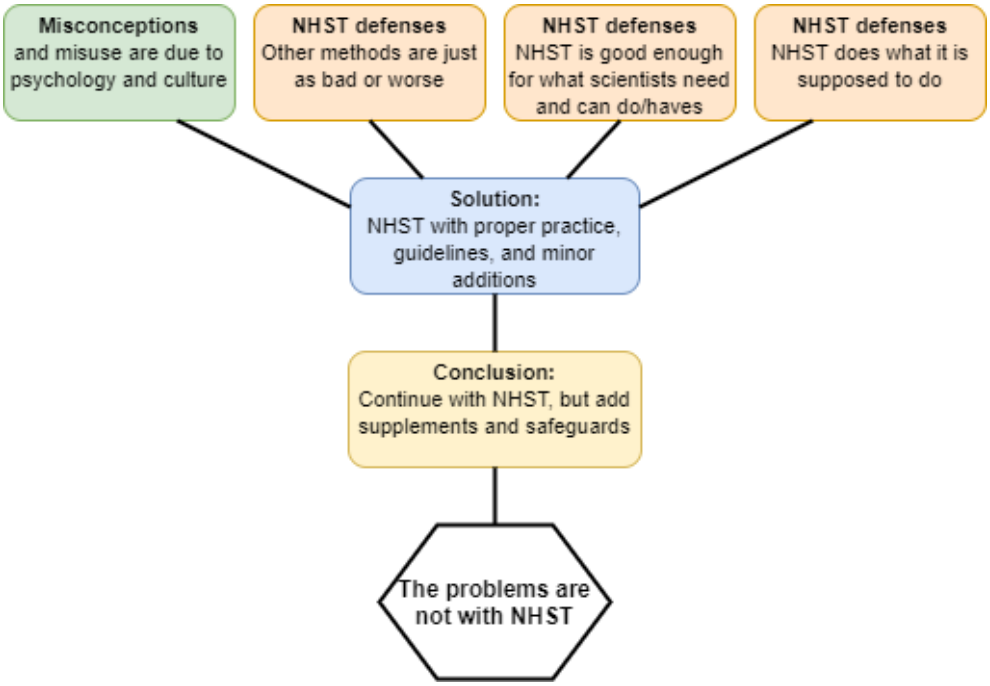


Figure 3.4: The mild zone in the NHST landscape: NHST itself is not the problem. This zone consists of three NHST defenses themes (orange), one misconceptions theme (green); one solution theme (blue), and one conclusion theme (yellow). 14 of the 42 papers ascribe to a conclusion that is part of this zone. There are an additional four papers that at least contain one misconception and at least one solution that are part of this zone. There are seven papers that only have one or more misconceptions, one or more deficiencies, or one or solutions that are part of this zone while also containing descriptive themes belonging to the other phenomena of interest whose analytic theme(s) are part of another zone (these papers are: Bradley, 2016; Engman, 2013; Gelman, 2013; Konijn, 2015; Trafimow, 2016; Cumming, 2013; Cumming, 2014).

adding *supplements and safeguards to NHST*, that mitigate the psychological and cultural problems. These solutions consist of policy changes (e.g., in education, funding, publication), focus on proper use of NHST (e.g., high enough power, adherence to strict guidelines), and additions to NHST, such as procedure verification, meta-analyses, and direct replication. For instance, Rouse (2016, p.132) writes:

> As the research community is examining and debating the most appropriate use of inferential statistics, we recommend that researchers continue to use the tools of NHST, but to use these tools carefully and effectively.

This zone also contains the three analytical themes concerning defenses of NHST and its use. These themes are 1) *other methods are just as bad or worse:* Bayesian hypothesis tests and estimations methods either come with additional and more

pernicious problems (for the former) or are based on the same principles and thus suffer from the same problems (for the latter). 2) *NHST does what it is supposed to do:* the shortcomings, such as reliance on sampling plan, are actually qualities and what NHST is supposed to do, control error rates, it does well. 3) *NHST is good enough for what scientists need and can do/have*: NHST is actually quite intuitive to understand or at least not more misunderstood than other methods and can be unproblematically operated when certain precautions are taken (e.g., sufficient power and replication studies). For instance, Miller (2017, p.664) writes:

> To the contrary, the heart of NHST is a simple, intuitive, and familiar "common sense" logic that most people routinely use when they are trying to decide whether something they observe might have happened by coincidence (a.k.a., "randomly," "by accident," or "by chance").

This zone can be seen containing the milder conclusions concerning NHST, such as the analytical subthemes *use NHST with care and proper practice, use complete reporting,* and *improve general practice*. In other words, from within this zone, there is nothing actually wrong with NHST. Specifically, the method works as it is supposed to work when used correctly. Policy and education changes are all that is required to remedy the existing problems.

The moderate zone

The second zone (see Figure 3.5) within the NHST landscape contains the proposals for moderate, but not fundamental, changes in light of serious misuse of NHST and mild shortcomings of the methods. Within the borders of this zone, the general conclusion is that researchers should *stick with frequentist hypothesis testing*, just in a different form than NHST. This differs from the mild zone in the sense that the NHST approach of sole reliance on *p*-values to accept unspecified alternative hypotheses should be dropped in favor of a procedure that is still based on the frequentist philosophy of probability, but has a different approach to hypothesis testing (e.g., interval-null hypothesis testing). Scientists misunderstand how NHST works which leads to wrongfully interpreted results. This misunderstanding is caused by deficiencies concerning the input for the analysis and the output. To illustrate, within this zone, the use of NHST is problematic because, for instance, if the sample size is large enough, the null hypothesis will always be rejected as a result (e.g., "Berkson (1938) was the first to notice dependence of significance testing on the sample size. He objected that it is possible to obtain a statistically significant

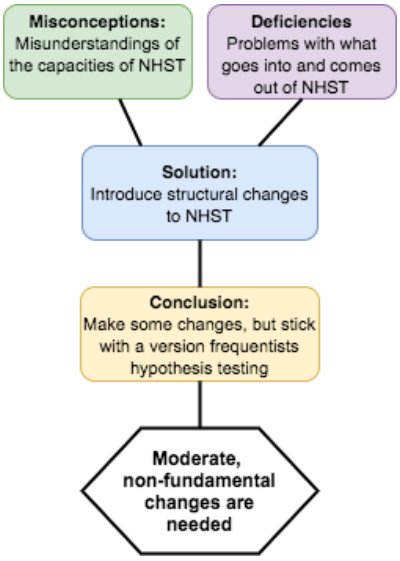


Figure 3.5: The moderate zone: Moderate, non-fundamental changes are needed. This zone consists of one deficiency theme (purple), one misconceptions theme (green); one solution theme (blue), and one conclusion theme (yellow). 9 of the 42 papers ascribe to a conclusion that is part of this zone. There are an additional six papers that contain at least one misconception, deficiency or one solution that is part of this zone and also on item that is part of another theme in this zone. There are ten papers that only have one or more misconceptions, one or more deficiencies, or one or solutions that are part of this zone while also containing descriptive themes belonging to the other phenomena of interest whose analytic theme(s) are part of another zone (these papers are: Häggström, 2017; Chen, 2017; Garamszegi, 2017; Gelman, 2013; Hasley, 2015; Hupé, 2015; Johanson, 2011; LeBel, 2011; Perezgonzalez, 2014; Trafimow, 2013).

chi-square test merely by increasing sample size"; Rao & Lovric, 2016, p.13) and, as LeBel (2011, p.374) states, it outputs ambiguous evidence:

> Thus, although it is well known that negative (null) results are ambiguous and difficult to interpret, exclusive reliance on NHST makes positive results equally ambiguous, because they can be explained by flaws in the way NHST is implemented rather than by a more theoretically interesting mechanism (Meehl, 1967).

Although these are important criticisms on NHST as it is currently practiced, not all frequentist hypothesis of statistical significance testing need to possess these deficiencies. As another example, the misconceptions, such as that statistical significance comes in degrees (e.g., '$0.05 < p < 0.1$' marginally significant, '$p < 0.001$' highly significant and that high p-values indicate that the null hypothesis is true, can be addressed by reconceptualizing several components of NHST (i.e., what error

probabilities are); as Haig (2017, p.496) writes about adopting Error statistics theory of Mayo:

> In her initial formulation of the error-statistical philosophy, Mayo (1999) modified, and built upon, the classical Neyman–Pearsonian approach to ToSS [tests of statistical significance]. However, in later publications with Spanos (e.g., Mayo & Spanos, 2011), and in writings with David Cox (Cox & Mayo, 2010; Mayo & Cox, 2010), her error-statistical approach has come to represent a coherent blend of many elements, including both Neyman– Pearsonian and Fisherian thinking. For Fisher, reasoning about p values is based on postdata, or after-trial, consideration of probabilities, whereas Neyman and Pearson's Type I and Type II errors are based on predata, or before-trial, error probabilities. The error-statistical approach assigns each a proper role that serves as an important complement to the other (Mayo & Spanos, 2011; Spanos, 2010). Thus, the error- statistical approach partially resurrects and combines, in a coherent way, elements of two perspectives that have been widely considered to be incompatible. In the post-data element of this union, reasoning takes the form of severe testing, a notion to which I now turn.

From which Haig (2017, p.504) concludes:

> In more than 50 years of preoccupation with these tests, psychology has concentrated its gaze on teaching, using, and criticizing NHST in its muddled hybrid form. It is high time for the discipline to bring itself up-to-date with best thinking on the topic, and employ sound versions of ToSS in its research.

In addition, deficiencies, such as that the null hypothesis is always false or that NHST overestimates the evidence against the null hypothesis, are remedied when minor structural changes are made. These changes are, for instance, using context dependent Type I and Type II errors (i.e., incorporating the costs of making a Type I and/or Type II error for the particular test at hand); interval null hypothesis testing (i.e., using a range of values around the null effect instead of a point null hypothesis); and using a minimum effect-size of interest null hypothesis (i.e., using as null hypothesis the smallest effect-size that would be of interest for the particular research question). As an example, Rao (2016, p.15) writes:

> So, what should we do? This article is an initial contribution to making a paradigm shift in testing statistical hypotheses. Instead of testing highly problematic and almost surely false point null hypotheses, as a natural replacement, test a negligible null hypothesis.

This collection of suggested solutions, and this zone in general, is succinctly represented by Baird (2016, p.113):

> By implementing the concepts from these sources, both traditional and contemporary, researchers are engaged in what could be described as "context-driven NHST" or CD-NHST. Instead of being driven by convention, which may or may not have much relevance, CD-NHST places the researcher in the driver's seat of inference.

Consequently, the deficiencies of NHST do require adjustments to the method, but the underlying assumption of falsification and controlling error rates remains intact.

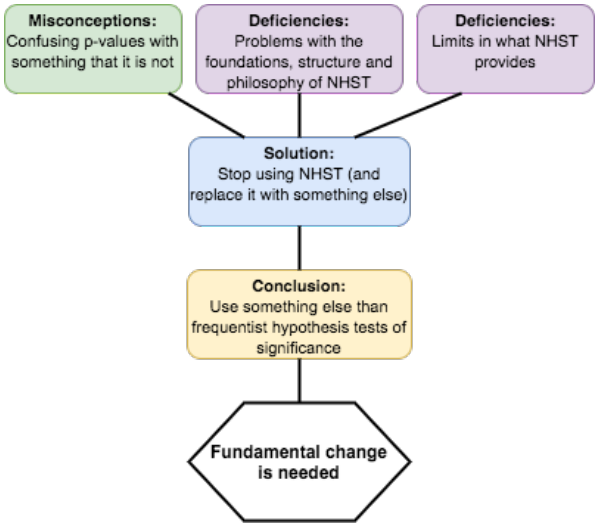


Figure 3.6: The critical zone on NHST: fundamental change is needed. This zone consists of two deficiency themes (purple), one misconceptions theme (green); one solution theme (blue), and one conclusion theme (yellow). 19 of the 42 papers ascribe to a conclusion that is part of this zone. There are an additional five papers that contain at least one misconception, deficiency, or one solution that is part of this zone and also on item that is part of another theme in this zone. There are ten papers that only have one or more misconceptions, one or more deficiencies, or one or solutions that are part of this zone while containing any descriptive themes belonging to the other phenomena of interest whose analytic theme(s) are part of another zone (these papers are: Häggström, 2017; Haig, 2017; Lu, 2015; Martin, 2014; Perezgonzalez, 2015; Rourse, 2016; Baird, 2016; Chang, 2017; Garamszegi, 2017; Perezgonzalez, 2014)

The critical zone

The critical zone (see Figure 3.6) calls for fundamental changes in the way science practices statistical inference. This is because the foundational problems of NHST,

such as its insensitivity to the specifics of the alternative hypothesis, are severe. If the alternative hypothesis is what the researcher is actually interested in, then a method is required that takes the characteristics of this hypothesis into account when evaluating the data. As Dienes and Mclatchie (2018, p.214) write:

> Bayes factors are sensitive to how vague or precise the theory is; [p]-values] are not. But, normatively, precise theories should be favored over vague ones when data appear within the predicted range.

In addition, NHST is incapable of providing that which research misconceive it to provide, such as interpreting p-values as the probability of the null hypothesis, the replication probability of the results, the realized Type I error probability, the methodological quality of the study, or that the results are due to chance. This is because, the method only provides, under the assumption that the null hypothesis is true, the probability of observing the data at hand or data deviating more strongly from the null hypothesis. As an example, Schneider (2015, p.423) writes:

> Another variant is to believe that p indicates the probability that a result is due to chance alone (i.e., sampling error). This is also not so, as [p-values] are calculated on the assumption that [$\mathcal{H}_0$] is true, this is the assumed 'chance model', so the probability that chance is the only explanation of the result is already taken to be 1. It is therefore illogical to view p as somehow measuring the probability of chance (Carver 1978).

Other methods might not suffer these deficiencies, making their result interpretations correct that would have been wrong for NHST results. For example, estimation methods and other types of hypothesis tests (e.g., Bayesian hypothesis tests) do not have a dichotomous decision rule (i.e., reject $\mathcal{H}_0$ versus retain $\mathcal{H}_0$) nor do they provide the probability of the data under the null hypothesis. They are said not suffer from the problems mention above, because these methods are not about testing a single null hypothesis and evidence comes in degrees. As Scheinder (2015, p.427) writes:

> First of all, there are inferential alternatives, which contrary to NHST do in fact assess the degree of support that data provide for hypotheses, [...] [such as] likelihood inference (e.g., Royall 1997).

The perspective on NHST from this zone is finds a clear expression in Szucs and Ioannidis (2017, p.13):

> NHST [...] does not allow for systematic knowledge accumulation. In addition, both because of its shortcomings and because it is subject to major misunderstandings it facilitates the production of non-replicable false positive reports. Such reports ultimately erode scientific credibility and result in wasting perhaps most of the research funding in some areas.

In summary of this zone, the deficiencies and their resulting misconception concerning NHST are severe to the extent that scientific progress is impeded to dangerous levels and radical methodological change is required.

3.4. Discussion

3.4.1. Summary

As this systematic review shows, NHST and its practice are diverse and intricate. We extracted data from 42 papers published between 2011 and 2018 in psychology and psychological methods journals. Combined, the papers discussed 39 deficiencies of NHST, 31 misconceptions or misuses, 20 arguments in defense of NHST, and 33 solutions to NHST problems in particular or statistical inference problems in general. Across the papers, these 123 items were supported by 508 unique pieces of evidence (i.e., references to other papers, examples, simulation studies, or mathematical proofs). The 123 arguments and solutions are employed to reach 14 types of conclusions. Collectively, these 137 items make up the landscape NHST (the unfiltered answers to our research question; Section 3.2.1), within which the author's positions, or perspective if you will, on NHST —in terms of its qualities, problems, and their solutions— can be interpreted as that which is visible to the author within this landscape from his/her point of view.

Within this landscape, we identified general zones. Based on their content, the arguments, solutions, and conclusions can be synthesized into three overarching zones enclosing certain practices and problems of NHST. The mild zone houses that which is actually in favor of NHST and only encloses problems in and solutions for policy, education, human psychology, and research culture. The moderate zone contains acknowledgments of some defects and misuse of NHST in combination with structural but not fundamental change in statistical inference (i.e., continue with frequentist hypothesis testing in a different form than NHST). The critical zone incorporates fundamental deficiencies and a systematic desire in the researchers that NHST will never be able to satisfy, in any shape or form. In this case, a complete break from NHST is required and a different statistical inference method needs to be adopted. The summary of the NHST landscape in these three zones provides a

filter/focus through which interpretation becomes more manageable.

3.4.2. Embedding an implications

At the moment, the replication crisis is not (yet) considered solved, there is still fierce debate on how to address the misuse of NHST, and disagreement continues on if and how NHST should be changed or replaced by which alternative. The contribution of our systematic review to this situation is twofold.

Firstly, our results offer a possible explanation of why these debates have not been resolved. The NHST landscape is vast. Vast enough that no single paper on the topic provides a complete overview of NHST's machination, its application, its consequences, and how it is perceived. There is more than enough space for its inhabitants to hold incompatible views without the threat of overt conflict, because their perspectives do not or minimally overlap and they have the internal coherence of their standpoint to fall back on. For instance, the proponents of preregistration and open science (Freese & King, 2018; Lindsay et al., 2016; Munafò, 2016; van 't Veer & Giner–Sorolla, 2016) build their position on arguments of psychology and research culture, which entail the improvement of replication rates. From another position in the landscape, these solutions are insufficient, because the inference method is considered at fault and these, from their perspective, superficial cosmetic fixes only give a false sense of betterment (e.g., Szollosi et al., 2019). However, they are too far apart for an actual conflict to ensue. Their positions are not build on the same or neighboring arguments, thus their opposing conclusions do not threaten that on which they are founded. Conflicts that do concern these foundations would be dialectic: each needs to improve or move (i.e., change) their positions in response to the other's attack, eventually leading to surrender/victory or consensus. Without such threats to one's position, one can continue to feel safe; out of the reach of the arguments of others and within the confines of internal consistency of one's arguments.

In such a case, published denouncements of each other's positions do not have to be perceived as carrying the force that would motivate change in one's position. As another example, there are several many-authors papers that call for different changes in the application of statistical methods. One such paper suggests leaving NHST as it is, but only lowering the threshold of statistical significance (Benjamin et al., 2018),[4] which is consistent with our mild zone. Another suggests taking

[4]It needs to be noted that (many of) the authors actually prefer Bayesian hypothesis tests, but see this solution as more feasible for the near future. However, taking this into account does not change the argument.

both Type I and Type II error rates into account and have scientist sets them with respect to the particularities of their research aim and subject (i.e., the Neyman-Pearson approach to frequentist hypothesis testing; Lakens, Adolfi, et al., 2018), which is consistent with our moderate zone. A third paper suggest abandoning hypothesis tests and thresholds in favor of continues measures of estimation and model adequacy (McShane et al., 2019), which is consistent with our critical zone. Hence, these papers draw on different parts of the NHST landscape (i.e., arguments). As a result, they argue more past each other than against each other. Consequently, as long as those trying to resolve these problems do not consider each other's position in the NHST landscape and do not target those areas where their confliction positions overlap, it is likely that conflicts will not be treating enough to motivate defense or change.

Secondly, our data can be used to move towards a resolution of these debates on how scientific practice should be improved. The overview our data provides could serve as a starting point for statisticians and methodologists to formulate alternatives that are more likely to have the support from a large group of researchers. Moreover, it provides a complete and neutral classification for academic researchers using NHST to reflect upon their own position and account for the advantages and drawbacks of using NHST. More particular, our results can be used to identify where in the NHST landscape particular positions overlap. The arguments and solutions in these overlapping areas could be the focus of evaluation or the subject of empirical research. This would allow a systematic comparison and assessment of the particular debate positions.

In addition, our data provides some indication of relevance and interconnection. Some deficiencies and misconceptions are more prevalent in the papers than other. One explanation for this prevalence is that these problems are more relevant and severe and thus garner more attention. Within the papers and our zones, these potentially relevant problems are connected to particular solutions, which could therefore be favored over their alternatives. Concretely, this, and our data set in general, could be of assistance to those intent on improving science (e.g., policy makers, journal editors, educators, methodologists, etc.) when deciding on how to allocate limited resource. For instance, those working towards the improvement of scientific practice through better statistics education could benefit from selecting those misconceptions of NHST that are most prevalent and evaluate if these curricular changes reduce these misconceptions.

Furthermore, an overview such as this could prove useful for those developing re-

search methods and want to take its adequate application by fallible human scientists into account. Scientists might be perceived by lay people and other scientists as more rational, objective, and open-minded than non-scientists (Veldkamp, Hartgerink, van Assen, & Wicherts, 2017), but, even if this is the case, scientists are not infallible. As our data shows, there are many misconceptions and misuses possible when it comes to statistical inference via NHST, which might result from psychological biases and research cultures that do not incentivize methodological rigor. Also, research show that reporting error in statistical results are prevalent in published papers (e.g., Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, 2014). It is thus advisable that these fallibilities of scientists should to be taken into account, to which our overview can be of assistance, when developing or implementing alternatives to NHST with the aim of improving the quality of scientific inference.

3.4.3. Limitations

As all research, this review suffers from several limitations. First and foremost, the methodology for the systematic review of opinion literature is not yet fully developed. We adopted methods for mixed-methods and qualitative systematic reviews, which offered guidelines. However, on several occasions, choices had to be made that had no precedent in the current methodological literature. The most prominent of these choices were the exclusion of empirical papers; using claims in the papers as our data; and categorizing the phenomena of interest in deficiencies, misconceptions, NHST defenses, solutions, and conclusions. Empirical studies of NHST and its use might also contain arguments and solutions; other types of text fragments could have been used as data; and the phenomena of interest could have been combined, further sub-divided, or differently delineated. For instance, the requirement of a positive element or alternative in a proposed solution would have precluded 'stop using NHST' from being an instance of that phenomenon of interest. Different choices would have led to different result. However, without proper methodological guidelines it is impossible to predict if these alternative choices would have let to meaningful differences in results and if these alternative results would be better or worse.

Secondly, a quality assessment of the included literature is lacking. It is general practice for systematic reviews to evaluate the quality of the data and either only includes data that meet certain quality criteria (e.g., Tong et al., 2012) or conduct sensitivity analysis of the review results with respect to the quality assessments of the data (e.g., Patsopoulos, Evangelou, & Ioannidis, 2008). However, at the moment,

there is no tool available for the evaluation of the quality of particular arguments, nor the structure of a set of arguments to their conclusion. Thus, we included all arguments even though some would consider them to be clearly false or of such a low quality that they should not be included in the overview. When comparing the arguments, some clearly contradict each other, such as Miller claiming that "the heart of NHST is a simple, intuitive, and familiar "common sense" logic" (Miller, 2017, p.664) while others claim that NHST is incoherent (and thus illogical). This is a clear indication that at least one of them must be wrong. In such a case, a quality assessment might be able to arbitrate between conflicting cases.

Thirdly, although the identified zones within the NHST landscape generally overlap with the perspectives of the authors, there is lack of complete correspondence. However, it does not have to be unexpected that some authors use arguments or solutions across more than one zone. It is not necessary that the authors' positions are exclusively dependent on what is present in the landscape. For instance, personal agenda, position within their academic field, or their particular education might play a role as well. To extent the landscape metaphor, borders are not always drawn along structure in the landscape, such as rivers, but logistic, political, historical, and other factors are also taken into account. consequently, this does not necessarily have to be considered as a limitation, but could be seen as highlighting the diversity and intricacies of the NHST landscape.

Finally, the scope of our review is limited. Time and resources required constraints. We only included essay and opinion literature from psychology and methodological journals, published between 2011 and 2018. The replication crisis and the consequent attention for methodological evaluation and reform made this, in our opinion, the most promising boundaries to work within. However, other fields might have a different perspective on NHST. Specifically, researchers in those fields might differ from psychologists in their misconceptions and misuses. Or, the authors of NHST papers in these fields might emphasize deficiencies of NHST that are ignored by authors of such papers working in psychology and vice versa. Furthermore, different deficiencies and misconceptions might have been in the spotlight before the advent of the replication crisis in 2011 and it could be possible that different problems are addressed in empirical literature concerning NHST. In short that is ample room for broadening and deepening the understanding of NHST in particular and statistical inference in general via systematic literature review.

3.4.4. Suggestions for further research

Our paper offers handholds for the extension of our review and the development of systematic review methodology for essay and opinion literature. The access we provide to our methods and progress notes will not only allow others to evaluate our work, but also extend and improve it. Specifically, this review can be extended to other academic fields and further back in time. These extensions will allow us to gain insight in the differences and similarities between fields and across time on the perspectives and practice of NHST. Such a nuanced overview will be of value in the continuing debate on how statistical inference is and should be practiced.

This paper also counts as a proof of principle. As far as we know, this is the first systematic review of essay and opinion literature. Its successful completion is an indication that such reviews are possible in principle. We expected that our methods can be developed into a systematic review methodology with general application, which itself can be investigated and evaluated. The academic literature is teeming with essays on a wide variety of topics. Many of them have a shared topic and argue against or support each other. At the moment, essays concerning open science, the replication crisis, and preregistration and methodological rigor are prime examples of topics on which many such essays are published. An adequate systematic review methodology could be used to catalog the copious amounts of such papers and get a clear perspective on topics that are of clear current relevance.

3.4.5. Closing remarks

In this paper we provided an overview on the perspectives on NHST and its practice. The extensive data set that resulted from our review contains more information than we could properly describe and analyze in one paper. Thus we conclude with the wish that others might use our data and further investigate this fascinating topic.

4

STATISTICAL INFERENCE

Abstract:
When data analysts operate within different statistical frameworks (e.g., frequentist versus Bayesian, emphasis on estimation versus emphasis on testing), how does this impact the qualitative conclusions that are drawn for real data? To study this question empirically we selected from the literature two simple scenarios –involving a comparison of two proportions and a Pearson correlation– and asked four teams of statisticians to provide a concise analysis and a qualitative interpretation of the outcome. The results showed considerable overall agreement; nevertheless, this agreement did not appear to diminish the intensity of the subsequent debate over which statistical framework is more appropriate to address the questions at hand.

This chapter is adapted from van Dongen, N. N.N., van Doorn, J. B., Gronau, Q. F., van Ravenzwaaij, D., Hoekstra, R., Haucke, M. N., ... & Wagenmakers, E.J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician*, 73(sup1), 328-339.

4.1. Introduction

When analyzing a specific data set, statisticians usually operate within the confines of their preferred inferential paradigm. For instance, frequentist statisticians interested in hypothesis testing may report *p*-values, whereas those interested in estimation may seek to draw conclusions from confidence intervals. In the Bayesian realm, those who wish to test hypotheses may use Bayes factors and those who wish to estimate parameters may report credible intervals. And then there are likelihoodists, information-theorists, and machine-learners — there exists a diverse collection of statistical approaches, many of which are philosophically incompatible.

Moreover, proponents of the various camps regularly explain why their position is the most exalted, either in practical or theoretical terms. For instance, in a well-known article 'Why Isn't Everyone a Bayesian?', Bradley Efron claimed that "The high ground of scientific objectivity has been seized by the frequentists" (Efron, 1986, p. 4), upon which Dennis Lindley replied that "Every statistician would be a Bayesian if he took the trouble to read the literature thoroughly and was honest enough to admit that he might have been wrong." (Lindley, 1986, p. 7). Similarly spirited debates occurred earlier, notably between Fisher and Jeffreys (e.g., Howie, 2002) and between Fisher and Neyman. Even today, the paradigmatic debates show no sign of stalling, neither in the published literature (e.g., Benjamin et al., 2018; McShane, Gal, Gelman, Robert, & Tackett, in press; Wasserstein & Lazar, 2016) nor on social media.

The question that concerns us here is purely pragmatic: 'does it matter?' In other words, will reasonable statistical analyses on the same data set, each conducted within their own paradigm, result in qualitatively similar conclusions (Berger, 2003)? One of the first to pose this question was Ronald Fisher. In a letter to Harold Jeffreys, dated March 29, 1934, Fisher proposed that "From the point of view of interesting the general scientific public, which really ought to be much more interested than it is in the problem of inductive inference, probably the most useful thing we could do would be to take one or more specific puzzles and show what our respective methods made of them" (J. H. Bennett, 1990, p. 156; see also Howie, 2002, p. 167). The two men then proceeded to construct somewhat idiosyncratic statistical 'puzzles' that the other found difficult to solve. Nevertheless, three years and several letters later, on May 18, 1937, Jeffreys stated that "Your letter confirms my previous impression that it would only be once in a blue moon that we would disagree about the inference to be drawn in any particular case, and that in the exceptional cases we would both be a bit doubtful" (J. H. Bennett, 1990, p. 162). Similarly, W. Edwards, Lindman, and

Savage (1963) suggested that well-conducted experiments often satisfy Berkson's *interocular traumatic test* – "you know what the data mean when the conclusion hits you between the eyes" (p. 217). Nevertheless, surprisingly little is known about the extent to which, in concrete scenarios, a data analyst's statistical plumage affects the inference.

Here we revisit Fisher's challenge. We invited four groups of statisticians to analyze two real data sets, report and interpret their results in about 300 words, and discuss these results and interpretations in a round-table discussion. The data sets are provided online at `https://osf.io/hykmz/` and described below. In addition to providing an empirical answer to the question 'does it matter?', we hope to highlight how the same data set can give rise to rather different statistical treatments. In our opinion, this method variability ought to be acknowledged rather than ignored (for a complementary approach see Silberzahn et al., in press).[1]

The selected data sets are straightforward: the first data set concerns a 2x2 contingency table, and the second concerns a correlation between two variables. The simplicity of the statistical scenarios is on purpose, as we hoped to facilitate a detailed discussion about assumptions and conclusions that could otherwise have remained hidden underneath an unavoidable layer of statistical sophistication. The full instructions for participation can be found online at `https://osf.io/dg9t7/`.

4.2. Data Set I: Birth Defects and Cetirizine Exposure

4.2.1. Study summary

Cetirizine is a non-sedating long-acting antihistamine with some mast-cell stabilizing activity. It is used for the symptomatic relief of allergic conditions, such as rhinitis and urticaria, which are common in pregnant women. In the study of interest, Weber-Schoendorfer and Schaefer (2008) aimed to assess the safety of cetirizine during the first trimester of pregnancy when used. The pregnancy outcomes of a cetirizine group (n = 181) were compared to those of the control group (n = 1685; pregnant women who had been counseled during pregnancy about exposures known to be non-teratogenic). Due to the observational nature of the data, the allocation of participants to the groups was non-randomized. The main point of interest was the

[1] In contrast to the current approach, Silberzahn et al. (ress) used a relatively complex data set and did notemphasize the differences in interpretation caused by the adoption of dissimilar statistical paradigms.

rate of birth defects.[2] Variables of the data set[3] are described in Table 4.1 and the data are presented in Table 4.2.

4.2.2. Cetirizine research question

Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects? In the next sections each of four data analysis teams will attempt to address this question.

Table 4.1: Variable names and their description.

Variable	**Description**
CetirizineExposure	Whether exposed to cetirizine
BirthDefect	Whether birth defects were detected
Counts	Count data for each cell

Table 4.2: Cetirizine exposure and birth defects.

	Birth Defects		
Cetirizine Exposure	No	Yes	Total
No	1588	97	1685
Yes	167	14	181
Total	1755	111	1866

4.2.3. Analysis and interpretation by Lakens and Hennig

Preamble

Frequentist statistics is based on idealised models of data generating processes. We cannot expect these models to be literally true in practice, but it is instructive to see whether data are consistent with such models, which is what hypothesis tests and confidence intervals allow us to examine. We do appreciate that automatic procedures involving for example fixed significance levels allow us to control error probabilities assuming the model, but given that the models do never hold precisely, and that there are often issues with measurement or selection effects, in most cases we think it is prudent to interpret outcomes in a coarse way rather than to read too much meaning into, say, differences between p-values of 0.047 and 0.062. We stick to

[2] The original study focused on Cetirizine-induced differences in major birth defects, spontaneous abortions, and preterm deliveries. We decided to look at all birth defects, because the sample sizes were larger for this comparison and we deemed the data more interesting.

[3] The data set is made available on the OSF repository: `https://osf.io/hykmz/`.

quite elementary methodology in our analyses.

Analysis and software

We performed a Pearson's Chi-squared test with Yates' continuity correction to test for dependence between exposure of pregnant women exposed to cetirizine and birth defects using the `chisq.test` function in R software version 3.4.3 (R Development Core Team, 2004). However, because Weber-Schoendorfer and Schaefer (2008) wanted "to assess the safety of cetirizine during the first trimester of pregnancy", their actual research question is whether we can reject the presence of a meaningful effect. We therefore performed an equivalence test on proportions (Chen, Tsong, & Kang, 2000) as implemented in the TOSTtwo.prop function in the TOSTER package (Lakens, 2017a).

Results and interpretation

The chi-squared test yielded $\chi^2(1, N = 1866) = 0.817$, $p = .366$, which suggests that the data are consistent with an independence model at any significance level in general use. The answer to the question whether the drug is safe depends on a smallest effect size of interest (when is the difference in birth defects considered too small to matter?). This choice, and the selection of equivalence bounds more generally, should always be justified by clinical and statistical considerations pertinent to the case at hand. In the absence of a discussion of this essential aspect of the study by the authors, and in order to show an example computation, we will test against a difference in proportions of 10%, which, although debatable, has been suggested as a sensible bound for some drugs (see Röhmel, 2001, for a discussion).

An equivalence test against a difference in proportions [$M_{dif} = 0.02$, 95% CI:−0.02; 0.06] of 10% based on Fishers exact z-test was significant, $z = -3.88$, $p < 0.001$. This means that we can reject differences in proportions as large, or larger, than 10%, again at any significance level in general use.

Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects? Based on the current study, there is no evidence that cetirizine exposure during pregnancy is associated with a higher incidence of birth defects. Obviously this does not mean that cetirizine is safe; in fact the observed birth defect rate in the cetirizine group is about 2% higher than without exposure, which may or may not be explained by random variation. Is cetirizine during the first trimester of

pregnancy 'safe'? If we accept a difference in the proportion of birth defects of 10%, and desire a 5% long run error rate, there is clear evidence that the drug is safe. However, we expect that a cost-benefit analysis would suggest proportions of 5% to be unacceptably high, which is in the 95% confidence interval and therefore well compatible with the data. Therefore, we would personally consider the current data inconclusive.

4.2.4. Analysis and interpretation by Morey and Homer

Fitting a classical logistic model with the binary birth defect outcome predicted from the cetirizine indicator confirmed the non-significant relationship ($p = 0.287$). The point estimate of the effect of taking cetirizine is to increase the odds of the birth defect by only 37%. At the baseline levels of birth defects in the non-cetirizine-exposed sample (approximately 6%), this would amount to about an extra two birth defects in every hundred pregnancies in the cetirizine-exposed group.

There are several problems with taking these data as evidence that cetirizine is safe. The first is the observational nature of the data. We have no way of knowing whether an apparent effect — or lack of effect — reflects confounds. Suppose, though, that we set this question aside and assess the evidence that birth defects are not more common in the cetirizine group. We can use a classical one-sided CI to determine the size of the differences we can rule out. We call the upper bound of the $100(1 - \alpha)\%$ CI the "worst case" for that confidence coefficient. Figure 4.1 shows that at 95%, the worst case odds increase is for the cetirizine group is 124%. At 99.5%, the worst case increase is 195%. We can translate this into more a more intuitive metric of numbers of birth defects: at baseline rates of birth defects, these would amount to additional 6 and 10 birth defects per 100, respectively (Figure 4.2).

The large p-value of the initial significance test suggests we cannot rule out that cetirizine group has lower rates of birth defects; the one-sided test assuming a decrease in birth defects as the null would not yield a rejection except at high α levels. But also, the "worst case" analysis using the upper bound of the one-sided CI suggests we also cannot rule out a substantial *increase* in birth defects in the cetirizine group.

We are unsure whether cetirizine is safe, but it seems clear to us that these data do not provide much evidence of its relative safety, contrary to what Weber-Schoendorfer and Schaefer suggest.

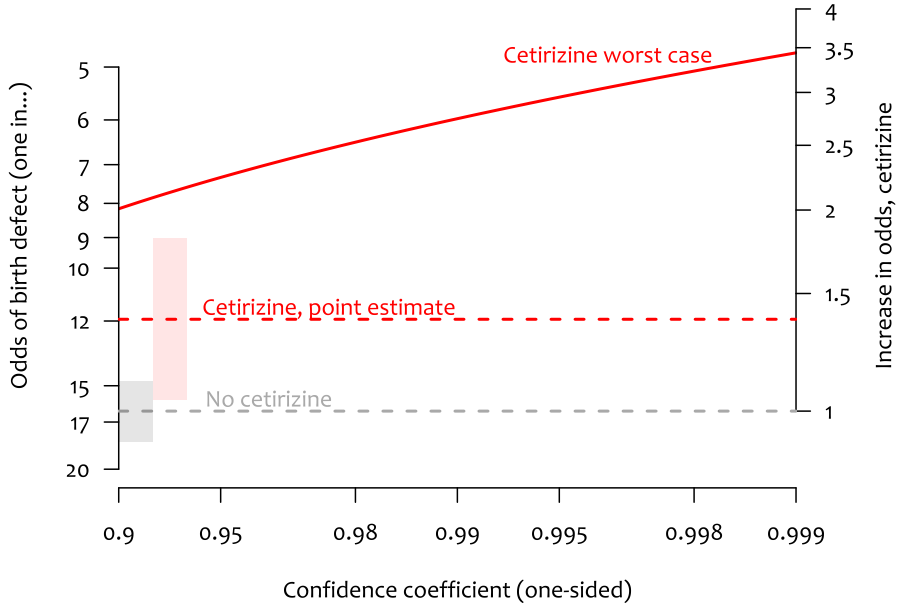


Figure 4.1: Estimates of the odds of a birth defect when no cetirizine (control) was taken during pregnancy and when cetirizine was taken. Horizontal dashed lines and shaded regions show point estimates and standard errors. The solid line labeled "Cetirizine worst case" shows the upper bound of the one-sided CI as a function of the confidence coefficient (x-axis). The right axis shows the estimated increase in odds of a birth defect for the cetirizine group compared to the control group.

4.2.5. Analysis and interpretation by Gronau, van Doorn, and Wagenmakers

We used the model proposed by Kass and Vaidyanathan (1992):

$$\begin{aligned} \log\left(\frac{p_1}{1-p_1}\right) &= \beta - \frac{\psi}{2} \\ \log\left(\frac{p_2}{1-p_2}\right) &= \beta + \frac{\psi}{2} \\ y_1 &\sim \text{Binomial}(n_1, p_1) \\ y_2 &\sim \text{Binomial}(n_2, p_2). \end{aligned} \tag{4.1}$$

Here, $y_1 = 97$, $n_1 = 1{,}685$, $y_2 = 14$, and $n_2 = 181$, p_1 is the probability of a birth defect in the control group, and p_2 is that probability in the cetirizine group. Probabilities p_1 and p_2 are functions of model parameters β and ψ. Nuisance parameter β corresponds to the grand mean of the log odds, whereas the test-relevant parameter ψ corresponds to the log odds ratio. We assigned β a standard normal prior and used a zero-centered normal prior with standard deviation σ for the log odds ratio

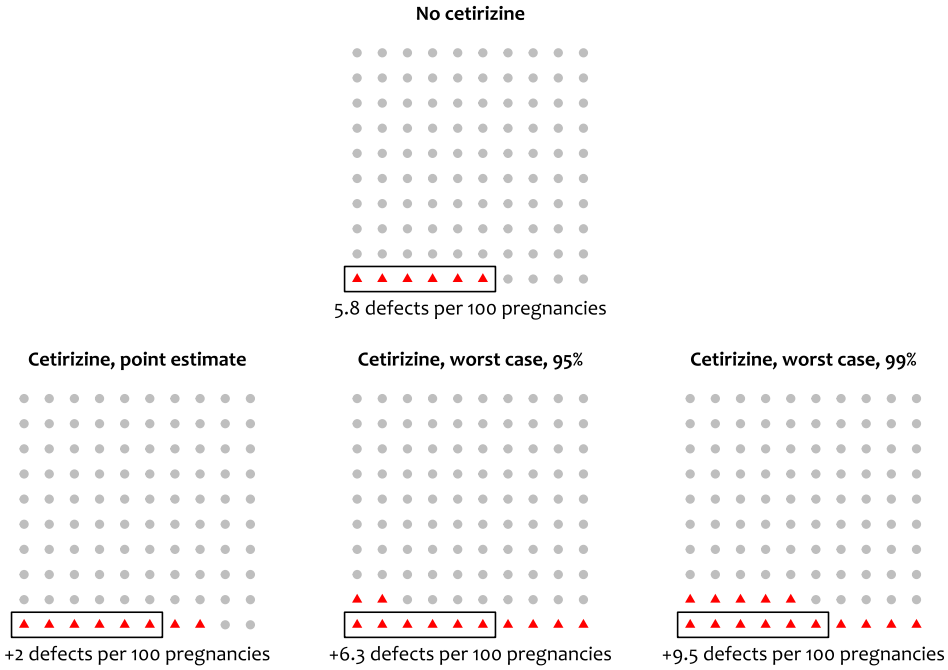


Figure 4.2: Frequency representations of the number of birth defects expected under various scenarios. Top: Expected frequency of birth defects when cetirizine was not taken (control). Bottom-left: Point estimate of the expected frequency of birth defects when cetirizine is taken. Bottom-middle (bottom-right): Upper bound of a one-sided 95% (99%) CI for the expected frequency of birth defects when cetirizine was taken. Because the analysis is intended to be comparative, in the bottom panels the no-cetirizine estimate was assumed to be the truth when calculating the increase in frequency.

ψ. Inference was conducted with `Stan` (Carpenter et al., 2017; Stan Development Team, 2016) and the `bridgesampling` package (Gronau, Singmann, & Wagenmakers, 2020). For ease of interpretation, the results will be shown on the odds ratio scale.

Our first analysis focuses on estimation and uses $\sigma = 1$. The result, shown in the left panel of Figure 4.3, indicates that the posterior median equals 1.429, with a 95% credible interval ranging from 0.793 to 2.412. This credible interval is relatively wide, indicating substantial uncertainty about the true value of the odds ratio.

Our second analysis focuses on testing and quantifies the extent to which the data support the skeptic's $\mathcal{H}_0 : \psi = 0$ versus the proponent's $\mathcal{H}_1$. To specify $\mathcal{H}_1$ we initially use $\sigma = 0.4$ (i.e., a mildly informative prior; Diamond & Kaul, 2004), truncated at zero to respect the fact that cetirizine exposure is hypothesized to cause a *higher* incidence of birth defects: $\mathcal{H}_+ : \psi \sim N^+(0, 0.4^2)$.

As can be seen from the right panel of Figure 4.3, the observed data are predicted about 1.8 times better by $\mathcal{H}_+$ than by $\mathcal{H}_0$. According to Jeffreys (1961, Appendix B),

this level of support is "not worth more than a bare mention". To investigate the robustness of this result we explored a range of alternative prior choices for σ under $\mathcal{H}_+$, varying it from 0.01 to 2. The results of this sensitivity analysis are shown in Figure 4.4 and reveal that across a wide range of priors, the data never provide more than anecdotal support for one model over the other. When σ is selected post-hoc to maximize the support for $\mathcal{H}_+$ this yields $BF_{+0} = 1.84$, which, starting from a position of equipoise, raises the probability of $\mathcal{H}_+$ from 0.50 to about 0.65, leaving a posterior probability of 0.35 for $\mathcal{H}_0$.

In sum, based on this data set we cannot draw strong conclusions about whether or not cetirizine exposure during pregnancy is associated with a higher incidence of birth defects. Our analysis shows an 'absence of evidence', not 'evidence for absence'.

4.2.6. Analysis and interpretation by Gelman

I summarized the data with a simple comparison: the proportion of birth defects is 0.06 in the control group and 0.08 in the cetirizine group. The difference is 0.02 with a standard error of 0.02. I got essentially the same result with a logistic regression predicting birth defect: the coefficient of cetirizine is 0.3 with a standard error of 0.3. I performed the analyses in R using rstanarm (code available at `https://osf.io/nh4gc/`).

I then looked up the article, "The safety of cetirizine during pregnancy: A prospective observational cohort study", by Weber-Schoendorfer and Schaefer (2008) and noticed some interesting things:

(a) The published article gives $N = 196$ and 1686 for the two groups, not quite the same as the 181 and 1685 for the "all birth defects" category. I couldn't follow the exact reasoning.[4]

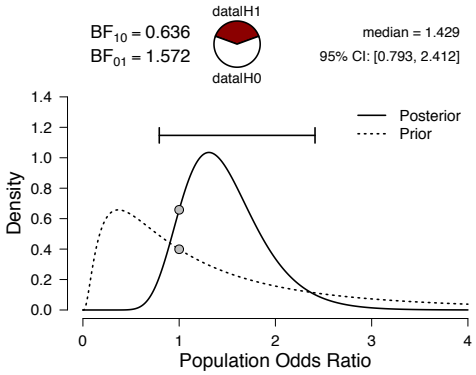
(b) The two groups differ in various background variables: most notably, the cetirizine group has a higher smoking rate (17% compared to 10%).

(c) In the published article, the outcome of focus was "major birth defects", not the "all birth defects" given for us to study.

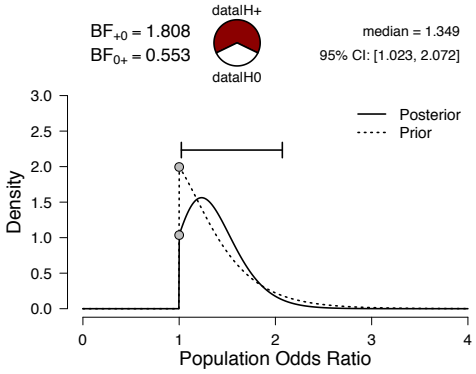
(d) The published article has a causal aim (as can be seen, for example, from the word "safety" in its title); our assignment is purely observational.

Now the question, "Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects?" I have not read the literature on the topic. To understand how the data at hand address this question, I would like to think of

[4]Clarification: the original paper does not provide a rationale for why several participants were excluded from the analysis.



(a) Estimation results.



(b) Testing results.

Figure 4.3: Gronau, van Doorn, and Wagenmakers' Bayesian analysis of the cetirizine data set. The left panel shows the results of estimating the log odds ratio under $\mathcal{H}_1$ with a two-sided standard normal prior. For ease of interpretation, results are displayed on the odds ratio scale. The right panel shows the results of testing the one-sided alternative hypothesis $\mathcal{H}_+ : \psi \sim \mathcal{N}^+(0, 0.4^2)$ versus the null hypothesis $\mathcal{H}_0 : \psi = 0$. Figures inspired by JASP (`jasp-stats.org`).

the mapping from prior to posterior distribution. In this case, the prior would be the distribution of association with birth defects of all drugs of this sort. That is, imagine a population of drugs, $j = 1, 2, \ldots$, taken by pregnant women, and for each drug, define θ_j as the proportion of birth defects among women who took drug j, minus the proportion of birth defects in the general population. Just based on my general understanding (which could be wrong), I would expect this distribution to be more positive than negative and concentrated near zero: some drugs could

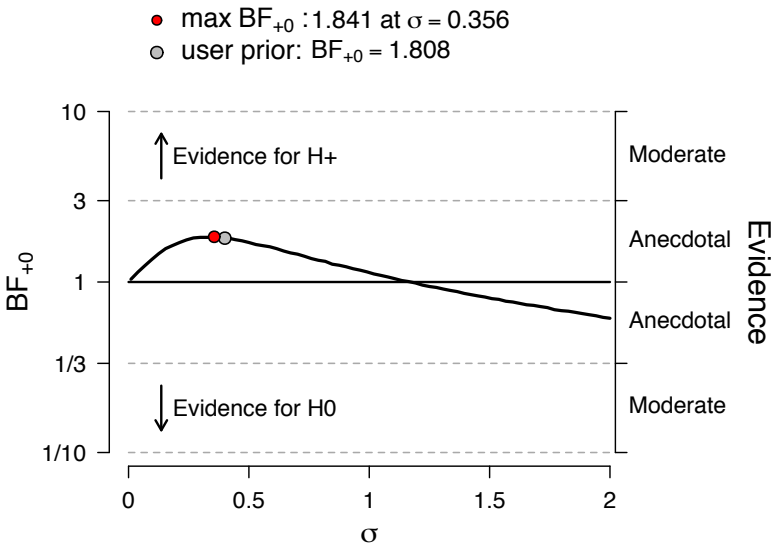


Figure 4.4: Sensitivity analysis for the Bayesian one-sided test. The Bayes factor BF_{+0} is a function of the prior standard deviation σ. Figure inspired by JASP.

be mildly protective against birth defects or associated with low-risk pregnancies, most would have small effects and not be strongly associated with low or high-risk pregnancies, and some could cause birth defects or be taken disproportionately by women with high-risk pregnancies. Supposing that the prior is concentrated within the range (−0.01, +0.01), the data would not add much information to this prior.

To answer, "Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects?", the key question would seem to be whether the drug is more or less likely to be taken by women at higher risk of birth defects. I'm guessing that maternal age is a big predictor here. In the reported study, average age of the exposed and control groups was the same, but I don't know if that's generally the case or if the designers of the study were purposely seeking a balanced comparison.

4.3. Data Set II: Amygdalar Activity and Perceived Stress

4.3.1. Study summary

In a recent study published in the *Lancet*, Tawakol et al. (2017) tested the hypothesis that perceived stress is positively associated with resting activity in the amygdala. In the second study reported in Tawakol et al. (2017), n = 13 individuals with an

increased burden of chronic stress (i.e., a history of post-traumatic stress disorder or PTSD) were recruited from the community, completed a Perceived Stress Scale (i.e., the PSS-10; S. Cohen, Kamarck, & Mermelstein, 1983) and had their amygdalar activity measured. Variables of the data set[5] are described in Table 4.3, the raw data are presented in Table 4.4, and the data are visualized in Figure 4.5.

4.3.2. Amygdala research question

Do PTSD patients with high resting state amygdalar activity experience more stress? In the next sections each of four data analysis teams will attempt to address this question.

Table 4.3: Variable names and their description.

Variable	Description
Perceived stress scale	Participant score on the PSS
Amygdalar activity	Intensity of amygdalar resting state activity

Table 4.4: Raw data as extracted from Figure 5 in Tawakol et al. (2017), with help of Jurgen Rusch, Philips Research Eindhoven.

Perceived Stress Scale	Amygdalar Activity
12.0103	5.2418
32.0350	6.8601
22.0296	6.4402
20.0079	5.4620
24.0155	5.4439
24.0155	5.3349
24.0155	5.4216
26.0082	5.5176
28.0120	5.1615
21.9872	4.7114
21.9872	4.1844
20.0138	4.3079
16.0088	3.3015

4.3.3. Analysis and interpretation by Lakens and Hennig

Analysis and software

We calculated tests for uncorrelatedness based on both Pearson's product-moment correlation and Spearman's rank correlation using the `cor.test` function in the stats

[5]The data set is made available on the OSF repository: `https://osf.io/hykmz/`.

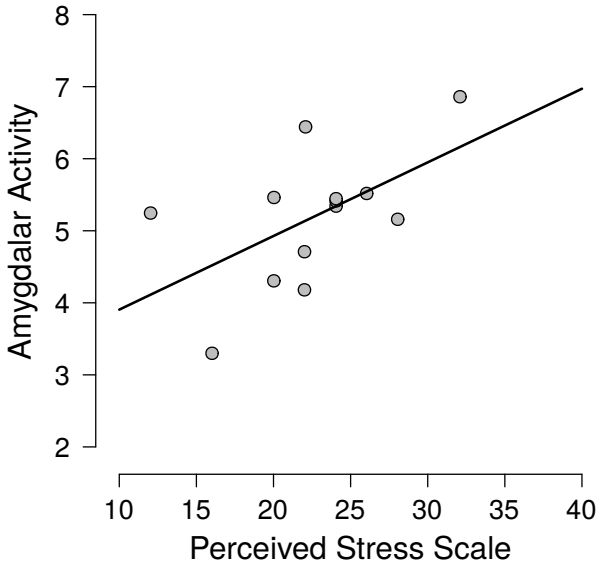


Figure 4.5: Scatter plot of amygdalar activity and perceived stress in 13 patients with PTSD. Data extracted from Figure 5 in Tawakol et al. (2017), with help of Jurgen Rusch, Philips Research Eindhoven.

package in R 3.4.3.

Results and interpretation

The Pearson correlation between perceived stress and resting activity in the amygdala is $r = 0.555$, and the corresponding test yields $p = 0.047$. Although this is just smaller than the conventional 5% level, we do not consider it as clear evidence for nonzero correlation. From the appendix it becomes clear that the reported correlations are exploratory: "Patients completed a battery of self-report measures that assessed variables that may correlate with PTSD symptom severity, including comorbid depressive and anxiety symptoms (MADRS, HAMA) and a well-validated questionnaire Perceived Stress Scale (PSS-10)." Therefore, corrections for multiple comparisons would be required to maintain a given significance level. The article does not provide us with sufficient information to determine the number of tests that were performed, but corrections for multiple comparisons would thus be in order. Consequently, the fairly large observed correlation and the borderline significant p-value can be interpreted as an indication that it may be worthwhile to investigate the issue with a larger sample size, but do not give conclusive evidence. Visual inspection of the data does not give any indication against the validity of using Pearson's correlation, but with $N = 13$ we do not have very strong information regarding the

distributional shape. The analogous test based on Spearman's correlation yields p = 0.062, which given its weaker power is compatible with the qualitative interpretation we gave based on the Pearson correlation.

Do PTSD patients with high resting state amygdalar activity experience more stress? Based on the current study, we can not conclude that PTSD patients with high resting state amygdalar activity experience more stress. The single p = 0.047 is not low enough to indicate clear evidence against the null hypothesis after correcting for multiple comparisons when using an alpha of .05. Therefore, our conclusion is: Based on the desired error rate specified by the authors, we can't reject a correlation of zero between amygdalar activity and participants' score on the perceived stress scale. With a 95% CI that ranges from r = 0 to r = 0.85, it seems clear that effects that would be considered interesting cannot be rejected in an equivalence test. Thus, the results are inconclusive.

4.3.4. Analysis and interpretation by Morey and Homer

The first thing that should be noted about this data set is that it contains a meager 13 data points. The linear correlation by Tawakol et al. (2017) depends on assumptions that are for all intents and purposes unverifiable with this few participants. Add to this the difficulty of interpreting the independent variable — a sum of ordinal responses — and we are justified being skeptical of any hard conclusions from these data.

Suppose, however, that the relationship between these two variables was best characterized by a linear correlation, and set aside any worries about assumptions. The large observed correlation coupled with the marginal p value should signal to us that a wide range of correlations are not ruled out by the data. Consider that the 95% CI on the Pearson correlation is [0.009,0.848]; the 99.5% CI is [−0.254,0.908]. Negligible correlations are not ruled out; due to the small sample size, any correlation from "essentially zero" to "the correlation between height and leg length" (that is, very high) is consistent with these data. The solid curve in Figure 4.6 shows the lower bound of the confidence interval on the linear correlation for a wide range of confidence levels; they are all negligible or even negative.

Finally, the authors did not show that this correlation is selective to the amygdala; it seems to us that interpreting the correlation as evidence for their model requires selectivity. It is important to interpret the correlation in the context of the relationship between amygdala resting state activity, stress, and cardiovascular disease. If one could not show that amygdala resting-state activation showed a substantially higher

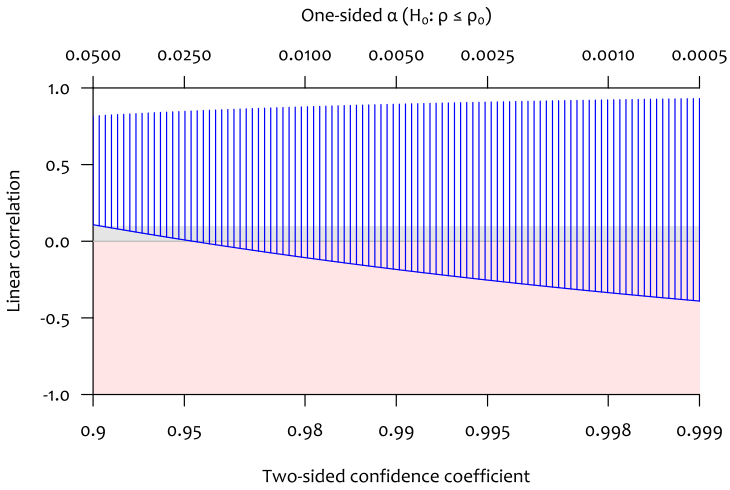


Figure 4.6: Confidence intervals and one-sided tests for the Pearson correlation as a function of the confidence coefficient. The vertical lines represent the confidence intervals (confidence coefficient on lower axis), and the curve represents the value that is just rejected by the one-sided test (α on the upper axis).

correlation with stress than other brain regions not implicated in the model, this would suggest that the correlation cannot be used to bolster their case. Given the uncertainty in the estimate of the correlation, there is little chance of being able to show this.

All in all, we're not sure that the information in these thirteen participants is enough to say anything beyond "the correlation doesn't appear to be (very) negative."

4.3.5. Analysis and interpretation by van Doorn, Gronau, and Wagenmakers

We applied Harold Jeffreys's test for a Pearson correlation coefficient ρ (Jeffreys, 1961; Ly, Marsman, & Wagenmakers, 2018) as implemented in JASP (`jasp-stats.org`; JASP Team, 2018).[6] Our first analysis focuses on estimation and assigns ρ a uniform prior from −1 to 1. The result, shown in the left panel of Figure 4.7, indicates that the posterior median equals 0.47, with a 95% credible interval ranging from −0.01 to 0.81. As can be expected with only 13 observations, there is great uncertainty about the size of ρ.

Our second analysis focuses on testing and quantifies the extent to which the data support the skeptic's $\mathcal{H}_0 : \rho = 0$ versus the proponent's $\mathcal{H}_1$. To specify $\mathcal{H}_1$ we initially use Jeffreys's default uniform distribution, truncated at zero to respect the

[6]JASP is an open-source statistical software program with a graphical user interface that supports both frequentist and Bayesian analyses.

directionality of the hypothesized effect: $\mathcal{H}_+ : \rho \sim U[0,1]$.

As can be seen from the right panel of Figure 4.7, the observed data are predicted about 3.9 times better by $\mathcal{H}_+$ than by $\mathcal{H}_0$. This degree of support is relatively modest: when $\mathcal{H}_+$ and $\mathcal{H}_0$ are equally likely a priori, the Bayes factor of 3.9 raises the posterior plausibility of $\mathcal{H}_+$ from 0.50 to 0.80, leaving a non-negligible 0.20 for $\mathcal{H}_0$.

To investigate the robustness of this result we explored a continuous range of alternative prior distributions for ρ under $\mathcal{H}_+$; specifically, we assigned ρ a stretched Beta(a,a) distribution truncated at zero, and studied how the Bayes factor changes with $1/a$, the parameter that quantifies the prior width and governs the extent to which $\mathcal{H}_+$ predicts large values of r. The results of this sensitivity analysis are shown in Figure 4.8 and confirm that the data provide modest support for $\mathcal{H}_+$ across a wide range of priors. When the precision is selected post-hoc to maximize the support for $\mathcal{H}_+$ this yields $\text{BF}_{+0} = 4.35$, which –under a position of equipose– raises the plausibility of $\mathcal{H}_+$ from 0.50 to about 0.81, leaving a posterior probability of 0.19 for $\mathcal{H}_0$.

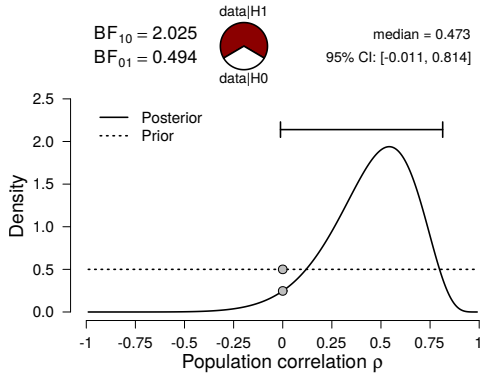
A similar sensitivity analysis could be conducted for $\mathcal{H}_0$ by assuming a 'perinull' (Tukey, 1995, p. 8) — a distribution tightly centered around $\rho = 0$ rather than a point mass on $\rho = 0$. The results will be qualitatively similar.

In sum, the claim that 'PTSD patients with high resting state amygdalar activity experience more stress' receives modest but not compelling support from the data. The 13 observations do not warrant categorical statements, neither about the presence nor about the strength of the hypothesized effect.

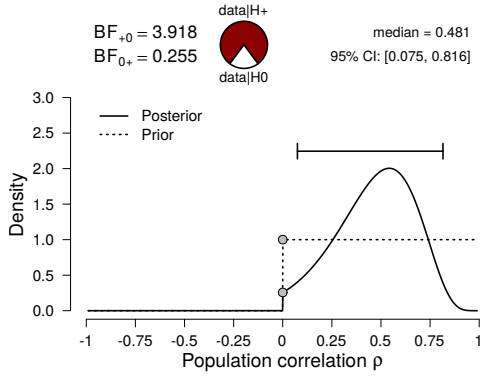
4.3.6. Analysis and interpretation by Gelman

I summarized the data with a simple scatterplot and a linear regression of logarithm of perceived stress on logarithm of amygdalar activity, using log scales because the data were all-positive and it seemed reasonable to model a multiplicative relation. The scatterplot revealed a positive correlation and no other striking patterns, and the regression coefficient was estimated at 0.6 with a standard error of 0.4. I performed the analyses in R using rstanarm (code available at `https://osf.io/nh4gc/`). and the standard error is based on the median absolute deviation of posterior simulations (see help("mad") in R for more on this).

I then looked up the article, "Relation between resting amygdalar activity and cardiovascular events: a longitudinal and cohort study", by Tawakol et al. (2017). The goal of the research is "to determine whether [the amygdala's] resting metabolic activity predicts risk of subsequent cardiovascular events." Here are some items



(a) Estimation results



(b) Testing results

Figure 4.7: van Doorn, Gronau, and Wagenmakers' Bayesian analysis of the amygdala data set. The left panel shows the result of estimating the Pearson correlation coefficient ρ under $\mathcal{H}_1$ with a two-sided uniform prior. The right panel shows the result of testing $\mathcal{H}_0 : \rho = 0$ versus the one-sided alternative hypothesis $\mathcal{H}_+ : \rho \sim U[0, 1]$. Figures from JASP.

relevant to our current task:

(a) Perceived stress is an intermediate outcome, not the ultimate goal of the study.

(b) Any correlation or predictive relation will depend on the reference population. The people in this particular study are "individuals with a history of post-traumatic stress disorder" living in the Boston area.

(c) The published article reports that "Perceived stress was associated with amygdalar activity ($r = 0.56$; $p = 0.0485$)." Performing the correlation (or, equivalently, the regression) on the log scale, the result is not statistically significant at the 5% level. This is no big deal given that I don't think that it makes sense to

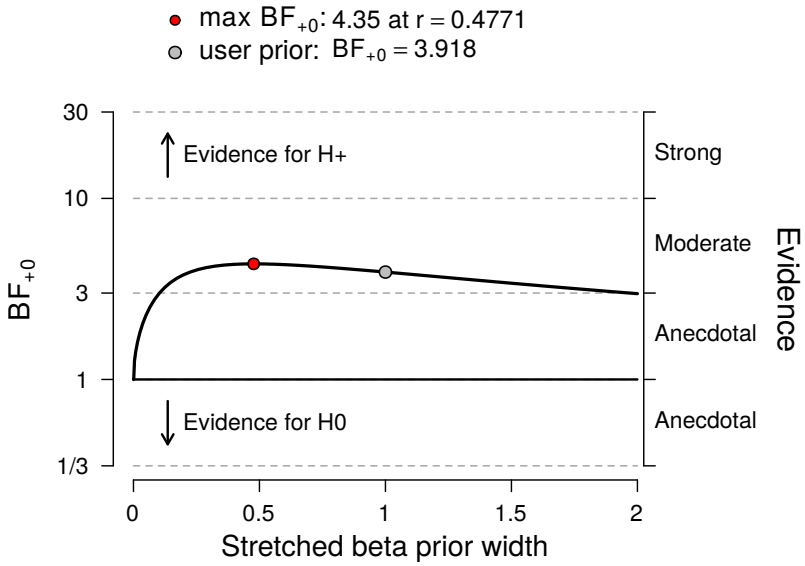


Figure 4.8: Sensitivity analysis for the Bayesian one-sided correlation test. The Bayes factor BF_{+0} is a function of the prior width parameter $1/a$ from the stretched Beta(a,a) distribution. Figure from JASP.

make decisions based on a statistical significance threshold, but it is relevant when considering scientific communication.

(d) Comparing my log-scale scatterplot to the raw-scale scatterplot (Figure 5A in the published article), I'd say that the unlogged scatterplot looks cleaner, with the points more symmetrically distributed. Indeed, based on these data alone, I'd move to an unlogged analysis–that is, the estimated correlation of 0.56 reported in the paper.

To address the question, "Do PTSD patients with high resting state amygdalar activity experience more stress?", we need two additional decisions or pieces of information. First, we must decide the population of interest; here there is a challenge in extrapolating from people with PTSD to the general population. Second, we need a prior distribution for the correlation being estimated. It is difficult for me to address either of these issues: as a statistician my contribution would be to map from assumptions to conclusions. In this case, the assumptions about the prior distribution and the assumptions about extrapolation go together, as in both cases we need some sense of how likely it is to see large correlations between the responses to a subjective stress survey and a biomeasurement such as amygdalar activity. It

could well be that there is a prior expectation of positive correlation between these two variables in the general population, but that the current data do not provide much information beyond our prior for this general question.

4.4. Round-Table Discussion

As described above, the two data sets have each been analyzed by four teams. The different approaches and conclusions are summarized in Table 5. The discussion was carried out via email and a transcript can be found online at `https://osf.io/f4z7x/`. Below we highlight and summarize the central elements of a discussion that quickly proceeded from the data analysis techniques in the concrete case to more fundamental philosophical issues. Given the relative agreement among the conclusions reached by different methodological angles, our discussion started out with the following deliberately provocative statement:

In statistics, it doesn't matter what approach is used. As long as you do conduct your analysis with care, you will invariably arrive at the same qualitative conclusion.[7]

In agreement with this claim, Hennig stated that "we all seem to have a healthy skepticism about the models that we are using. This probably contributes strongly to the fact that all our final interpretations by and large agree. Ultimately our conclusions all state that 'the data are inconclusive'. I think the important point here is that we all treat our models as tools for thinking that can only do so much, and of which the limitations need to be explored, rather than 'believing in' our relying on any specific model" (Email 29). On the other hand, Hennig wonders "whether differences between us would've been more pronounced with data that wouldn't have allowed us to sit on the fence that easily" (Email 29) and Lakens wonders "about what would have happened if the data were clearer" (Email 32). In addition, Morey points out that "none of us had anything invested in these questions, whereas almost always analyses are published by people who are most invested" (Email 33). Wagenmakers responds that "we [referring to the group that organized this study] wanted to use simple problems that would not pose immediate problems of either one of the paradigms...[and] we tried to avoid data

[7] The statement is based on Jeffreys' claims "[a]s a matter of fact I have applied my significance tests to numerous applications that have also been worked out by Fisher's, and have not yet found a disagreement in the actual decisions reached" (Jeffreys, 1939, p. 365) and "it would only be once in a blue moon that we [Fisher and Jeffreys] would disagree about the inference to be drawn in any particular case, and that in the exceptional cases we would both be a bit doubtful" (J. H. Bennett, 1990, p. 162).

sets that met Berkson's 'inter-ocular traumatic test' (when the data are so clear that the conclusion hits you right between the eyes) where we would immediately agree" (Email 31). In addition, the focus was on the analyses and discussion as free as possible from other consideration (e.g., personal investment in the questions).

However, differences between the analyses were emphasized as well. First, Morey argued that the differences (e.g., research planning, execution, analysis, etc.) between the Bayesian and frequentist approach are critically important and not easily inter-translatable (Email 2). This gave rise to an extended discussion about the frequentist procedures' dependence of the sampling protocol, which Bayesian procedures lack. While Bayesians such as Wagenmakers see this as a critical objection against the coherent use of frequentist procedures (e.g., in cases where the sampling protocol is uncertain), Hennig contends that one can still interpret the *p*-value as indicating the compatibility of the data with the null model, assuming a hypothetical sampling plan (see Email 5-11, 16-18, 23, and 24). Second, Lakens speculated that, in the cases where the approaches differ, there "might be more variation within specific approaches, than between" (Email 1). Third, Hennig pointed out that differences in prior specifications could lead to discrepancies between Bayesians (Email 4) and Homer pointed out that differences in alpha decision boundaries could lead to discrepancies between frequentists' conclusions (Email 11). Finally, Gelman disagreed with most if what had been said in the discussion thus far. Specifically, he said: "I don't think 'alpha' makes any sense, I don't think 95% intervals are generally a good idea, I don't think it's necessarily true that points in 95% interval are compatible with the data, etc etc." (Email 15).

A concrete issue concerned the equivalence test used by Lakens and Hennig for the first data set. Wagenmakers objects that it does not add relevant information to the presentation of a confidence interval (Email 12). Lakens responds that it allows to reject the hypothesis of a >10% difference in proportions at almost any alpha level, thereby avoiding reliance on default alpha levels, which are often used in a mindless way and without attention to the particular case (Email 13).

A more foundational point of contention with the two data sets and their analysis was about the question of how to formulate Bayesian priors. For these concrete cases, Hennig contends that the subject-matter information cannot be smoothly translated into prior specifications (Email 27), which is the reason why Morey and Homer choose a frequenstist approach, while Gelman considers it "hard for me to imagine how to answer the questions without thinking about subject-matter information" (Email 26).

Lakens raised the question of how much the approaches in this paper are representative of what researchers do in general (Email 43 and 44). Wagenmakers' discussion of *p*-values echoes this point. While Lakens describes *p*-values as a guide to "deciding to act in a certain way with an acceptable risk of error" and contends many scientists conform to this rationale (Email 32), Wagenmakers has a more pessimistic view. In his experience, the role of *p*-values is less epistemic than social: they are used to convince referees and editors and to suggest that the hypothesis in question is true (Email 37). Also Hennig disagrees with Lakens, but from a frequentist point of view: they should not guide binary accept/reject-decisions, they just indicate the degree to which the observed data is compatible with the model specified by the null hypothesis (Email 34).

The question of how data analysis relates to learning, inference and decision-making was also discussed regarding the merits (and problems) of Bayesian statistics. Hennig contends that there can be "some substantial gain from them [priors] only if the prior encodes some relevant information that can help the analysis. However, here we don't have such information" and the Bayesian "approach added unnecessary complexity" (Email 23). Wagenmakers reply is that "prior is not there to help or hurt an analysis: it is there because without it, the models do not make predictions; and without making predictions, the hypotheses or parameters cannot be evaluated" and that "the approach is more complex, but this is because it includes some essential ingredients that the classical analysis omits" (Email 24). In fact, he insinuates that frequentists learn from data through "Bayes by stealth": the observed *p*-values, confidence intervals and other quantities are used to update the plausibility of the models in an "inexact and intuitive" way. "Without invoking Bayes' rule (by stealth) you can't learn much from a classical analysis, at least not in any formal sense." (ibid.) According to Hennig there is more to learning than "updating epistemic probabilities of certain parameter values being true. For example I find it informative to learn that 'Model X is compatible with the data' " (Email 25). However, Wagenmakers considers Hennig's example of learning as a synonymous to observing. Though he agrees that "it is informative to know when a specific model is or is not compatible with data; to learn anything, however, by its very definition requires some form of knowledge updating" (Email 30). This discussion evolved, finally, into a general discussion about the philosophical principles and ideas underlying schools of statistical inference. Ironically, both Lakens (decision-theoretically oriented frequentism) and Gelman (falsificationist Bayesianism) claim the philosophers of science Karl Popper and Imre Lakatos, known for their ideas

of accumulating scientific progress through successive falsification attempts, as one of their primary inspirations, although they spell out their ideas in a different way (Emails 42 and 45).

Hennig and Lakens also devoted some attention to improving statistical practice and either directly or indirectly questioned the relevance of foundational debates. Concerning the above issue with using *p*-values for binary decision-making, Hennig suspects that "if Bayesian methodology would be in more widespread use, we'd see the same issue there ... and then 'reject' or 'accept' based on whether a posterior probability is larger than some mechanical cutoff" (Email 34) and "that much of the controversy about these approaches concerns naive 'mechanical' applications in which formal assumptions are taken for granted to be fulfilled in reality" (Email 29). In addition, Lakens points out that "whether you use one approach to statistics or another doesn't matter anything in practice. If my entire department would use a different approach to statistical inferences (now everyone uses frequentist hypothesis testing) it would have basically zero impact on their work. However, if they would learn how to better use statistics, and apply it in a more thoughtful manner, a lot would be gained" (Email 32). Homer provides an apt conclusion to this topic by stating "I think a lot of problems with research happen long before statistics get involved. E.g. Issues with measurement; samples and/or methods that can't answer the research question; untrained or poor observers" (Email 35).

Finally, an interesting distinction was made between a prescriptive use of statistics and a more pragmatic use of statistics. As an illustration of the latter, Hennig has a more pragmatic perspective on statistics, because a strong prescriptive view (i.e., fulfillment of modeling assumptions as a strict requirement for statistical inference) would often mean that we can't do anything in practice (Email 2). To clarify this point, he adds: "Model assumptions are never literally fulfilled so the question cannot be whether they are..., but rather whether there is information or features of the data that will mislead the methods that we want to apply" (Email 23). The former is illustrated by Homer: "I think that assumptions are critically important in statistical analysis. Statistical assumptions are the axioms which allow a flow of mathematical logic to lead to a statistical inference. There is some wiggle room when it comes to things like 'how robust is this test, which assumes normality, to skew?' but you are on far safer ground when all the assumptions are/appear to be met. I personally think that not even checking the plausibility of assumptions is inexcusable sloppiness (not that I feel anyone here suggested otherwise)" (Email 11). From what has been said in the discussion, there is general consensus that not all

assumptions need to be met and not all rules need to be strictly followed. However, there is great disagreement about which assumptions are important; which rules should be followed and how strictly; and what can be interpreted from the results when (it is uncertain if) these rules and assumptions are violated. The interesting subtleties of this topic and the discussants' views on use of statistics can be read in the online supplement (model assumptions: Emails 4, 11, 23, and 24; alpha-levels, *p*-values, Bayes factors, and decision procedures: Emails 2, 9, 11, 14, 15, 24, and 31-40; sampling plan, optional stopping, and conditioning on the data: Emails 2, 5-11, 16-18, 23, and 24).

In summary, dissimilar methods were used that resulted in similar conclusions and varying views were discussed on how statistical methods are used and should be used. At times it was a heated debate with interesting arguments from both (or more) sides. As one might expect, there was disagreement about particularities of procedures and consensus on the expectation that scientific practice would be improved by better general education on the use of statistics.

4.5. Closing Remarks

Four teams each analyzed two published data sets. Despite substantial variation in the statistical approaches employed, all teams agreed that it would be premature to draw strong conclusions from either of the data sets. Adding to this cautious attitude are concerns about the nature of the data. For instance, the first data set was observational, and the second data set may require a correction for multiplicity. In addition, for each scenario, the research teams indicated that more background information was desired; for instance, "when is the difference in birth defects considered too small to matter?"; "what are the effects for similar drugs?"; "is the correlation selective to the amygdala?"; and "what is the prior distribution for the correlation?". Unfortunately, in the routine use of statistical procedures such information is provided only rarely.

It also became evident that the analysis teams not only needed to make assumptions about the nature of the data and any relevant background knowledge, but they also needed to interpret the research question. For the first data set, for instance, the question was formulated as "Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects?". What the original researchers wanted to know, however, is whether or not cetirizine is safe – this more general question opens up the possibility of applying alternative approaches, such as the equivalence test, or even a statistical decision analysis: should pregnant women be advised not to

take cetirizine? We purposefully tried to steer clear from decision analyses because the context-dependent specification of utilities adds another layer of complexity and requires even more background knowledge than was already demanded for the present approaches. More generally, the formulation of our research questions was susceptible to multiple interpretation: as tests against a point null, as tests of direction, or as tests of effect size. The goals of a statistical analysis can be many, and it is important to define them unambiguously – again, the routine use of statistical procedures almost never conveys this information.

Despite the (unfortunately near-universal) ambiguity about the nature of the data, the background knowledge, and the research question, each analysis team added valuable insights and ideas. This reinforces the idea that a careful statistical analysis, even for the simplest of scenarios, requires more than a mechanical application of a set of rules; a careful analysis is a process that involves both skepticism and creativity. Perhaps popular opinion is correct, and statistics is difficult. On the other hand, despite employing widely different approaches, all teams nevertheless arrived at a similar conclusion. This tentatively supports the Fisher-Jeffreys conjecture that, regardless of the statistical framework in which they operate, careful analysts will often come to similar conclusions.

Table 4.5: *Overview of the approaches and results of the research teams.*

Research Team	Data Set I: Cetirizine & Birth Defects	Data Set II:Amygdalar Activity
Lakens & Hennig	– Frequentist test of equivalence – 10% equivalence region – Data deemed inconclusive	– Frequentist correlation, $p = .047$ – Concerns about multiple comparisons – Data deemed inconclusive
Morey & Homer	– Frequentist logistic model, $p = .287$ – Observational, so possible confounds – Data deemed inconclusive	– Frequentist correlation, $p = .047$ – Small n means assumptions unverifiable – Is the effect specific for amygdala? – Data deemed inconclusive
Gronau, Van Doorn, & Wagenmakers	– Default Bayes factor $BF_{01} = 1.6$ – Evidence "not worth more than a bare mention" – Data deemed inconclusive	– Default Bayes factor $BF_{10} = 2$ – Small n – Data deemed inconclusive
Gelman	– Bayesian analysis needs good prior – Data likely to be inconclusive – A key question is who takes the drug in the population	– Bayesian analysis needs good prior – Problems with generalizing to population – Data likely to be inconclusive

5

OBJECTIVITY

Abstract:
In the last decade, many problematic cases of scientific conduct have been diagnosed; some of which involve outright fraud (e.g., Stapel, 2012) others are more subtle (e.g., supposed evidence of extrasensory perception; Bem, 2011). These and similar problems can be interpreted as caused by lack of scientific objectivity. The current philosophical theories of objectivity do not provide scientists with conceptualizations that can be effectively put into practice in remedying these issues. We propose a novel way of thinking about objectivity for individual scientists; a negative and dynamic approach. We provide a philosophical conceptualization of objectivity that is informed by empirical research. In particular, it is our intention to take the first steps in providing an empirically and methodologically informed inventory of factors (e.g., Simmons, Nelson & Simonsohn, 2011) that impair the scientific practice. The inventory will be compiled into a negative conceptualization (i.e., what is not objective), which could in principle be used by individual scientists to assess (deviations from) objectivity of scientific practice. We propose a preliminary outline of a usable and testable instrument for indicating the objectivity of her scientific practice.

This chapter is adapted from van Dongen, N. N.N., & Sikorski, M. (under review). Objectivity for the Research Worker. *European Journal for Philosophy of Science*.
Both authors contributed equally.

5.1. Introduction

Despite its undeniable success (e.g., electricity, space flights, etc.), science seems to be in a difficult position today. In the last few years, many problematic cases of scientific conduct were diagnosed, some of which involve outright fraud (e.g., Stapel, 2012) while others are more subtle (e.g., supposed evidence for precognition; Bem, 2011). These particular issues and the general lack of replicability of scientific findings (e.g., Open Science Collaboration, 2015) have contributed to what has become known as the *replication crisis* (e.g., Harris, 2017). In addition, the general public has become aware of these problems, which has shaken the general trust in science (e.g., Anvari & Lakens, 2018; Lilienfeld, 2012; Pashler & Wagenmakers, 2012).

Let us imagine a scientist, Dr. Jane Summers. Dr. Summers[1] does research in cognitive psychology. One day, she reads about the low replicablity rate of the results of psychological studies (e.g., Open Science Collaboration, 2015). She becomes very concerned about the value of scientific results in general and her own research in particular. As a result, she is resolved to investigate to what extent her work is at risk of irreplicability and to ensure that her current and future work is as resilient as possible against such a fate. She decides that, apart from ensuring the accuracy and precision of her measurements, the methods she employs should not be significantly influenced by her feelings, values, biases and other idiosyncrasies. To her, this would mean that they are objective (Hawkins & Nosek, 2012; Stegenga, 2011; Ziman, 1996).[2] Objectivity can be attributed, among others, to scientific measurements, tools for development/improvement of scientific theories, and/or to true-to-nature explanations. It ensures that study outcomes are not biased (e.g., over estimation of drug efficacy, under estimation of risk; Goldacre, 2014), positive research results are not false-positives (to a larger proportion than is allowed by the statistical method; Simmons et al., 2011), and are independently reproducible by other scientists (Altmejd et al., 2019; Lindsay, 2015; Simons, 2014; van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). Dr. Summers considers objectivity to be essential to science[3] and its absence to be a cause of the crisis that threatens the foundations of her research field. In short, Dr. Summers considers the assessment and safeguarding of scientific objectivity as being of vital importance. Plausibly, such sentiment toward

[1] Dr. Summers is a fictional character, though we postulate that she is representative of current practitioners who work on the improvement of research practice and her considerations of potential problems result from a synthesis of methods textbooks that have recently fallen from grace for recommending questionable practice.

[2] A similar description of objectivity can be found on Wikipedia (2019).

[3] Interestingly, there are not many references on the importance of objectivity for science. Scientists we spoke to consider its relevance obvious and self-explanatory to such an extent that it does not warrant explicit explanation and justification.

objectivity is common among actual scientists. For example, we can easily imagine Prof. Bem wanting to present results as solid and close to incontrovertible evidence in favor of precognition as possible (Bem, 2011). Specifically, ensuring objectivity of his experiments would ensure that his claims are on solid ground.

It is therefore somewhat puzzling to Dr. Summers that a proper explication of objectivity appears to be lacking in science. She is unable to find tools for the qualitative and/or quantitative assessment of objectivity. Methodological reforms are inspired by problematic cases, for instance, measurement or impossible results (e.g., precognition; C. M. Bennett, Baird, Miller, & Wolford, 2011) or failures to reproduce established experimental results (e.g., Klein et al., 2018), rather than a clear understanding of objectivity. She could attempt to replicate her own work, have it replicated by others, and/or review her publication with respect to guidelines of statistical methods (see for instance Carney, 2016; Gervais, 2017) and, as a result, declare a lack of confidence in her own work (see for instance Rohrer et al., 2018), but nothing more systematic is available. Similarly, Prof. Bem would have a hard time providing an objectivity assessment of his precognition experiments with the currently available tools. Thus, Dr. Summers realizes that science could greatly benefit from having a definition of 'objectivity' that can be explicated in a quantitative or qualitative assessment of scientific practice.

Dr. Summers has a hunch that philosophy might be of assistance in defining objectivity. After a short review of the philosophical literature, she does not manage to find a notion of objectivity ready for use in scientific practice. Typically, such proposals are descriptive and therefore without guiding force, because they are not supported by normative considerations. Other proposals are difficult or impossible to test, thus prohibiting scientists from assessing objectivity (Section 5.2). In effect, Dr. Summers becomes disheartened and contemplates quitting her quest for objectivity.

It is our opinion that we, philosophers, should not disappoint scientists like Dr. Summers in this respect and that philosophy can and should do better. We believe that the philosophical literature currently lacks a scientifically useful conceptualization of objectivity and we intend to fill this gap. In this manuscript, we present a conceptualization of objectivity of scientific practice that is practicable by the individual scientist. We understand scientific practice as pertaining to empirical research, which include all activities done by scientist essential for this endeavor. These include study design, data collection and measurement, data analysis, result reporting etc.[4] We recognize that the social and cultural conditions play a role in, for

[4]For reasons clarified in Sections 5.3 and 5.4, we restrict our definition to research that works with non-qualitative (quantitative or countable) data.

instance, determining what kind of research gets funded, and recognize the value of social epistemology and literature on non-epistemic values (e.g., Biddle, 2007; Bueter, 2015; Elliott & McKaughan, 2009; Longino, 1990). However, much of this is beyond the purview of what an individual researcher can control and therefore beyond the scope of our paper.[5]

Our aim is to provide a scientifically useful notion of objectivity. In order to be useful such a conceptualization must be both based on normative considerations and testable. For if it is not based on normative and reliable methodological results it is not clear if it possesses any guiding force and if it is not testable, it cannot be used to assess the objectivity of a given practice. In the next section, we will briefly discuss several philosophical views on objectivity and highlight where there is room for improvement. In the third section, we present a novel version of a negative approach to scientific objectivity and provide a testable conceptualization of objectivity that is based on robust empirical results and methodological considerations. Finally, we defend the fruitfulness of our notion by demonstrating how it can be used in scientific practice (Section 5.4.2) and provide a sketch of a tool for assessing objectivity inspired by our new conceptualization (Appendix C).

5.2. Philosophy on Objectivity

In philosophy of science, scientific objectivity is a well discussed notion. Following Reiss and Sprenger (2014), we can list three main ways of conceptualizing it. Firstly, objectivity can be understood as a faithfulness to facts. Secondly, something can be understood as objective when it is free from value commitments. Thirdly, objectivity can be understood as being free from scientists' personal biases. Recently, proposals which have gained much popularity are pluralist notions of objectivity (e.g., Douglas, 2004; Megill, 1994; Wright, 2018). Such notions encompass some or all of mentioned individual notions (e.g., the value-free objectivity, value neutral objectivity, procedural objectivity etc.). Finally, there are negative conceptions of objectivity (e.g., Daston & Galison, 2007; Hacking, 2015; Koskinen, 2018) which claim that the objectivity consist of the absence of certain factors. In case of Daston and Galison (2007), these are factors of scientific subjectivity which are recognized by the scientific community as particularly troubling or important in a given time period. In the case of Koskinen (2018), these factors are epistemic risks which arise from the imperfections of epistemic agents.

Despite this effort, it seems that a conceptualization of scientific objectivity that

[5] We address this limitation in more detail in Section 5.4.

can be easily used by scientists has not yet been proposed. The literature is comprised of proposals that were not designed to fulfill such a practical role. Instead, they were designed to describe how the concept is used. Following Searle (1975) we will understand the difference between descriptive and normative discourse in terms of the direction of fit. The descriptive claims aim at describing reality (e.g., 'there is no poverty in the world'). In contrast, normative claims are not descriptions of how things are, they are intended to describe how the word should be (e.g, 'there should not be poverty in the world'). In other words descriptive claims have *language-to-reality direction of fit* while normative one *reality-to-language direction of fit*. Consequently, a descriptive theory of a given concept describes in precise terms (the meaning of) the concept, which is actually used by natural language speakers (or some sub-group of them). Such theories can be assessed empirically by comparing it with the intuitions of a target group. The examples of such theories are the semantics for conditionals (see e.g., Douven, Elqayam, Singmann, & van Wijnbergen-Huitink, in press). On the other hand, a normative theory of a given concept presents (a meaning of) the concept which, when used, will be beneficial for the hypothetical users. For example, some of the formal theories on truth present replacements for the concept of 'truth' used in natural language (see e.g., Scharp, 2013; Tarski, 1936). Authors of these proposals argue that new concepts are superior to the concept present in natural language, because, for example, they are not susceptible to notorious semantic paradoxes. In our article, we are interested in a normative theory of objectivity. Hence, we are less concerned with how the new conceptualization corresponds to how objectivity is used in natural language and more concerned with how the conceptualization promotes the methodological quality of science (e.g., replicability, lack of bias, etc.) and its results (e.g., approximately true / highly corroborated theories, theories with high predictive accuracy).

In light of the conflicting intuitions and conceptual confusion surrounding objectivity, the descriptive conceptualizations of objectivity are clearly useful. However, it is distinct from our aim of formalizing a notion that is normatively useful. Due to their descriptive aim, it is not clear if these theories can fulfill the normative task of guiding scientific practice and it would not be fair to assess them in this context. Some authors are explicit about the descriptive nature of their proposals. For example, the aims of Heather Douglas (2004) famous article seems to be primarily[6]

[6]Despite this descriptive goal, there are some normative remarks in the paper. For example, Douglas refers to her earlier work (Douglas, 2000) in which she argued that one of the conceptualizations of objectivity, value-free objectivity, is impossible to realize in practice. However, this does not detract from the main aim.

descriptive:

> In this paper, I will lay out a complex mapping of the senses of objectivity. This mapping will make two contributions to current discussions. First, it will dissect objectivity along operationally distinct modes.[...] Second, the mapping will allow me to cogently argue that the different meanings of objectivity I explore here are not logically reducible to one core meaning. (Douglas, 2004, p. 454-455)

Similarly, Koskinen (2018) is explicit about the descriptive aim of her proposal:

> In this article I defend a risk account of scientific objectivity. The account is meant to be a largely descriptive or even a semantic one; my aim is to draw together ideas presented in recent discussions, and to clarify what we philosophers of science do when we identify distinct, applicable senses of objectivity or call something objective. (Koskinen, 2018, p. 1)

These quotes indicate that Douglas (2004) collects applicable notions of objectivity (*procedural objectivity*, *value free objectivity*, etc.) while Koskinen unifies those distinct meanings. Their aims are descriptive. In the case of other proposals, it is clear that they are descriptive due to their methodological approach. For instance, Datson and Galison's (2007) historical methodology makes it a descriptive proposal.

Secondly, some of the proposed normative conceptualizations of objectivity are not suitable to be used by scientists. Such notions need to be testable. Otherwise, how can we assess if given scientific practice is objective or not? An example of a notion that fails in this respect is *value free objectivity*. Value-free objectivity is based on a more general *value-free ideal*. The value-free ideal claims that scientists should not use their non-epistemic values, like 'equality' of 'fairness', when they justify their claims (e.g., Betz, 2013). This conception of objectivity claims that a scientific justification is objective as long as it is not influenced by non-epistemic values. There might be reasons to believe that value-free ideal should be followed (e.g., Betz, 2013; Sober, 2007) or that the corresponding notion of objectivity is compelling. However, many problems of value-free objectivity have been diagnosed. For instance, Douglas (2004), after Rudner (1953), argued that value free-ideal is unrealizable. Similarly, Longino (1996) claims that the distinction between epistemic and non-epistemic values, on which value-free objectivity is based, is ill-defined, making this conceptualization of objectivity problematic. Additionally, there are clear difficulties in assessing the value-free objectivity of scientific practice. Most glaringly, we do not have access to scientists' intentions, thus we cannot judge what motivated their decisions and actions. Therefore, we cannot use a theory which defines objectivity in terms of values used by scientist to, for example, assess the

objectivity of a procedure or research result. In short, the rich and fruitful discussion concerning the role of values in science (see e.g., M. Brown, 2013; Douglas, 2009; Hicks, 2014; Longino, 1990; Steel, 2010) and other notions of objectivity inspired by it (see e.g., Douglas, 2009; Longino, 2004) are not directly applicable to our problem.

A detailed discussion of the practical usability of other notions of objectivity presented in literature is beyond the scope of our paper. However, we expect that this cursory sketch provides an overview of the problems with using these notions and motivates the value of a new conceptualization of objectivity of scientific practice.

5.3. Problems in Science and the Via-Negativa approach to Objectivity

There is no generally accepted positive definition of 'health' in health care and the medical sciences.[7] Fortunately, this does not prevent doctors from healing ailments and researchers from developing new drugs and technologies. A positive definition of health is unnecessary, when the instances that reduce or endanger health can be defined and addressed. In brief, health is what remains when the particular infirmities are removed.[8] Health care and medical science appears to be successful, even in the face of changing definitions, diagnostics, and disagreements about ailments.[9] We believe that this via-negativa approach can also be applied to the concept of scientific objectivity.

Our negative approach resembles other negative proposals in philosophy (e.g., Daston & Galison, 2007; Koskinen, 2018). Just like these approaches, we conceptualize objectivity as what remains in the absence of certain factors, however, our aim and identified factors are different. Specifically, the purpose of our notion is to be testable and practicable by scientists. Hence, we base our conceptualization on empirical research. We postulate that non-objectivity consists of factors that have been empirically or methodologically identified as making scientific practice susceptible to the actions and decisions of scientist which can inadvertently or intentionally influence research results. Such practices have the propensity to reduce

[7] The World Health Organization defines health as 'a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.' (Organization et al., 1950, p.100). Which is essentially a negative definition augmented with an well-being requirement. The definition is considered controversial and with little added benefit over the original negative definition (e.g., Jadad & O'Grady, 2008). Adding 'well-being' just kicks the can down the road.

[8] In this paper, the negative definition of 'health' is used as an analogy to clarify our approach. Our conceptualization has no stake in this definition or its controversies.

[9] Some might argue that medical science is not as successful as it purports to be and suffers from diminishing returns (Stegenga, 2018). However, this does not detract from the fact that medical science is successful to at least some degree and the mentioned problems do not necessarily result from a lack of a positive definition of health (see for instance Firestein, 2015).

reliability, validity, and replication rates of the results (e.g., Simmons et al., 2011). We assert that the factors constituting non-objectivity translates to a conceptualization of objectivity, which not only preserves some of our intuitions about objectivity but also and more importantly, can be put into practice by scientists (Section 5.4.2).

The general sense of how the objectivity of scientific practice can be compromised is as follows. Researchers make certain decisions when they design their study and collect, process, and analyze their data. The possibility of choosing between two or more options in these instances are called *researchers' degrees of freedom* (Simmons et al., 2011; Wicherts et al., 2016). The misuse of which can result in biased[10] and/or irreproducible outcomes. Kinds of such misuse are identified, for instance, by questioning scientists on their behavior and the behavior of others (e.g., John et al., 2012; Kerr, 1998) or case studies (e.g., Schimmack, 2020) in comparison to data simulation or principles of statistical analysis. As an example of misuse identification, Schimmack (2020) reanalyzed the data of Bem's *Feeling the Future* experiments (Bem, 2011) and uncovered that effects of precognition were very large if only the data of the first few included participants were analyzed, but decreased raptly to just above the statistical significance threshold towards the end of participant inclusion. Schimmack showed that this pattern is produced by starting many studies on a non-existing effect and discontinuing all but those that show 'promise' (i.e., an initial strong positive effect).

The ways in which scientists can misuse this freedom can be grouped into two categories. Firstly, a scientist can make a priori decisions concerning the research design and data collection, which can preclude certain outcomes or make them more/less likely (i.e. introduce systematic bias in a certain direction). Secondly, a scientist has to make decisions on how to process and analyze the data, which allows her to try all possible combinations of decisions until a positive/desired result is found. In this section, we first focus on problematic practices taking place before or during conducting the study (Section 5.3.1). Then, we discuss problematic data management and analysis practices (Section 5.3.2). The section is concluded with a testable conceptualization of objectivity as resilience to such problematic practices.

[10]For clarification, bias, as discussed in this paper, is used to indicate systematic error introduced by behavior of the scientist (e.g., Ioannidis, 2005). This is different from bias in psychology (e.g., MacCoun, 1998) where it is used to classify cognitive heuristics (e.g., confirmation bias, bandwagon effect, anchoring, etc.). These heuristics might indirectly influence results, but these distant causes are irrelevant to our approach.

5.3.1. Problems before and during Research: Design, Data collection, and Measurements

During the early stages of a scientific experiment (e.g., designing the observational study or experiment, sampling, measurement, etc.), a scientist has to make several decisions, which could influence the final result. In some cases, a scientist might make such choices with the aim of obtaining a specific result in mind. Such decisions introduce bias (e.g., Fanelli, Costas, & Ioannidis, 2017).

In science, biased research seems to results from the influence of beliefs or prejudices of the scientist on her methodological decisions. For example, scientists make methodological choices that increase the likelihood of getting results that align with the preferences of those that provide the research funds (this is known as 'funding bias' Jones & Sugden, 2001; J. A. Nelson, 2014). Similarly, a scientist can adjust the design of her experiment or observational study, consciously or subconsciously, in order to increase the probability that the results will support her prior beliefs. Typically, biased outcomes(s) only require(s) a single decision or a small number of decisions during the experiment design phase of the research. Taking pharmaceutical research as an illustration, positive results for tested medication is boosted by, for instance, selecting only unrepresentative 'ideal' patients, comparing the drug to an ineffective alternative, or using different effective doses for the treatment and control group (e.g., Rothwell, 2005; Safer, 2002; Travers et al., 2007). As another example, the biased studies presented in Wilholt (2008) involve scientists choosing a specific strain of experimental animals, which made the experiments significantly less likely to show the toxicity of the tested substance (in line with the preference of the funding institution). Next to sample selection, bias can be introduced through many of the other decision that a scientist has to make when designing and conducting research, specifically:

1. Which measurement (outcome measure) to use?
2. Which kind of independent variable (experimental manipulation) to use?
3. Which sample to select and how?
4. Setting of the experiment or observational study (when and where)?
5. How and to what extent do researcher and research subject interact?
6. How to perform the measurement (e.g., blinded or unblinded)?

Recognition of features that can introduce bias is reflected in proposals concerning how to counter it. For example, Wilholt (2008) proposed establishing conventions which regulate the way scientist should conduct their studies as a remedy to funding bias. In the case of choosing insensitive animals, he proposed to adopt the following convention:

> Because of clear species and strain differences in sensitivity, animal model selection should be based on responsiveness to endocrine active agents of concern (i.e. responsive to positive controls), not on convenience and familiarity (US Department of Health and Human Services, 2001, p.vii).

Different conventions are and can be implemented in order to impose methodological restrictions on scientists. Some of them force scientists to measure the direct outcome of interest instead of a proxy, use standardized tests or measurements, use random sampling from the population, use random allocations of participants to conditions, use equal group treatment, use blind or double blind design (experimental studies), and/or use data collectors that are blind to the research aim (observational studies). All of these conventions restrict the range of biasing decisions a scientist can make. In addition, these conventions can be empirically tested with respect to prohibiting potentially biasing actions by the scientists and reducing bias in research outcomes.

5.3.2. Problems After Experiments or Observations: Data Management, Analysis Specification, and Result Reporting

After a researcher has run the experiment and the data has been collected, several decisions have to be made. For instance, the data needs to be processed (e.g., removing outliers, combining variables, binning variable values, etc.), the statistical model needs to be specified (e.g., linear model, multilevel model, structural equation model, etc.), and finally the dependent and predictor variables for the model need to be selected. The assumption is that for each step only one (and the most appropriate) of the possible options is selected. However, research has shown that the general rate of false-positive results[11] is increased when, instead of taking a single option for each step, several possible combinations of options are explored and only the combinations that culminate in positive results are reported (e.g., John et al., 2012; Simmons et al., 2011; Szucs, 2016; Wicherts et al., 2016). These behind-the-scenes practices that covertly influence results go by the name of *questionable research practices*. The causes of these practices may be the scientists' (sub)conscious beliefs or preferences, the ambiguity or ignorance about how the methods works and what

[11] No matter how precise an instrument is and no matter how stringent the evidence requirements are, there is always a non-zero probability that the result of a study does not reflect reality (e.g., positive outcome of a HIV test when the person is actually HIV negative). Statistical methods of analysis come with certain rules and assumptions, which must be followed in order for this probability to have a known maximum. In other words, if a study is performed according to its rules, none of the assumptions are violated, and it is repeated a large number of times, the proportion of false-positive results (i.e., an effect is observed while actually no effect exists) is at most equal to this probability, which is or can be known.

the statistics are/mean, or the desire to find/see associations and structure in what is being studied. Concretely, at least the following decisions need to be made by a researchers when dealing with quantitative data and performing statistical analyses (this incomplete list is adapted from: Bakker, van Dijk, & Wicherts, 2012; Kass et al., 2016; L. D. Nelson, Simmons, & Simonsohn, 2018; Simmons et al., 2011; Wicherts et al., 2016):

1. How to handle incomplete or missing data?
2. How to pre-processes data (e.g., cleaning, normalizing, etc.)?
3. How to process data, deal with violations of statistical assumptions (e.g., normality, homoscedasticity, etc.)?
4. How to deal with outliers?
5. Which measured construct to select as primary outcome?
6. Which variable to select as dependent variable out of several that measure the same construct?
7. How to score, bin, recode the chosen dependent variable?
8. Which variables to select as predictors out of the set of measured variables?
9. How to recode or restructure these predictors (e.g., combining variables, combining levels of a variable, etc.)?
10. If and which variables to additionally include as covariates, mediators, or moderators?
11. Which statistical model to use?
12. Which estimation method and computation of standard errors to use?
13. If and which correction for multiple testing to use?
14. Which inference criteria to use (e.g., p-values and alpha level, Bayes factor, etc.)?

Note that, if such decisions needs to be made and how many option the scientists has to choose from depends on how the study was designed and the structure and size of the data that was collected.

Currently, there are already some potential strategies for restricting uses of questionable research practices (i.e., ad hoc decision making in order to get positive results, also known as p-hacking). For instance, a) preregistration of the study from design to analysis (e.g., Chambers, 2013; Nosek et al., 2018; Wicherts et al., 2016); b) data and analysis blinding (e.g., MacCoun & Perlmutter, 2015); and c) run several/all of the (theoretically) possible tests in a *multiverse analysis* (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016).[12] The effectiveness of these strategies

[12]In-depth discussion of these strategies is beyond the scope of this paper. Function, benefits, and limitations of these strategies can be found in the cited papers.

can be empirically verified by researching their effects in, for instance, replication studies. It should be noted that such strategies are not mutually exclusive and that combinations are possible, because they all restrict researcher's degrees of freedom without introducing new ones. For instance, not all decisions can be made in advance, precluding their preregistration. In such a case, some of these can be caught by data blinding, because the scientist might not know what the data will look like in advance, though has an analysis plan that can be communicated to the independent data analysis. In addition, the multiverse analysis can be employed for those elements of the research that have an exploratory nature that do not allow for data blinding and handing the analysis to someone else.

5.3.3. To Sum up: a conceptualization of objectivity

Our negative version of conceptualizing objectivity ties it to scientific problems that result from the decisions and actions of individual scientists. These problems are notoriously hard to detect. For instance, a report of a study during which questionable research practices were used can be indistinguishable from a report of a study where scientists actively tried to avoid influencing the results. If objectivity is just absence of these problems, then testing it is extremely difficult to impossible. On the other hand, we can easily tell if precautions against such problems (e.g., preregistration) are present and thus how resilient a given practice is. Therefore, we state that a scientific practice becomes more objective when it becomes more resilient to actions and decisions that have the potential to influence its outcome; concretely, when:

a) the study design and data collection becomes more resilient to the scientists' influence on the data;
b) and the data processing and analysis become more resilient to ad hoc decision making and selective reporting of positive results.

In the limit, a practice is objective when it is impervious to biasing influences and precludes ad hoc decisions and actions.

Our approach has two clear advantages. 1) It is empirically verifiable. 2) It does not require universally agreement about factors that reduce objectivity nor does the procedure for identifying these factors need to be objective. Our notion, in opposition to traditional conceptualization (e.g., value-free objectivity), ties objectivity to features of scientific practice, the existence of which can be empirically tested (e.g., was the study preregistered or not).[13] These features can, for instance, be collected in a

[13]Note that it is not the testing of whether or not certain potentially biasing actions were made

form of a checklist (see the Appendix for a first setup). Concretely, such a checklist could in principle be used by reviewers to evaluate submitted manuscripts on the precautions taken against biasing influences; or by readers of published papers who want to assess their trustworthiness; or by reviewers (writers) of grant applications to evaluate (show) that future results will be as insulated as possible against biasing effects.[14] In addition, objectivity according to this conceptualization can be verified by assessing the extent of systematic bias and inflated false-positive rates in a body of literature. The presence of objectivity promoting features like preregistration decreases the chance of a given study being a false positive. Therefore we can indirectly test the objectivity of studies, for instance, by testing consistency of results between preregistered experiments in comparison to consistency of results between non-preregistered experiments.

The second advantage follows from the first. We do not claim that the list of objectivity reducing factors on which our conceptualization is based is exhaustive. Moreover, some factors might be considered controversial as objectivity reducing or it may not be objective how factors are included, while other are not. This is not problematic for our proposal, because a) the identification and inclusion of factors is based on robust empirical results and methodological considerations; and b) their impact on the quality of the study, as explained in the previous paragraph, can be empirically verified.

5.4. Discussion

In this paper, we have offered a novel and practicable conceptualization of scientific objectivity. We have argued that many of the popular philosophical attempts at defining objectivity are not practicable and are likely to be impossible to implement by individual scientists. As we have argued, some of the theories aim at reconstructing the way the philosophers or scientists understand objectivity rather than proposing a normatively compelling notion. Secondly, some of the normative proposals define objectivity in terms of features that are prohibitively difficult to test empirically and hence use in practice. For example, testing conceptualizations that define objectivity in terms of the intentions of scientists, like value-free objectivity, would require real-time access to the mind of scientists during research.

during the research, but what precautions were in place to preclude such actions.

[14] Similar checklists have been developed and are in wide use as tools for assessing methodological quality of studies (e.g., Downs & Black, 1998b; Sindhu, Carpenter, & Seers, 1997) when, for instance, appraised for inclusion into a systematic review (e.g., Haidich, 2010). Recently, a checklist to assess scientific transparency published (Aczel et al., 2020) that awaits application.

In our approach, we have used findings from empirical research and methodological considerations to identify features of scientific practice considered to be problematic (i.e., potential causes of bias and inflated false-positive rates). We postulate that resilience to these features constitute objectivity. Given these features, scientific practice approaches objectivity when it becomes less vulnerable to decisions and actions of scientists that can influence its outcome.

In this section, we discuss the limitations and implications of our conceptualization. In the appendix, we present a draft for a tool that can be used to assess the objectivity of scientific endeavours (e.g., published papers, submitted manuscripts, proposed research in grant applications, etc.). In addition, we suggest investigations into a tool such as ours to test and improve its validity and reliability. We close this paper with a detailed illustration of how such a tool could be usefully implemented.

5.4.1. Limitations

Incompleteness. Plausibly, in our paper we do not reach a complete list of ways in which scientific practice can be compromised. Therefore, it is most likely that we did not reach a complete definition of objectivity, though rather a number of currently identified necessary conditions. However, our approach does provide a framework for learning from empirical research and methodological developments when, where, and how particular factors compromise scientific objectivity. Even with this limitation, we believe that our conceptualization is an improvement over previous attempts of conceptualizing objectivity and can still be used in a fruitful way (Sections 5.3.3 and 5.4.2).

Ritualization. Some might argue that restricting researchers in the proposed way will actually reduce objectivity. For instance, the (faulty) use of the Null Hypothesis Significance Testing procedure (NHST) has been described as a restrictive ritual; a practice that discourages informed reasoning and prescribes certain actions and decisions. The NHST ritual has been considered to be the main cause of the inflated number of false-positive results in science (Gigerenzer, 2004; Ioannidis, 2005; Stark & Saltelli, 2018), which is the opposite of what an objective method should achieve. However, the NHST ritual only appears to restrict researchers and provides just the illusion of objectivity. In particular, apart from inference criterion (i.e., an observed statistic lower than a conventional threshold), this ritual does not restrict (mis)use of degrees of freedom (mentioned in Section 5.3.2) at any point during the research process. Specifically, and in contrast to recommendations of our proposal, ad hoc decision-making in data management, analysis, and result reporting are

not prohibited in the NHST ritual. It might even be considered that this partial formalization enshrines a false sense of objectivity that is actually harmful to the quality of scientific results (e.g., Gigerenzer, 2004; Simmons et al., 2011). In other words, if the ritual had been restrictive in ruling out questionable research practices, it would actually promote objectivity. Our conceptualization does recommend these additional restrictions. Also, in contrast to the conservative nature of a ritual, our conceptualization is (meant to be) adaptive; developed in accordance with novel discoveries concerning problematic scientific practices and methodological changes in science.

Restricted. Our conceptualization is restricted to practice of quantifiable or countable research, which precludes qualitative research and non-empirical practices. Qualitative research is currently omitted from our definition, because, to our knowledge, empirical research and methodological considerations on the particulars of systematic bias and false-positive rate inflation in the use of qualitative methods are currently absent in the academic literature. It remains an open question if our or a analogous notion can be applied to qualitative research.

Scientist-independent problems. In some cases, a source of negative influence on research results is independent of the decisions of a scientist (e.g., Biddle, 2007; Bueter, 2015; Harding, 2015; Leuschner, 2012; Longino, 1990). For example, a scientist may be restricted in access to particular instruments, samples, or treatments of research subjects for external reasons (e.g. ethical, political, financial, practical, etc.). Therefore, the results can be compromised, though not because the scientist misused degrees of freedom. It is also possible that some internal features of a research field or used methodology cause the results to be systematically biased. In such a case, the culture and conventions of a particular area of research may restrict individual scientists to particular measurement instruments and research subjects, which could produce spurious and biased findings. For instance, culture and politics can influence which research projects get funded and thus carried out (e.g., Bueter, 2015; Elliott & McKaughan, 2009). These factors might also influence which research results get published (i.e., publication bias). Specifically, at the moment it seems that most scientific journals prefer to publish articles describing experiments with positive results and/or scientists submit only positive results to these journals. This bias against negative results precludes some research from entering the scientific literature, which inflates the rate of published false-positive results. Consequently, even if the scientific practice of each individual scientist is (as) objective (as possible), the false-positive rate will still be inflated to an unknown degree. Publication bias

(e.g., Malički & Marušić, 2014) and other similar scientist-independent problems (e.g., Biddle, 2007; Leuschner, 2012) are discussed extensively in the literature and some solutions were proposed (see e.g., H. A. Carroll, Toumpakari, Johnson, & Betts, 2017; Harding, 2015; Longino, 1990). These problems are larger than the individual scientists and thus the proposed solutions typically involve changing the social arrangement of science rather than practices and procedures used by individual scientists. For example, Biddle (2007) proposes to implement a system of institutionalized criticism to counter the corrupting effect of financial stakes on the integrity of research, another major scientist-independent problem. However, the two types of problems are distinct and therefore they require different solutions. Misuse of degrees of freedom requires the improvement in objectivity as understood in a way we have described above. On the other hand, external limitation requires improvement in the social structure of science and possibly general improvements in scientific methodology. Thus, we acknowledge the existence of these social, cultural, political, and technical problems and are in favor of programs addressing these issues. Additionally, the solutions on both social and individual level problems are complementary and might be combined into a more complete proposal.

Exploratory research and serendipitous discoveries. Many (if not most of the) famous scientific breakthroughs have been serendipitous discoveries. These discoveries were most likely the product of exploratory research that were neither done by unbiased scientist nor completely free from practices that would now be labeled as 'questionable'. It should be noted that we do not object to these practices and even see them as a vital part of science. However, when it comes to verifying these findings and integrating them in the rest of science, we firmly believe that these discoveries should be tested with a practice that is as objective as possible.

Too demanding. Clearly, our conceptualization is very exacting. Not many or maybe even no scientific practice, past or present, is objective in this sense. This is a criticism that has also been leveled at procedural objectivity (Jukola, 2017).[15] Be that as it may, this does not prevent our notion from being useful. As we will demonstrate in the next section (Section 5.4.2) we can compare the relative objectivity of two methods even if neither of them is fully objective according to our conceptualization. Secondly, the notions give us a clear idea of which modifications of a given practice increase its objectivity. In light of this, we believe that the usefulness of our conceptualization is not impaired because it is hard to satisfy. It is

[15]Procedural objectivity is the claim that there is objectivity when different scientists using the same procedure get the same/similar results (e.g., Porter, 1995). This is one of the notions included in Douglas' (2004) pluralist account of objectivity.

something to strive for, not necessarily something to reach.

Objective research does not guarantee true nor trustworthy results. Even if the work of a scientist did not suffer from anything that could jeopardize the research's objectivity, it is still possible that the results are not true (i.e., do not reflect or represent reality). It could be as innocent as a false-positive or it might be that the measurement instrument is not adequate for investigating the phenomenon at hand. Either way, we should be clear that objectivity of a practice cannot be equated with scientific truth generation. Similarly, even when scientific practice is (as close to) objective (as possible), maybe of a low reliability (i.e., noisy measurement) or lacks validity (i.e., does not measure what it is supposed to measure). Validity and reliability might be necessary to guarantee the quality and trustworthiness of results. Furthermore, the possibility of trustworthy results without *procedural objectivity* has been leveled as a criticism against this type of objectivity (Jukola, 2017). However, according to our conceptualization, perfect/high reliability, validity and thus trustworthiness are neither necessary nor sufficient conditions for the objectivity of the scientific practice that produced the results. That being said, we should still care about objectivity, because validity and reliability are promoted by it (Section 5.4.2).

5.4.2. Implications and Applications

The primary advantage of our approach to objectivity is that, in contrast to traditional theories of objectivity, it can be applied in science. For instance, our notion can be used to assess and address currently salient problems in science (i.e., the replication crisis: Harris, 2017) and evaluate suggested solutions to problematic scientific practices. Concretely, our conceptualization of objectivity can be captured in an tool that can be tested and calibrated (for an example, see Appendix C).

Increasing objectivity of scientific methods is a necessary step in remedying problems, such as the replication crisis (e.g., Harris, 2017). This crisis is constituted by the fact that results from many scientific experiments are not reproduced in replication studies (for a discussion see: Open Science Collaboration, 2015; Romero, 2016). Concretely, that experiments with similar or identical designs conducted by different scientists (or by the same researchers for the second time) delivered widely different results. The exact percentage of replicablility is unknown, though some indication might be gleaned from large scale replication projects (e.g., Klein et al., 2018; Open Science Collaboration, 2015). In the case of the Open Science Collaboration (2015), hundreds of scientists collaborated to attempt replication of one-hundred experiments published in prestigious psychological journals. Less than

half of the attempts were successful;[16] clearly a disappointing result.

Replicability can be compromised by many factors. One of them is the misuse of degrees of freedom (e.g. Simmons et al., 2011; Wicherts et al., 2016). Specifically, biased studies are more likely to deliver results which fit the particular interest of the scientist (Section 5.3.1), or general interest in positive results or absence of negative results (Section 5.3.2), which therefore will likely disagree with the results of unbiased experiments; decreasing the overall replicability. Now, if the objectivity of scientific practice (i.e., resistance against bias and questionable research practices) is increased, then replicability on any reasonable metric will increase. In light of that, increasing objectivity seems to be a necessary steps toward solving the replication crisis and its effectiveness will be clearly observable in the published scientific literature.

In addition, our notion gives clear indications of which suggested solution to problematic scientific practices will most likely be successful. Some of these restrict scientists directly (e.g., preregistration requirement, random sampling, randomization, etc.), while others make it harder to exploit degrees of freedom (e.g. blind analysis). Because of that, they improve the objectivity to a certain extent. On the other hand, for some of the proposals it is not clear if they are capable of improving objectivity. The *Reformist Package* is an example of such a proposal. It requires that the first author of a paper on a scientific experiment states all potential conflicts of interest. This amounts to explicitly listing all sources of funding that supported his/her work and claiming full responsibility for the result and decision to publish it. The Reformist Package has some proponents in scientific literature (e.g., Stelfox, Chua, O'Rourke, & Detsky, 1998) and some of the most important scientific journals (e.g., Lancet, Journal of the American Medical Association, etc.) adopted it in their publishing policy. However, according to our conceptualization, it is not clear at all if the proposal improves the objectivity. The Package is forcing scientists to reveal potential causes of systematic bias in the form of financial ties, but it does not safeguard the experiment against actions that can introduce this bias. Our conceptualization predicts that the Reformist Package is ineffective in dealing with the influence funding agencies have, via their researchers, on the results. This is corroborated by the dissatisfaction concerning its ineffectiveness common in current literature (e.g., Schafer, 2004), and is supported by the results of empirical research (e.g., Cain, Loewenstein, & Moore, 2005).

Additionally, our notion can be used to assess the objectivity of research practices

[16] A clear and formal definitions of replication is still absent and several benchmarks were used in this paper. On none of them did the replication rate exceed 50%.

reported in scientific papers (e.g., through a checklist; Appendix C). As an example, we can use the previously mentioned, notorious precognition paper by Bem (2011). This article reports nine experiments that allegedly provide evidence for the hypothesis that future events affect human beliefs (precognition). These results are treated by scientists with skepticism because the existence of precognition is inconsistent with laws of nature (e.g., the second law of thermodynamics), common sense, and everyday experience. Not surprisingly, the subsequent replication attempt failed (e.g., Galak, Leboeuf, Nelson, & Simmons, 2012; Ritchie, Wiseman, & French, 2012) and evidence of the use of QRPs was found (e.g., Francis, 2014; Schimmack, 2012, 2020).

The procedures employed for the nine experiments would not score high on our conception of objectivity. Some aspects of the design of the experiment promote objectivity, the outcome measure and intervention are directly connected to the studied phenomenon and the allocation of subjects was random. On the flip side, participants were exclusively psychology students, the study was not preregistered, and neither the blind analysis nor multiverse analysis was used.[17] The absence of such countermeasures makes the experiment susceptible to the QRPs. An example of such a practice is looking at the initial data of many started experiments and continuing only those that look 'promising'; i.e., only continue with studies that show high initial effect sizes that are due to random chance alone (for evidence for this claim, see Schimmack, 2020).

This is an intuitive result given the skepticism concerning the results of precognition. Moreover, as we have seen, the subsequent replication failed to replicate the original result and evidence suggesting the the QRP were used during the experiment. The objectivity of many other older experiments will be similarly disappointing. The methodological problems central to our conceptualization of objectivity were not widely acknowledged and the countermeasures against them were rarely implemented. This may seem to be a disappointing consequence but it is consistent with low rates of replicability of classical studies (e.g., Klein et al., 2018) and acknowledges the recent rapid development in scientific methodology.

Finally, our conceptualization of objectivity is compatible with, and follows the spirit of many traditional theories of objectivity. Our notion is based on the intuition that objectivity is essentially about minimizing the influences that the individual traits of a scientist have on her research (results). This intuition inspired many other conceptualizations of objectivity, for instance, value-free objectivity and procedural

[17] We acknowledge that these practices were not widely known at the time. However, this does not mean that we cannot assess previous practice by current standards.

objectivity (Section 5.2). Specifically, the value-free conception of objectivity claims that a scientific justification is objective as long as it is not influenced by non-epistemic values. However and in contrast to our conception, the value-free objectivity is hard to assess and therefore use in practice. This is the case because there is no reliable way to assess and test what was the motivation behind any methodological choice.

The same goes for procedural objectivity. This proposal has been previously criticized in Jukola (2017) and we identify two additional problems. First, as in case of VFI, it is prohibitively difficult to verify if a given process is objective in this sense. Secondly, the conceptualization is too restrictive. For example, when statistical methods are used to analyze data it is always the case that the result of an experiment will be different when conducted second time at some level of precision. Furthermore, there is always the possibility of false-positives and false-negatives. Therefore, it seems that no such study can be objective in the sense of the procedural objectivity. Our conceptualization does not suffers from those two difficulties.

Furthermore, our notion is consistent with all descriptive theories, because we do not claim anything about how the concept is used and understood by scientists or natural language users. Besides, some of these descriptive conceptualizations seem to be based on the above mentioned intuition as well. For example, the epistemic risk account of objectivity of Koskinen (2018), seems to be similar in spirit to our proposal. It claims that objectivity consists in averting epistemic risks arising from imperfections of epistemic agents. Adhering to the recommendations of our proposal averts some of such risks, for example, the risk of delivering a biased result due to study design choices (Section 5.3.1). In other words, her description of how objectivity is understood fits to a certain extent with our recommendations. Regulatory objectivity, described in Cambrosio, Keating, Schlich, and Weisz (2006), is another example of a descriptive conceptualization based on the same intuition. It is built on the historical analysis of objectivity from Daston and Galison (1992, 2007). Regulatory objectivity consists of conventions which aim to ensure research quality, specifically (Cambrosio et al., 2006, p.1):

> *Regulatory objectivity*, that is based on the systematic recourse to the collective production of evidence. Unlike forms of objectivity that emerged in earlier eras, regulatory objectivity consistently results in the production of conventions, sometimes tacit and unintentional but most often arrived at through concerted programs of action.

Recent developments are interpreted as the emergence of a new type of objectivity. Implementing and developing such conventions fit our recommendations for the

prevention of methodological choices that can bias results or inflate false-positive rates. Again, there is coherence between our normative proposal and the descriptive theory which describes how scientists understand the objectivity.

5.4.3. Conclusions

Let us once again imagine our scientists, Dr. Jane Summers. Dr. Summers is starting a new experiment (e.g., the effects of caffeine on attention, short-term memory, and long-term memory in psychologically healthy adults) but this time she has a grasp on the notion of objectivity and will include (some of) the objectivity promoting precautions. Specifically, when she designs the study, she ensures that for all intents and purposes the participants selection is random from the population of interest (e.g., males and females, age 21 and up that do not suffer from psychological disorders) and that the non-response rate is not biased (e.g., equal non-response in age and gender), that the measurement instruments come with published validation (i.e., standardized test for attention and memory), the participants' allocation to conditions (e.g., coffee with a high dose of caffeine or decaffeinated coffee) is random, and the experiment is double-blinded (i.e., both participant and experimenter are unaware of experiment condition and purpose). Dr. Summers preregisters the study design and the analysis (e.g., structural equation model) of the main effect of interest (e.g., caffeine positively affects long-term memory, mediated by attention and short-term memory). She will have her data blinded and processed by an independent researcher. In addition, she reserves a room for a multiverse analysis. In Dr. Summers' case, not much is known about the complex relation between dependent and independent variables and its mediation or moderation by participant characteristics (e.g., sex, age, daily caffeine consumption, etc.). Thus, apart from the main model suggested by theory and previous research, she wishes to explore other theoretically possible options. Specifically, she performs and reports the results of the analyses of all theoretically possible models and summarizes their results in a multiverse analysis. By taking these steps, Dr. Summers restricts many ways in which her study can be biased and thereby improves the objectivity of her work.

Similar steps may be taken in order to improve the objectivity of Bem's (2011) experiments. The main problem with the experiments is the (possible) use of QRPs. In particular, he seemed to have started many experiments and only continued collecting data on those that showed 'promising' results (Schimmack, 2020). This could be countered by ensuring preregistration of all initial studies and requiring an appropriate analysis plan if Bem intended to apply sequential analyses. If the

diagnosis by Schimmack (2020) is correct, this alone would significantly increase the reliability of the experiments to the point that it would be highly improbable that they would deliver the suspicious result. In addition, one could require a multiverse analysis over control variables (e.g., gender) and experimental variation (e.g., subcategories of stimuli). As an example for such a requirement, in one of the experiments the precognition effect was observed for pornographic stimuli, but not for neutral stimuli (Bem, 2011). In brief, it is to be expected that implementing the safeguards proposed in this paper would have prevented Bem from getting his results that humans have the ability to feel the future.

To summarize, in this paper we have presented a practicable notion of scientific objectivity. In our opinion, popular disquisitions on objectivity are focused on what the concept means and how it is used, but they do not provide scientists with any guidance on how to improve or assess the objectivity of their work. We presented our empirically informed version of via negativa approach to objectivity and conceptualization of objectivity as methodological resilience. Finally, we showed that and how this new conceptualization can plausibly be used by scientists. In the present form, our theory is far from perfect or complete. At the same time, like science itself, it has the potential to be adjusted and developed to move ever closer to adequacy and completeness.

6

SEVERITY

Abstract:
The Bayesian framework for statistical inference is increasingly popular in psychology. However, one of the main objections to Bayesian inference is that it gives a poor account of the severity of a test, and whether a theory has survived repeated and stringent attempts to prove it wrong. Severity forms the foundation of the error-statistical account of inference, which is often placed in opposition to the Bayesian framework. Here we show that the error-statistical approach leads to a problematic concept of statistical evidence. We then discuss the value of severity within the Bayesian framework via the value of specific, risky predictions and illustrate the practical use of Bayesian severity-based reasoning by means of various practical examples.

This chapter is adapted from Dongen, N.N.N., Sprenger, J.M., & Wagenmakers, E.-J. (under review) A Bayesian perspective on severity: Risky predictions and specific hypotheses. *Psychonomic Bulletin & Review*.

6.1. Introduction

The Bayesian framework for statistical inference—expressing one's uncertainty about hypotheses and parameters by means of subjective probabilities and updating them through Bayes' theorem—is increasingly popular in psychology (e.g., Lee & Wagenmakers, 2013; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Vandekerckhove, Rouder, & Kruschke, 2018). This popularity is easy to explain: Bayesian inference is based on a general theory of uncertain reasoning, its basic principles are simple and easily remembered, statistical evidence is quantified straightforwardly by means of relative predictive performance (i.e., Bayes factors), and statistical inferences are connected to our beliefs and the practical consequences of our decisions (e.g, Bernardo & Smith, 1994; Evans, 2015; Howson & Urbach, 2006; Jeffrey, 1971; Jeffreys, 1961; Lindley, 2000; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016; Savage, 1972; Sprenger & Hartmann, 2019).

An intuitively appealing competitor to Bayesian inference is *severe testing*. According to that approach, the genuine confirmation of a scientific theory does not consist in agreement of a theory with experimental data since post-hoc confirmations are too easy to obtain. For instance, in extreme cases, researchers can design experiments that are guaranteed to produce confirming results. Instead, a theory should count as confirmed only if it has survived repeated and stringent attempts to prove it wrong (e.g., Mayo, 1996, 2018; Meehl, 1978, 1986, 1990a, 1990b; Popper, 2002[1959]). In other words, what matters for a theory's status is its *probative value*, or, in other words, whether it has "proved its mettle" (Popper, 2002[1959], p. 264). In Bayesian inference, however, the (posterior) plausibility of a theory does not directly depend on the severity of a test, or the extent to which one has tried to prove the theory wrong. Prima facie, the idea of severe testing is thus at odds with Bayesian inference while it appears to be a natural match for frequentist inference, where the idea of controlling error in inference stands central (Mayo, 1996; Neyman, 1977; Neyman & Pearson, 1933, 1967).

Bayesians can respond to this challenge in one of two ways. The first response is to deny its relevance and to refer to the *Likelihood Principle*: all the evidence that an experiment provides about an unknown quantity is expressed by the likelihood function of the various hypotheses on the observed data (Berger & Wolpert, 1984; Birnbaum, 1962; W. Edwards, Lindman, & Savage, 1992). This likelihood function is amalgamated with the prior probability distribution, which represents our background knowledge, for producing posterior probabilities and determining the correct inferences. "Consequently the whole of the information contained in the

observations that is relevant to the posterior probabilities of different hypotheses is summed up in the values that they give to the likelihood" (Jeffreys, 1961, p. 57). Therefore, the severity of a test, as expressed by whether the hypothesis of interest stood a risk of being refuted (e.g., whether unobserved data could have proven it wrong), cannot directly enter the evaluation of an experiment. According to this line of response, critics of Bayesian inference are just mistaken when they believe that severity should enter the assessment of a theory.

The second Bayesian response is more conciliatory in spirit; it aims at showing that the concept of severity is naturally accommodated within the Bayesian framework. In this paper we elaborate on this strategy, which has not yet been developed in a systematic way. First, we expound the rationale for severe testing from general methodological considerations. Then we show that frequentist explications of severity run into several substantive problems. Moreover, we show that contrary to popular opinion, severe testing is not uniquely anchored in frequentist inference: we construe severity within the Bayesian framework as testing only those hypotheses that make specific and risky predictions. Specifically, we explicate the notion of severity as the specificity of the prediction of a hypothesis in relation to the potential data that could be observed. The value of severity is reflected in the *expected evidential value* of a test, quantified by the expected absolute log-Bayes factor (Cavagnaro, Myung, Pitt, & Kujala, 2010; Good, 1979; Lindley, 1956; Schönbrodt & Wagenmakers, 2018; Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019).

This approach can be put into practice using the "encompassing prior" framework (e.g., Hoijtink, Klugkist, & Boelen, 2008; Klugkist & Hoijtink, 2005; Klugkist, Kato, & Hoijtink, 2005). Researchers who are in the position to test more specific predictions are rewarded because they need smaller sample sizes in order to obtain the same amount of evidence (or more evidence for the same sample size) and can expect on average more evidence from their tests. The paper concludes with sketching the advantages, applications and limitations of severity-based Bayesian inference.

6.2. Theoretical Background: Scientific Theories and Severe Tests

Our scientific theories are imperfect descriptions and explanations of reality. They allow us, with a certain degree of accuracy, to predict future observations or to retrodict structure in data already available. According to Popper (2002[1959]), the value or *empirical content* (p. 96) of a scientific theory lies in the combination of its *universality* and its *precision* (i.e., generality to what it pertains and specificity of what it predicts/explains; pp. 105–106). A theory has high empirical content when it has

a vast scope in which many observations are possible in principle (e.g., all planetary motions in the universe) yet only a few of these possible observations are consistent with the theory (e.g., ellipses around a center of gravity). This is consistent with the common intuition that a successful risky prediction is more impressive than a successful vague prediction. Meehl (1978) expressed this intuition in an example of two meteorological theories that both make predictions on next month's weather. Theory A predicts: in Turin it will rain on three days in the next month. Theory B predicts: in Turin it will rain on the third, fourth, and the seventh day of next month. We expect that most people agree with Meehl that a success of Theory B's prediction is more impressive than the success of Theory A's prediction. This difference can be made explicit in terms of its *falsifiablity* (Popper, 2002[1959], p. 96): Theory B's prediction is only correct in $\binom{30}{1} = 1$ out of 2^{30} possibilities, while Theory A's prediction is correct in $\binom{30}{3} = 4060$ out of 2^{30} possibilities. What is important for the standing of a scientific theory is not its ability to fit the data, but how specific it is with respect to all possible data (see also Roberts & Pashler, 2000). Popper (2002[1959]) gave the examples of Freudian psychoanalysis and Marxist sociology as theories that could fit any pattern of data, thus being low in empirical content and even unscientific. For these 'theories', both the presence and absence of a personality trait or both presence and absence of worker unrest could be considered as consistent with the theory.

When testing competing theories, they are falsifiable to the extent that they make sharply contrasting predictions. The outcome of a test should either favor one theory and contradict the other, or vice-versa (Platt, 1964). In other words, one needs to maximize the expected information gain (Myung, Cavagnaro, & Pitt, 2013; Myung & Pitt, 2009; Oaksford & Chater, 1994) regarding the assessment of the competing theories:

> The scientific method is always comparative and there are no absolutes in the world of science. It follows from this comparative attitude that a good theory is one that enables you to think of an experiment that will lead to data that are highly probable on [the theory], highly improbable on [the complement of the theory], or vice versa, so that the likelihood ratio is extreme and your odds substantially changed. (Lindley, 2006, p. 209)

A prime example is Eddington's 1919 test of Einstein's General Theory of Relativity versus Newton's classical theory of gravity (Dyson, Eddington, & Davidson, 1920). Both theories make sharp mutually exclusive predictions about the degree to which passing massive bodies, like our sun, bend light from distant sources, like other stars.

The predictions of GTR were famously verified by Eddington during the 1919 solar eclipse.

The benefits to science from increasing the falsifiability of theories and riskiness of predictions have been presented in numerous critiques of social science methodology (e.g., Meehl, 1978, 1986, 1990a, 1990b; Roberts & Pashler, 2000). In particular, the specificity of the predictions of a theory is tightly linked to the *capacity for testing it severely* (and thus, to its falsifiability). The more possible observable outcomes are consistent with a theory, the less informative the theory becomes. An example of uninformativeness would be a meteorological theory that predicts that it will rain in Amsterdam on some days in the next year. On the opposite side we have theories that only allow a set number of parameters with fixed values and interrelations. Again, an excellent example of a theory with high informative content is General Relativity. This theory makes specific predictions about how much faster time moves for our GPS satellites relative to the earth's surface, which allows continued navigational accuracy. The more informative a theory is, the more falsifiable it is, because fewer possible observations are consistent with it. Only such informative theories can be severely tested (Jeffreys, 1973, pp. 39–40):

> We may say that to make predictions with great accuracy increases the probability that they will be found wrong, but in compensation they tell us much more if they are found right. [...] The best procedure, accordingly, is to state our laws as precisely as we can, while keeping a watch for any circumstances that may make it possible to test them more strictly than has been done hitherto.

Popper's definition of *degree of corroboration*, proposed for evaluating the standing of scientific theories, reflects the value of falsifiability and adds a record of severe testing as a key component (Popper, 1979, p. 18, our emphasis):

> By the degree of corroboration of a theory I mean a concise report evaluating the state (at a certain time *t*) of the critical discussion of a theory, with respect to the way it solves its problems; its degree of testability; *the severity of tests it has undergone; and the way it has stood up to these tests*.

Notably, severe testing can be even more important than success in problem-solving and fit with the data.[1] We now investigate how severity manifests itself in various statistical frameworks, and whether it indeed plays the role that Popper, Meehl, and other methodologists have reserved for it.

[1] Compare also the following quote: "it is not so much the number of corroborating instances which determines the degree of corroboration than *the severity of the various tests* to which the hypothesis in question can be, and has been, subjected" (Popper, 2002[1959], p. 266, original emphasis).

6.3. Severity in Statistical Inference Frameworks

6.3.1. From Significance Testing to Error Statistics

At first glance there is a striking resemblance between Popper's falsificationist approach and *null hypothesis significance testing* (NHST). We will now explain the logic of NHST using *parametric hypotheses*, that is, hypotheses that make claims about the value of an unknown parameter of interest.

The basis of NHST is a *point null hypothesis* postulating that a parameter takes a precise value (e.g., the mean of a population, $\mathcal{H}_0 : \mu = \mu_0$)—usually corresponding to a complete absence of a causal effect in an experimental intervention, the equality of two medical treatments, and so on. The significance tester then evaluates the tenability of that null hypothesis by means of the p-value: the probability that data X, for the given experimental design and the (hypothetical) truth of the null hypothesis, have deviated from the null hypothesis more than realized data x (i.e., $p(t(X) \geq t(x); \mathcal{H}_0)$; the statistic t measures divergence of the data from $\mathcal{H}_0$). If this probability is sufficiently small, the null hypothesis is rejected: "either an exceptionally rare chance has occurred, or the theory [i.e., the null hypothesis] is not true" (Fisher, 1956, p. 39).[2]

In various dimensions, the NHST methodology squares well with Popper's falsificationism. First, the point null hypothesis is maximally falsifiable. Second, the idea of submitting the point null to a severe test and of rejecting it when it fails to explain the data, has a distinct Popperian note. However, the scientific theory that we would like to submit to a severe test is usually *not* the null hypothesis (but see Gallistel, 2009; Rouder et al., 2009; Sprenger & Hartmann, 2019, ch. 9). Rather, we would like to severely test the hypothesis that there *is* an effect: that a medical drug is better than a placebo, that playing a musical instrument makes people happier, that video games improve reasoning skills, and so on. This 'alternative' hypothesis usually takes the generic form $\mathcal{H}_1 : \mu > \mu_0$ (or even $\mathcal{H}_1 : \mu \neq \mu_0$). It does not commit the scientist to precise predictions; rather it states that the unknown parameter lies in a *range* of values.

The significance tester applies a form of *eliminative inference*: when the point null hypothesis $\mathcal{H}_0$ obtains a low p-value, it is eliminated from consideration and the significance tester infers the presence of an effect. However, note that the evidence against $\mathcal{H}_0$ does not tell us anything about the *magnitude* of the effect; in fact, some

[2] Note that contemporary significance testing should not be identified with the methodology that Fisher (1935b, 1956) developed; rather, contemporary practice is a hybrid of Fisherian inference and the behavioral approach of Neyman and Pearson (1933). For a discussion see Gigerenzer (1998) and Hubbard and Bayarri (2003).

elements of $\mathcal{H}_1$ (i.e., specific parameter values close to μ_0) would neither pass the test that $\mathcal{H}_0$ has failed. Thus, the inference to $\mathcal{H}_1 : \mu > \mu_0$ or $\mathcal{H}_1 : \mu \neq \mu_0$ on the basis of "rejecting" $\mathcal{H}_0 : \mu = \mu_0$ is both indiscriminate and invalid from a severity-based perspective: it neither informs us about the size of the effect, nor have any particular parameter values contained in $\mathcal{H}_1$ been well probed. It is then questionable to what extent NHST provides for a severe test of the hypothesis of interest. In fact, the indiscriminate characterization of alternative hypotheses, and the neglect of effect magnitude in theorizing and testing is a classical criticism of psychological methods and social science in general (e.g., J. Cohen, 1994; Meehl, 1967; Ziliak & McCloskey, 2008). Plausibly it is also among the causes of the notorious replication crisis.[3]

Mayo's (1996; 2018) *error-statistical approach* addresses this problem and tries to explicate how parametric hypotheses can be tested severely. Mayo develops a general philosophy of assessing hypotheses based on the severity with which they have been tested. Error statistics embodies to a certain extent Popper's perspective on theory testing as the ability to detect and control error between data and hypothesis. Like Popper, Mayo believes that the standing of a theory is primarily determined by the severity of the tests it has survived. The extent to which the theory has proved its mettle constitutes statistical evidence. Mayo (2018, p. 14) states this explicitly in her Strong Severity Principle:

> **Severity Principle (strong):** We have evidence for a claim C just to the extent it survives a stringent scrutiny. If C passes a test that was highly capable of findings flaws or discrepancies from C, and yet none or few are found, the passing result, x, is evidence for C.

However, we believe this principle is problematic, for three general reasons.

First, the context in which the test is performed is not specified. It is not clear how a claim can "survive a stringent scrutiny" without a specific context in which it is embedded. For instance, a DNA test on partially recovered materials can be considered a severe test to identify the perpetrator of a crime from, say, ten suspects. However, this is not the case when all inhabitants of the state or country are included in the suspect pool, or when two of the ten suspects are monozygotic twins and test positive. Context is essential for evaluating whether a test is "highly capable of findings flaws or discrepancies from [claim] C". If C encompasses all that is

[3]Recent attempts to improve psychological methods based on equivalence tests or "regions of practical equivalence" (of effect size) address this criticism—and acknowledge the value of severe testing—by making the alternative hypothesis more specific, based on what would count as a meaningful deviation from the null (Kruschke & Liddell, 2018; Lakens, 2017a; Lakens, Scheel, & Isager, 2018).

possible, no test —no matter how probative the test usually is— will be capable of finding flaws in it. For example, a stringent scrutiny of the claim "all swans are white" requires only a single swan if the alternative is "all swans are black"; when the alternative is that "most swans are white, but a few are black", however, the observation of a single white swan is relatively uninformative. The Severity Principle must not be interpreted in absolute, but in relative terms. This aspect is missing in Mayo's formalization.

Second, Mayo's perspective does not explicate an essential element of severe testing that Popper has time and again stressed: the universality of the tested hypothesis $\mathcal{H}$ and the specificity of the predictions it makes (Popper, 2002[1959], p. 266). Nowhere in Mayo's severity principle or in the more specific explications discussed below is the riskiness or specificity of a hypothesis, and the predictions it makes, incorporated as a requirement for a severe test (Meehl, 1978, 1986, 1990a, 1990b; Roberts & Pashler, 2000).

Third, Mayo's operationalization of the Severity Principle leads to an inference structure that is similar to NHST, thus shares some of their shortcomings, and has the potential to lead to problematic and counter-intuitive conclusions. Mayo (2018, p. 92) proposes the following requirements for when data support hypothesis $\mathcal{H}$:

> **Severity Requirement:** for data to warrant hypothesis $\mathcal{H}$ requires not just that
> (S-1) $\mathcal{H}$ agrees with the data ($\mathcal{H}$ passes the test), but also
> (S-2) with high probability, $\mathcal{H}$ would not have passed the test so well, were $\mathcal{H}$ false.

Suppose that the hypothesis of interest is $\mathcal{H}_1 : \mu \geq \mu_0 + \delta$, where $\delta > 0$ expresses the effect size relative to the default value μ_0. The application of (S-1) is then similar to the Neyman-Pearson (and NHST) approach (see Mayo, 2018, p. 142): $\mathcal{H}_1$ passes a statistical test when $\mathcal{H}_0 : \mu \leq \mu_0$ is rejected at the pre-specified Type I error level α. This happens when the probability of the observed data, or data deviating more extremely from $\mathcal{H}_0$, is lower than α (i.e., $p(t(X) \geq t(X); \mathcal{H}_0) < \alpha$).[4]

To assess the evidence in favor of $\mathcal{H}_1$, Mayo applies (S-2) and computes a severity function (SEV). The severity function takes as arguments the statistical test, the observed data, and the target hypothesis $\mathcal{H}_1$ (Mayo calls it the "Claim"). As a representative of the negation of $\mathcal{H}_1$, we choose the point hypothesis in $\mathcal{H}_0$ that is closest to $\mathcal{H}_1$, that is, $\mu = \mu_0$. If this hypothesis is rejected with test T^+ based on data

[4]Note that, due to the δ added to μ_0, $\mathcal{H}_1$ is not (necessarily) the complement of $\mathcal{H}_0$.

x, the severity with which $\mathcal{H}_1$ has passed the test is defined as

$$SEV(T^+, x, \mathcal{H}_1) = p(t(X) \leq t(x); \neg\mathcal{H}_1) = p(t(X) \leq t(x); \mu = \mu_0 + \delta),$$

which is (a lower bound on) the probability of observing the actually observed data, or data more deviating from $\mathcal{H}_1$, in the direction of $\mathcal{H}_0$ if $\mathcal{H}_1$ were false.[5] This probability needs to be high in order to meet severity requirement (S-2) "with high probability, [$\mathcal{H}_1$] would not have passed the test so well, were [$\mathcal{H}_1$] false" (Mayo, 2018, p. 92). In other words, the larger this probability, the more probable it is that one would have observed data less congruent with $\mathcal{H}_1$ (i.e., the more unlikely it is that $\mathcal{H}_1$ "would have passed the test so well'), *if* $\mathcal{H}_1$ *were false*. Mayo calls the probability that results from the SEV function (degrees of) "severity". Note that this probability is actually a one-sided p-value, testing $\mathcal{H}_1$ rather than $\mathcal{H}_0$. This means that Mayo's account is open to several statistical objections that have been raised against p-values in NHST (e.g., Benjamin et al., 2018; W. Edwards et al., 1992; Jeffreys, 1961; Robinson, 2019): ubiquitous misinterpretation, openness to manipulation through multiple testing, dependence on data never observed, mechanical application of all-or-none decision rules in the face of weak evidence, and overestimation of evidence in general. Given the doubts that all these objections shed on the veracity of evidential support as expressed by p-values, we doubt that the SEV function is a good indicator of the severity of a test of $\mathcal{H}_1$, and of its resilience to further experiments. Moreover, the SEV function operationalizes the degree of severity with which (a particular component of) $\mathcal{H}_1$ passes the test as a quantity that is *independent of* $\mathcal{H}_0$ (given that $\mathcal{H}_0$ is rejected). Concretely, $SEV(T^+, x, \mu = \mu_0 + \delta;\ \delta > 0)$ has the same outcome whether $\mathcal{H}_0 : \mu = \mu_0$, $\mu = \mu_0 - 1$, $\mu = \mu_0 - 2$ or any other value sufficiently small to have $\mathcal{H}_0$ be rejected by data x. As observed above, Mayo's formalization of severity neglects the context of the research question: is the actual deviation from $\mathcal{H}_0$ sufficiently large to reject that hypothesis with a high degree of severity?

The basic idea of a frequentist hypothesis test consists in looking for deviations from the normal or default state of affairs, as expressed by the null hypothesis $\mathcal{H}_0$. Is an effect due to chance or should it be given a causal interpretation (e.g., Fisher, 1935a, 1956)? This rationale has motivated most of frequentist theory and philosophy, but it gets lost in Mayo's approach. If the null hypothesis $\mathcal{H}_0$ does not matter for an evaluation of the severity with which a specific alternative passes the test, we

[5] For convenience, the SEV function is evaluated at $\mu = \mu_0 + \delta$, because the probability value that it produces will be even greater for $\mu < \mu_0 + \delta$ (for detailed explanation, see Mayo, 2018, p. 144).

In the case of a unidirectional test of $\mathcal{H}_1$, the severity function becomes $SEV(T^-, x, \mathcal{H}_1) = p(t(X) \geq t(x); \neg\mathcal{H}_1) = p(t(X) \geq t(x); \mu \geq \mu_0 + \delta)$.

basically give up the contrastive nature of hypothesis testing. But then we find it unclear why the error statistician would still like to talk about severity. A worked-out example of Mayo's severity function in practice is given in Appendix D.

6.3.2. Bayesian Inference

The Bayesian counts observations x as supporting a hypothesis $\mathcal{H}$ if and only if x raises the probability of $\mathcal{H}$ (e.g., Carnap, 1950; Evans, 2015; Horwich, 1982; Howson & Urbach, 2006; Sprenger & Hartmann, 2019):

$$p(\mathcal{H} \mid x) > p(\mathcal{H}).$$

Equivalently, in the case of a null hypothesis $\mathcal{H}_0$ and an alternative $\mathcal{H}_1$, x is evidence for $\mathcal{H}_1$ if and only if

$$\mathrm{BF}_{10}(x) := \frac{p(x \mid \mathcal{H}_1)}{p(x \mid \mathcal{H}_0)} > 1.$$

In other words, x raises the probability of $\mathcal{H}_1$ if and only if x is better predicted under $\mathcal{H}_1$ than under $\mathcal{H}_0$, that is, the Bayes factor BF_{10} exceeds 1. The higher the Bayes factor, the stronger the evidence for $\mathcal{H}_1$, and the closer it is to zero, the stronger the evidence for $\mathcal{H}_0$. See also Table 6.1.

Table 6.1: Classification of Bayes Factors according to Lee and Wagenmakers (2013), adjusted from Jeffreys (1961).

Bayes Factor BF_{10}	Interpretation
>100	Extreme evidence for H_1
30–100	Very strong evidence for H_1
10–30	Strong evidence for H_1
3–10	Moderate evidence for H_1
1–3	Anecdotal evidence for H_1
1	No evidence for either hypothesis
1/3–1	Anecdotal evidence for H_0
1/3–1/10	Moderate evidence for H_0
1/10–1/30	Strong evidence for H_0
1/30–1/100	Very strong evidence for H_0
<1/100	Extreme evidence for H_0

The Bayesian idea of measuring evidence by means of the Bayes factor avoids the main problem that we stated for Mayo's severe testing: it is relative to an explicit choice of context (i.e., the alternative hypothesis in the calculation of the Bayes factor) and its quantification of statistical evidence is conditional on that context. Even so, the inferences of the Bayesian and Mayo's severe tester will often agree in practice. Figure 6.1 provides a graphical illustration: the severity with which a hypothesis passes a test is close to unity (and more specifically, above the $1 - \alpha$ threshold) if

and only if the Bayes factor indicates strong or overwhelming evidence for that hypothesis.

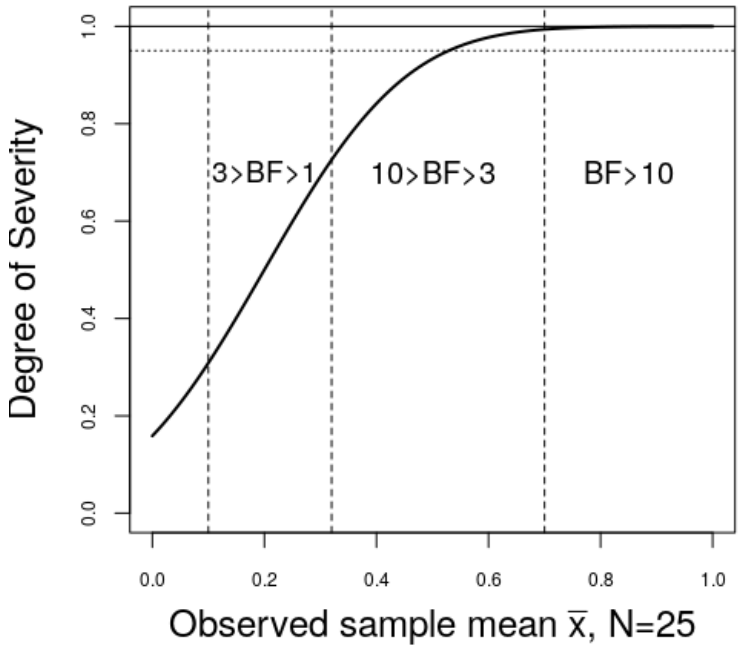


Figure 6.1: Severity function for the inference about the mean of a normal distribution μ with known standard deviation $\sigma = 1$, as a function of the observed sample mean for $N = 25$. The tested hypothesis is $\mu > .2$ and the dashed horizontal line indicates the 95% severity label to facilitate orientation. Vertical lines indicate which ranges of the observed data correspond to different ranges of Bayesian evidence, calculated as BF_{10} with the two point hypotheses $\mathcal{H}_0 : \mu = 0$ and $\mathcal{H}_1 : \mu = 0.20$.

The posterior probability of a theory $\mathcal{H}_0$ given data x is calculated as $p(\mathcal{H}_0 \mid x) = p(\mathcal{H}_0) \cdot p(x \mid \mathcal{H}_0)/p(x)$. This allows us to write the posterior odds in favor of $\mathcal{H}_0$ over $\mathcal{H}_1$ as

$$\frac{p(\mathcal{H}_0 \mid x)}{p(\mathcal{H}_1 \mid x)} = \frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)} \cdot \frac{p(x \mid \mathcal{H}_0)}{p(x \mid \mathcal{H}_1)}.$$

Thus, keeping the prior probability $p(\mathcal{H}_0)$ fixed, the posterior probability of $\mathcal{H}_0$ will be the larger (1) the better $\mathcal{H}_0$ predicts x and (2) the more surprising or the worse predicted x is under $\mathcal{H}_1$ (e.g., Howson & Urbach, 2006; Roberts & Pashler, 2000). Note the two different kinds of counterfactual reasoning: the Bayesian only considers the probability of the specific evidence provided by data x under $\mathcal{H}_0$ and $\mathcal{H}_1$ while

Mayo considers which values had been less probable than x under $\mathcal{H}_1$.

At first glance, severity-related aspects appear to be lacking in the Bayesian paradigm. However, they are lurking just beneath the surface: the Bayesian cashes out severity in terms of the expected success of the predictions that Theory $\mathcal{T}$ makes with respect to its negation, $\neg\mathcal{T}$. Specifically, the Bayesian answers the question "... what constitutes a good theory?" (Lindley, 2006, p. 196) in terms of its potential to yield high likelihood ratios in testing $\mathcal{T}$ against its negation $\neg\mathcal{T}$ (Lindley, 2006, p. 197, notation changed):

> [...] A good theory is one that makes lots of predictions that can be tested, preferably predictions that are less probable were the theory not true. [...] what is wanted are data that are highly likely when the theory is true, and unlikely when false. A good theory cries out with good testing possibilities.

In the next section, we bring these aspects into the limelight and show how the virtue of severity in scientific inference translates, within Bayesian statistics, to the virtue of making precise, specific predictions, and how the specificity of predictions boosts the evidential value of an experiment for a tested hypothesis.

6.4. Specific Hypothesis Testing

We now discuss a straightforward approach of incorporating severe testing into Bayesian inference: for a test to be severe, the tested hypothesis needs to *impose substantial restrictions on the range of potential data* that are consistent with it.

This notion is also reflected in the proposal of Roberts and Pashler (2000) for a stricter evaluation of model-fit as evidence. They argue that a good fit between data and model does not have to be convincing as evidence when the parameter values of the model are fully adjustable to accommodate the data. As Roberts and Pashler (2000, p. 359) put it:

> Theorists who use good fits as evidence seem to reason as follows: if our theory is correct, it will be able to fit the data; our theory fits the data; therefore it is more likely that our theory is correct. However, if a theory did not constrain possible outcomes, the fit is meaningless. A prediction is a statement of what a theory does and does not allow. When a theory has adjustable parameters, a particular fit is just one example of what it allows. To know what a theory predicts for a particular measurement you need to know all of what it allows (what else it can fit) and all of what it does not allow (what it cannot fit).

Figure 6.2 shows for a theory concerned with the relation between the values of two measures that the data can provide strong support only if a narrow range of possible values is consistent with the theory and the data fall in this narrow range.

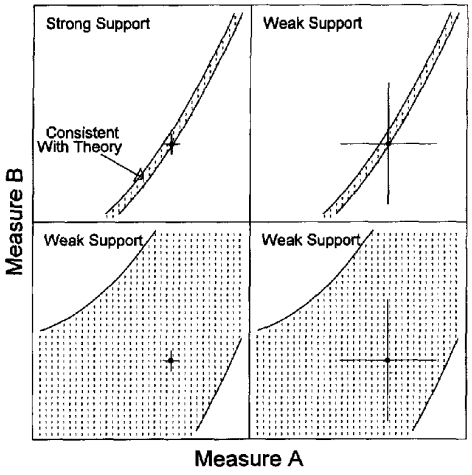


Figure 6.2: Figure from Roberts and Pashler (2000): "Four possible relationships between theory and data. Measures A and B are both measures of behavior. For both measures, the axes cover the whole range of possible values. The dotted areas indicate the range of outcomes that would be consistent with the theory. The error bars indicate standard errors. In every case, the theory can closely fit the data, but only when both theory and data provide substantial constraints does this provide significant evidence for the theory." (Roberts & Pashler, 2000, p. 360).

What does this imply for parametric hypothesis tests? Popper notes that a hypothesis is the more testable, and the more restrictive on the data, the simpler it is, where simplicity is explicated as paucity of parameters (Popper, 2002[1959], pp. 126–128 and 392–394). In the context of statistical testing, this means that the most simple, and most severely testable hypothesis is a point null hypothesis $\mathcal{H}_0$ which assigns zero degree of freedom to the parameter of interest. More generally, there is a systematic connection between the degree to which a hypothesis restricts its parameters, and the degree to which it restricts the range of data consistent with it (see our example in Section 3.2).

From a Bayesian perspective, the above considerations are reflected in the *expected evidential value* of a hypothesis test, which can be made precise as the expected absolute log-Bayes factor relative to one of the two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ (Good, 1979; Lindley, 1956; Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019):

$$\begin{aligned} \mathbb{E}_0\left[|\log \mathrm{BF}_{10}|\right] &= \int p_0(y)|\log \mathrm{BF}_{10}(y)|\,\mathrm{d}y. \\ \mathbb{E}_1\left[|\log \mathrm{BF}_{10}|\right] &= \int p_1(y)|\log \mathrm{BF}_{10}(y)|\,\mathrm{d}y, \end{aligned} \tag{6.1}$$

where p_0 and p_1 are the probability density functions that $\mathcal{H}_0$ and $\mathcal{H}_1$ postulate for the sample space. The log-Bayes factor symmetrically quantifies evidence for (against) $\mathcal{H}_1$ with respect to $\mathcal{H}_0$, with $\log \text{BF}_{10} = 0$ being the indifference point. The expected absolute log-Bayes factor quantifies the amount of evidence one can expect, for and against combined, where values close to $\mathbb{E}|\log \text{BF}_{10}| = 0$ constitute uninformative tests (i.e., the hypotheses make similar predictions).

When a null hypothesis is very specific, its predictions will differ notably from the alternative, and this means that we can expect strong evidential support for either one or the other hypothesis. Furthermore, the expected evidential value is higher for the more restricted hypothesis: it is easier to find evidence against it when it is false, and one obtains more easily evidence for it when it is true. Testing simple, restrictive hypotheses is thus valuable from a Bayesian perspective, too. Conversely, the Bayesian can explain why tests with high expected evidential value are more severe: such tests are, in Mayo's words, more capable of findings flaws or discrepancies from the tested hypothesis. We will illustrate this reasoning with a simple Binomial model example in Section 3.2.

6.4.1. Specific Hypothesis Testing: The General Principle

The Bayesian *informative hypothesis testing* approach (e.g., Klugkist & Hoijtink, 2005; Klugkist et al., 2005) can be used to illustrate how Bayesian inference incorporates the concept of severity as the ability of an experimental design to discriminate the hypotheses at hand. Within this approach, each hypothesis to be tested is derived from a more general *encompassing* model. This model consists of all pertinent parameters free from any constraints (e.g., $\mathcal{H}_e : \mu_a,\ \mu_b$); all specific hypotheses are nested in this model (e.g., $\mathcal{H}_0 : \mu_a = \mu_b$, $\mathcal{H}_1 : \mu_a > \mu_b$, $\mathcal{H}_2 : \mu_a < \mu_b$). Only a prior for the parameters in the encompassing model needs to be specified and the priors from the nested hypotheses follow from this encompassing prior. Specifically, the hypotheses occupy particular sections of the parameter space, allotting each hypothesis a segment of the encompassing prior (for further explanation of the encompassing prior, see Klugkist & Hoijtink, 2005).

We can denote the encompassing model as $\mathcal{H}_e$ and the encompassing prior as $g(\boldsymbol{\theta} \mid \mathcal{H}_e)$, where $\boldsymbol{\theta}$ is the vector of parameters of interest. The prior distribution of any hypothesis $\mathcal{H}_i$ nested in $\mathcal{H}_e$ can be obtained from the encompassing prior by restricting the parameter space according to the limits of the hypothesis, given by

$$g(\boldsymbol{\theta} \mid \mathcal{H}_i) = \frac{g(\boldsymbol{\theta} \mid \mathcal{H}_e) I_{\mathcal{H}_i}(\boldsymbol{\theta})}{\int g(\boldsymbol{\theta} \mid \mathcal{H}_e) I_{\mathcal{H}_i}(\boldsymbol{\theta}) d\boldsymbol{\theta}}.$$

Here, $I_{\mathcal{H}_i}(\boldsymbol{\theta})$ is an indicator function which equals 1 if the parameter value is within the limits of $\mathcal{H}_i$ and 0 if it is outside the limits of $\mathcal{H}_i$.

The evaluation of the hypothesis is based on the Bayes factor. Specifically, support provided by the data for one versus another hypothesis is quantified as the ratio of marginal likelihoods of hypotheses (Kass & Raftery, 1995). For data x and hypotheses $\mathcal{H}_i$ and $\mathcal{H}_e$, the Bayes factor (BF) is

$$\text{BF}_{ie} = \frac{p(x \mid \mathcal{H}_i)}{p(x \mid \mathcal{H}_e)},$$

that is, the quotient of the marginal likelihoods for the hypotheses $\mathcal{H}_i$ and $\mathcal{H}_e$. Due to Bayes' Theorem

$$p(\boldsymbol{\theta} \mid x, \mathcal{H}) = \frac{p(x \mid \boldsymbol{\theta}, \mathcal{H}) g(\boldsymbol{\theta} \mid \mathcal{H})}{p(x \mid \mathcal{H})},$$

the marginal likelihood can, for any value of the unknown parameter $\boldsymbol{\theta}$ consistent with $\mathcal{H}$, also be written as

$$p(x \mid \mathcal{H}) = \frac{p(x \mid \boldsymbol{\theta}, \mathcal{H}) g(\boldsymbol{\theta} \mid \mathcal{H})}{p(\boldsymbol{\theta} \mid x, \mathcal{H})}.$$

In this formulation, the denominator is the posterior density of $\boldsymbol{\theta}$ under model $\mathcal{H}_i$ and the numerator is the product of the likelihood function and the prior distribution. As a consequence, the Bayes factor of $\mathcal{H}_i$ and $\mathcal{H}_e$ can be written as

$$\text{BF}_{ie} = \frac{p(x \mid \mathcal{H}_i)}{p(x \mid \mathcal{H}_e)} = \frac{p(x \mid \boldsymbol{\theta}, \mathcal{H}_i) g(\boldsymbol{\theta} \mid \mathcal{H}_i) / p(\boldsymbol{\theta} \mid x, \mathcal{H}_i)}{p(x \mid \boldsymbol{\theta}, \mathcal{H}_e) g(\boldsymbol{\theta} \mid \mathcal{H}_e) / p(\boldsymbol{\theta} \mid x, \mathcal{H}_e)}. \tag{6.2}$$

If we then suppose, more specifically, that the value $\boldsymbol{\theta}$ is allowed by $\mathcal{H}_i$, (6.2) delivers

$$\text{BF}_{ie} = \frac{g(\boldsymbol{\theta} \mid \mathcal{H}_i) p(\boldsymbol{\theta} \mid x, \mathcal{H}_e)}{g(\boldsymbol{\theta} \mid \mathcal{H}_e) p(\boldsymbol{\theta} \mid x, \mathcal{H}_i)}.$$

Through $\mathcal{H}_i$ being nested in $\mathcal{H}_e$, the densities $g(\boldsymbol{\theta} \mid \mathcal{H}_i)$ and $p(\boldsymbol{\theta} \mid x, \mathcal{H}_i)$ can be rewritten as follows:

$$g(\boldsymbol{\theta} \mid \mathcal{H}_i) = c_i \times g(\boldsymbol{\theta} \mid \mathcal{H}_e)$$
$$p(\boldsymbol{\theta} \mid x, \mathcal{H}_i) = f_i \times p(\boldsymbol{\theta} \mid x, \mathcal{H}_e),$$

where c_i and f_i are constants. Thus, $\mathcal{H}_i$'s prior and posterior densities can be given in terms of the densities of the encompassing model $\mathcal{H}_e$. $1/c_i$ denotes how much of the prior distribution of $\mathcal{H}_i$ is in agreement with the encompassing model $\mathcal{H}_e$, and $1/f_i$ denotes the same quantity for the posterior distribution. $1/c_i$ is therefore

a natural measure of the complexity of the nested model $\mathcal{H}_i$, $1/f_i$ of its post hoc fit. Since $\mathcal{H}_i$ is nested within $\mathcal{H}_e$, we can write the Bayes factor simply as

$$\text{BF}_{ie} = \frac{1/f_i}{1/c_i} = \frac{\text{fit}}{\text{complexity}}$$

From this specification of the Bayes factor, it follows that good model fit is not enough as evidence for one's hypothesis. Near perfect fit is easily reached when close to all possible parameter values and functional forms are allowed. Only when one's hypothesis is restrictive enough that it occupies only a small fraction of possible parameter values and functional forms a priori, though still is in accordance with the data a posteriori, does one have evidence in favor of the hypothesis under consideration when supporting observations are made.

6.4.2. Specific Hypothesis Testing: An Example

Let us clarify this specification of the Bayes factor with a simple fictitious example. We imagine that military veterans receive either one of two treatments to address their post traumatic stress disorder (PTSD). Treatment A is taken to be a regular psychotherapy (e.g, Cognitive Processing Therapy Monson et al., 2006). Treatment B is taken to consists of the same psychotherapy but enhanced with regulated dosages of the drug 3,4-Methylene-dioxymethamphetamine (Sessa, 2017).

Based on the theory and previous research, success rates of these treatments should fall in a particular range. The more uncertainties there are in the theory about what could influence the success rate (e.g., inter-patient psychological variability) and the more measurement error one can expect (e.g., low reliability of PTSD assessment), the wider the interval around these expected values that are still considered consistent with the theory.

Specifically, assume that either treatment has a true success rate, θ_A and θ_B. Under an encompassing model these rates can have any value between 0 and 1, $\mathcal{H}_e : \theta_A = [0.0, 1.0],\ \theta_B = [0.0, 1.0]$. For simplicity, we consider these values equally likely and adopt a uniform prior over θ_A and θ_B in the encompassing model. The scenario can be modelled as two independent binomial distributions. We start treating participants, randomly divided in equal amount to each treatment, and count the number of successful treatments. In this case, for a treatment we would have a specific number of participants N_i, where $i \in \{A, B\}$, and the number of successes for this treatment S_i would be determined by the actual value and sampling error of θ_i; $S_i \sim Binomial(N_i, \theta_i)$.

We illustrate this example in four scenarios to allow comparison with the four

cases identified by Roberts and Pashler (2000) in Figure 6.2. Specifically, we consider two hypotheses, a vague and a specific one, and two different data sets. In the first and second scenario, corresponding to the bottom panels in Figure 6.2, the hypothesis makes vague predictions: $\mathcal{H}_v : \theta_A = [0.50, 1.00]$, $\theta_B = [0.00, 0.50]$.[6] Thus, $\mathcal{H}_v$ takes up 25% of the encompassing model and it predicts that the success rates for Treatment A and B will very likely be above and below 0.5, respectively (see Figure 6.3).

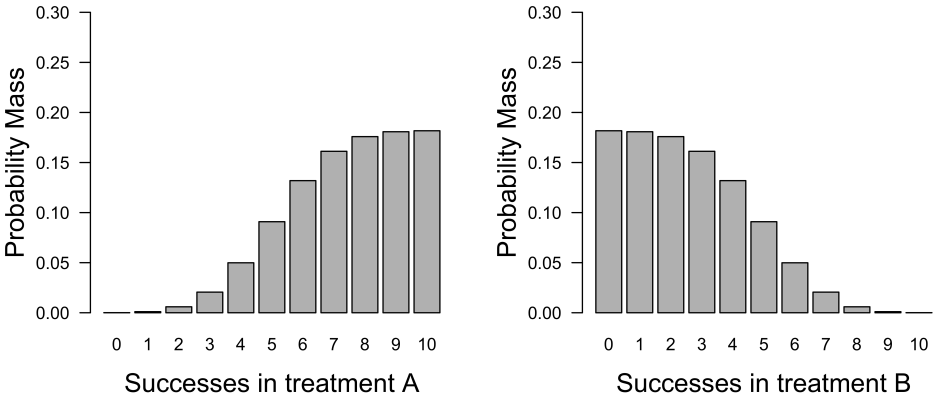


Figure 6.3: Data predictions for the vague hypothesis. If ten patients were to be tested in both Treatment A and Treatments B, then according to $\mathcal{H}_v$ we would expect to see these numbers of success with these probabilities.

In the third and fourth scenario, corresponding to the two top panels in Figure 6.2, the hypothesis makes specific predictions and strongly restricts the possible parameter values: $\mathcal{H}_s : \theta_A = [0.70, 0.80]$, $\theta_B = [0.20, 0.30]$. Thus, $\mathcal{H}_s$ takes up 1% of the encompassing model and it predicts that success rates between 0.7 and 0.8 for Treatment A and between 0.2 and 0.3 for Treatment B are most probable while success rates above and below those values are increasingly unlikely (see Figure 6.4).

These hypotheses are compared on two data sets. A small data set of four participants per treatment, $N_A = N_B = 4$ represents the two right panels of the Roberts and Pashler (2000) quartet. A relatively large data set of twenty participants per treatment, $N_A = N_B = 20$, represents the two left panels of the Roberts and Pashler (2000) quartet. If the data align well with each hypothesis (e.g., $S_A = 3/4 N_A$ and $S_B = 1/4 N_B$), the specific hypothesis $\mathcal{H}_s$ is better supported than the vague hypothesis

[6] Locations of these hypotheses are chosen for symmetry, not much changes when different locations are used.

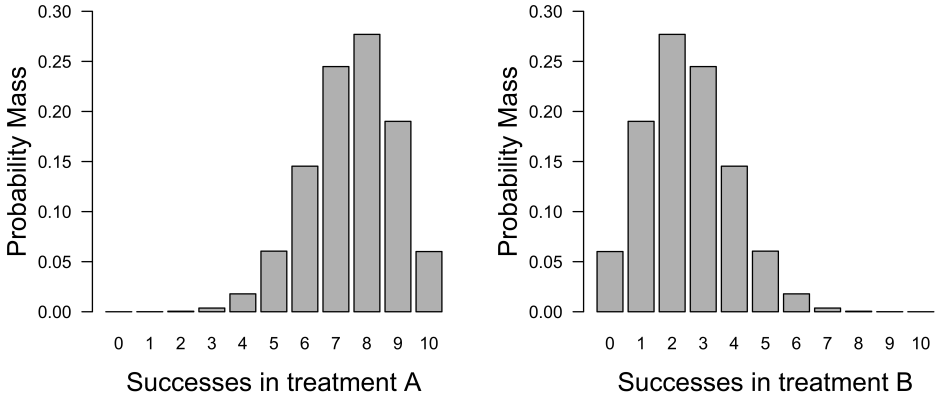


Figure 6.4: Data predictions for the specific hypothesis. If ten patients were to be tested in both Treatment A and Treatments B, then according to $\mathcal{H}_s$ we would expect to see these numbers of success with these probabilities.

$\mathcal{H}_v$. It is even the case that the specific hypothesis is better supported by the small data set ($\text{BF}_{se} = 4.37$) than the vague hypothesis is supported by the large data set ($\text{BF}_{ve} = 3.89$). As shown in Figure 6.5, this scenario reproduces the figure of Roberts and Pashler (2000).

However, the benefit of a specific hypothesis when right comes at a cost when wrong. As is visible in Figure 6.7, $\mathcal{H}_s$ only has a small amount wiggle room and the Bayes factor quickly drops towards zero when the number of success deviates from the predicted range, while $\mathcal{H}_v$ has a large area to move in with respect to possible success rates that support it.

This is also reflected in the expected evidential value of the experiment in the four situations of the example—compare Equation (6.1). As explained on page 134, the expected absolute log-Bayes factor describes the evidential value one can expect either for or against the specific hypothesis in relation to the encompassing model, where the expectation is taken over all possible data y:

$$\mathbb{E}\left[\left|\log \text{BF}_{se}\right|\right] = \int p(y)\left|\log \text{BF}_{se}(y)\right| \mathrm{d}y$$

where $p(y)$ is

$$p(y) = p(\mathcal{H}_i)p(y \mid \mathcal{H}_i) + p(\mathcal{H}_e)p(y \mid \mathcal{H}_e)$$

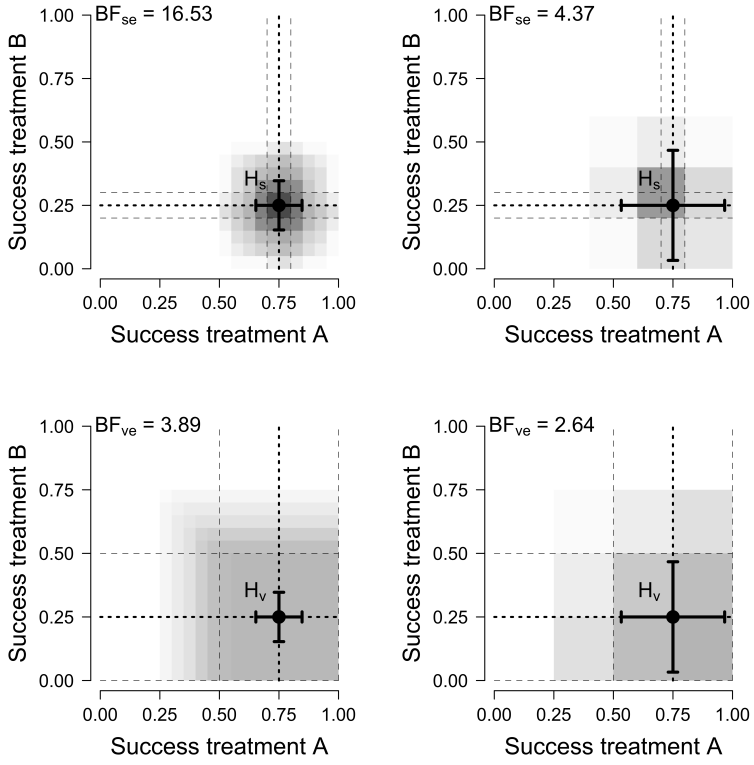


Figure 6.5: Four relations between hypothesis and data. The top two graphs show the results of specific hypothesis $\mathcal{H}_s$. The bottom two graphs show the results of the vague hypothesis $\mathcal{H}_v$. The gradient gray areas depict the probabilities mass with respect to the hypotheses' predicted outcomes (a top view of Figures 6.4 and 6.3). The dotted-lines in gray visualize the hypotheses' restrictions on the parameter values. The point estimates and standard errors of the data are visualized as black crosses. The graphs in the left column display the results of the large data set (N_i = 20) and the graphs in the right column display the results of the small data set (N_i = 4). In this example, the data align perfectly with the hypotheses. The evidential support for the hypotheses in comparison to the encompassing model is quantified as Bayes factors (top-left of each plot). From top-left to bottom-right, these Bayes factors are 16.53, 4.37, 3.89, and 2.64 respectively.

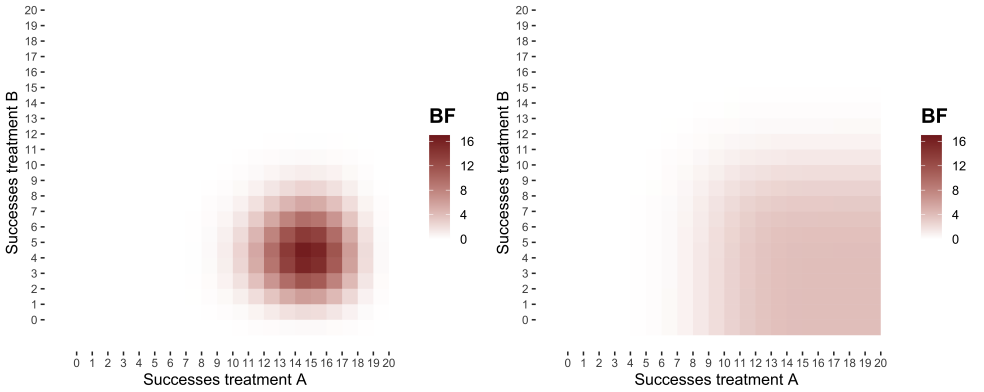


Figure 6.7: Comparison of hypotheses with respect to diverging data. The graphs displays the size of the Bayes factor with respect to data diverging from the predicted values for the specific hypothesis $\mathcal{H}_s$ (left) and the vague hypothesis $\mathcal{H}_v$ (right). The x-axis and y-axis indicate, for treatment A and treatment B respectively, the number of successful treatments (S_i) out of the 20 participants that are treated (N_i = 20). The values within the lattice indicate the evidence for the hypothesis that results from the particular combination of S_A and S_B. This evidence is quantified in terms of the Bayes factor for the hypothesis (top: specific, bottom: vague) with respect to the encompassing model. The size of the Bayes factor is also indicated as the intensity in color.

and $p(y \mid \mathcal{H})$ is

$$p(y \mid \mathcal{H}) = \int p(y \mid \theta, \mathcal{H}) g(\theta \mid \mathcal{H}) \, \mathrm{d}\theta$$

If no data has been observed yet and any proportion of successful treatments is still possible with the predetermined sample, the expected evidential value is higher for the specific hypothesis, that is, $\mathbb{E}_y\left[|\log \mathrm{BF}_{se}|\right] > \mathbb{E}_y\left[|\log \mathrm{BF}_{ve}|\right]$. The expected evidential value is highest for the specific hypothesis with the predetermined sample size of 20 per treatment, $\mathbb{E}[|\log \mathrm{BF}_{se}|] = 6.32$. This is followed by the vague hypothesis with $N_i = 20$ per treatment, $\mathbb{E}[|\log \mathrm{BF}_{ve}|] = 3.14$; the specific hypothesis with $N_i = 4$ per treatment, $\mathbb{E}[|\log \mathrm{BF}_{se}|] = 1.61$; and the vague hypothesis with $N_i = 4$ per treatment, $\mathbb{E}[|\log \mathrm{BF}_{ve}|] = 1.25$. The evidence that we obtain is on average more telling and discriminatory for the specific hypothesis (compare Etz, Haaf, Rouder, & Vandekerckhove, 2018). These numbers illustrate the value of severe testing—presupposing our explication of severity as specificity of predictions—in the framework of Bayesian inference. Examples can be generalized straightforwardly to other statistical models.

6.5. Discussion

This article presents a Bayesian proposal on how to accommodate the concept of severity when testing statistical hypotheses. It provides a translation of Popper's (2002[1959]) falsificationist philosophy and the intuitive impressiveness of risky

predictions in terms of a test of specific hypotheses. The specificity of a hypothesis is explicated by the degree to which predictions are spread out across the sample space, while evidence is defined by relative predictive performance, that is, how much probability mass is allocated to the data under the competing hypotheses. A complex or vague hypothesis spreads out its predictive mass across a wide range of options, and by hedging its bets will lose out against a more restrictive hypothesis that makes a more precise (and accurate) prediction. The more precise the predictions from the competing hypotheses, the higher the expected diagnosticity, and the more severe the test. This approach clearly quantifies Popper's idea of evaluating theories on the basis of their empirical content and degree of falsifiablity (Popper, 2002[1959]). In the specific case of a parameter space whose regions correspond to different statistical hypotheses, the specificity of a hypothesis can often be measured by the proportion of the parameter space it occupies (weighted with the prior probability density); compare our explanations in Section 6.4 on the encompassing-prior model.

Popper and Bayes can thus be reconciled: the evaluation of hypotheses in terms of Bayes factors is influenced by their specificity and Bayesian inference has the conceptual resources to reward specific predictions. Notably, obtaining this conclusion does *not* require any blending of Bayesian and frequentist inference; our account stays faithful to the principles of subjective Bayesianism.

6.5.1. Advantages of specific hypothesis testing

There are several clear advantages to testing specific hypotheses in a Bayesian framework. First and foremost, it is to the researchers' benefit to make more specific predictions. As outlined earlier, the more specific the hypothesis is, the more evidence one can expect either for or against the hypothesis: either more evidence is obtained from the same amount of data, or less data are needed for the same amount of evidence. This is in stark contrast with orthodox frequentist methods where "more evidence" requires more data if the discrepancy between null hypothesis and alternative hypothesis is fixed. In experimental psychology, the use and benefit of informed/specific Bayes factor hypothesis tests includes Vohs et al. (in press), Gronau et al. (2017), Ly, Etz, Marsman, and Wagenmakers (2019) and in particular the work of Zoltan Dienes (e.g., Dienes, 2008, 2011, 2014, 2016, 2019).[7] For cognitive

[7]Vohs et al. (in press) assign a positive-only Gaussian prior for effect size with mean .30 and standard deviation .15 – the Vohs prior. Gronau et al. (2017) assign to effect size a positive-only t-distribution with location 0.350, scale 0.102, and three degrees of freedom – the Oosterwijk prior, see Gronau, Ly, & Wagenmakers, 2020. Finally, Ly et al. (2019) outline a test to assess replication success, where the posterior distribution of the original experiment functions as the prior distribution of the replication experiment; see also Verhagen & Wagenmakers, 2014.

models, the value of informed prior distributions has been highlighted by Vanpaemel and Lee (2012) and Lee and Vanpaemel (2018).

Second, our account makes explicit why vague predictions lack diagnostic value. This is most clearly expressed in cases where (almost) everything is consistent with the theory. In such cases where all parameter values are allowed by the theory, one can never expect to obtain strong evidence either in favor or against the theory. Only when the theory is restrictive, strong confirming or undermining evidence can plausibly be expected.

Consequently, evidence is limited by the specificity of the hypothesis. Intuitively, the possible evidence in favor of a theory should be limited by its strength and how well it is tested. The specific hypothesis testing approach provides this feature, because the maximum amount of evidence in terms of the Bayes factor is one divided by the specificity of the hypothesis.

In addition, this approach can be extended to the research planning phase. Specifically, the specificity of one's hypothesis, or the restrictiveness of one's theory, can inform the researcher about the sample size required to be fairly certain of strongly supporting or contradicting results. Vice versa, in the case where the sample size is predetermined, the researcher can infer how specific the tested hypothesis needs to be in order to be fairly certain of obtaining high degrees of confirmation or disconfirmation.

Explicitly, our approach can be integrated into a Bayes Factor Design Analysis (BFDA; e.g., Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019). Given background assumptions concerning the rival models and the data-generating process, BFDA provides either the distribution of Bayes factors for a fixed sample size, or the distribution of sample sizes to obtain a fixed Bayes factor (`http://shinyapps.org/apps/BFDA/`). The use of more specific models will affect the BFDA such that the same data will become more diagnostic: a fixed sample size is projected to yield more evidence, and a given level of evidence is reached with smaller samples. Similarly, in Adaptive Design Optimization (ADO; e.g., Ahn et al., 2020; Myung et al., 2013) the next stimulus to be presented is determined by maximizing expected information gain. As the observations accumulate, the rival hypotheses become increasingly specific (i.e., their constituent posterior distributions become more peaked), and therefore easier to discriminate. Our research thus connects to the growing literature that considers Bayesian inference not only a method of hypothesis evaluation, but also uses it in planning and designing experiments that are both reliable and efficient.

Disclaimers and Conclusion

Our approach is, of course, susceptible to any objection that may be raised against Bayesian statistics as a whole (e.g., Mayo, 2010, 2018; Moyé, 2008; Senn, 2011). Our claim here is just that the epistemic value of severity is not a particularly good argument for preferring frequentist to Bayesian statistics. Severity is a notoriously elusive and hard-to-quantify concept in the frequentist paradigm, too, and the state-of-the-art explications are arguably unsatisfactory (Section 6.3.1, Appendix D). At the same time, it is possible to give a Bayesian account of the evidential value of severity (Sections 6.4.1 and 6.4.2). We hope that our approach provides a useful tool that can fill an empty spot in the researcher's toolbox of statistical methods, specifically with regard to the role of severe testing Bayesian inference. At the very least, we expect that this paper will inspire criticism, which might fallow to fruitful debate on how theories can show their probative value irrespective of the preferred statistical paradigm.

7

CONCLUSION

7.1. General Summary

The aim of this project was to contribute to the improvement of scientific practice. Both empirical research and theoretical investigations were applied to this end. The first chapters were more concerned with critical appraisal and perspective taking on the current state of the art. The last two chapters were more concerned with potential contribution to the process of scientific improvement.

In Chapters 2, 3, and 4, a review approach was adopted. In particular, statistical inference was investigated in Chapters 3 and 4. For the former, the methodological literature was systematically scrutinized to map out the purported qualities and problems with the Null Hypothesis Significance Test (NHST) method. NHST is the standard method for statistical inference in the social sciences, but also has been criticized for decades (e.g., Berkson, 1942). Over more than seventy years, many papers have been published on NHST's deficiencies, how it has been misinterpreted and misused, and how it has been responsible for many problems in science. The number of such publications has intensified since the advent of the replication crisis (Open Science Collaboration, 2015), which started with the publication of Bem's precognition paper in 2011. These essays form an ongoing debate on what the qualities and problems of NHST are and what should be done to solve these problems. However, a clear overview of this debate and the proposed solutions is lacking. To this end, we systematically reviewed the existing literature that was published between 2011 and 2018. From this body of essay literature we extracted, in the form of arguments, 70 problems with NHST and its use, 20 properties that speak to NHST's defense, 33 solutions proposed to remedy the perceived problems, and 14 conclusions. These 137 items can be condensed into three zones in the NHST landscape. The *mild zone* contains those arguments, solutions, and conclusions which maintains that the problems are not with NHST itself, but with the people that use it and how it is used. In this case, the problems are supposed to be solved by implementing policy changes and adequate safeguards. The *moderate zone* contains the acknowledgments that NHST has some serious defects, but these can be addressed by some structural changes that leave its foundation in frequentist statistics in tact. The *critical zones* encloses the deficiencies of NHST and its misuse are insurmountably severe without a whole-sale switch to an alternative method (e.g., estimation, Bayesian inference). The final contribution of this study is the data itself that was extracted from the papers. The information on the NHST's qualities, defects, and solutions has been made publicly available and is expected to be of use in the ongoing effect to understand statistical inference and its practice.

Chapter 4 provides additional insight into the actual practice of statistical inference. For this study, several teams of statisticians were provided with the same research questions and data sets. These teams were representative of the frequentist or Bayesian school of statistics and the different perspectives within these school (e.g., Bayesian hypothesis testing with default priors and Bayesian estimation with subjective priors). The teams were instructed to analyse two existing data sets from published studies in the attempt to answer the research question that were reported in these publications. The particular aim of the study was to observe differences and similarities in analysis and conclusions between the teams. The noteworthy similarities and agreements were that 1) the teams drew close to identical conclusions from the data with respect to the research questions; 2) we need better education of statistics at the undergraduate and graduate level; and 3) most statistical inference procedures, when adequately used, will provide similar results in most cases. The noteworthy differences were 1) the analysis procedures that the teams used to research their close to identical conclusions; 2) the way we learn from data (analysis); 3) the value and problems of sampling protocols in frequentist approaches; 4) prior specification in the Bayesian approaches; and 5) to what extent model assumptions should be adhered to for adequate statistical inference. These results provide additional insight into diversities and similarities in the practice of statistical inference and what could possibly be done to remedy some of the problems identified in the papers that were reviewed in Chapter 3.

A more focused gaze was used in Chapter 2 where a particular body of literature was reviewed. Machery et al. (2004) is a seminal paper in the experimental philosophy literature. It sparked an ongoing debate on the role of intuition in philosophical methodology and inspired others to replicate and extent the research. The original study reported results on the difference between semantic intuitions concerning reference of proper names, which showed a difference between Western and East-Asian participants. By some, this was considered a devastation blow to philosophy, because they considered the content of philosophical intuitions to be invariant and thus reliable as evidence for philosophical arguments. The results of this study were bolstered by the results of several replication and extension studies until a high-powered preregistered replication attempt failed (Cova et al., 2018). This was considered a curious outcome, thus we decided to take a closer look at the entire body of research. Specifically, a systematic review and meta-analysis approach was adopted to investigate the evidential value and methodological quality of the experimental philosophy literature on semantic intuitions (Chapter 2). We

included all studies conducted since the first study by (Machery et al., 2004) on semantic intuition concerning reference of proper names between people of Western and East-Asian cultures. Our results indicate that there is some difference between Westerners and East-Asians. However, these differences were small and there was much variance between studies. In addition, our assessment of the methodological quality of the studies revealed that much is either lacking or unreported in the papers. In the general sense, as this body of research is not atypical, these results can be considered as an anecdotal indication that there is room for methodological improvement in experimental philosophy.

In the last two chapters (Chapters 5 and 6), proposals were made with the aim to contribute to the improvement of scientific practice. These proposal were developed from a philosophically sound foundation and honed with the aid of empirical evidence and methodological investigations. Specifically, Chapter 5 offers an conceptualization of scientific objectivity that can be practiced by scientists, investigated on validity and reliability, and improved upon as sciences progresses. In this chapter, we assessed the notions of objectivity that are currently in vogue in philosophy of science. Furthermore, we scrutinized the literature on failed replication studies and the questionable research practices that have been identified as detrimental to the veracity of research results. This led us to adopt a via-negative approach where (approximate) objectivity is what remains after the removal or safeguarding against aspects of research that have the potential to bias results or inflate the rate of false-positive discoveries. Our conception of objectivity is one of *practice*: the study, how it was conducted, how the data were analyzed, and which/how results were reported are considered as objective (as is currently possible) if the proper safeguards are in place. The main benefit of our conceptualization is that it is easily operationalized as a checklist that can be implemented, calibrated, and improved (Appendix C).

In the last chapter, we proposed a formulation of specific hypothesis tests (Chapter 6). This test was embedded in Popper's (2002[1959]) conception of severity and the purported necessity for scientific theories to pass such tests. According to Popper, the degree to which a theory can be tested and thus falsified, depends on its *universality* (i.e., the reach of a theory) and its *precision* (i.e., the proportion of possible observations that are consistent with the theory). We evaluated both frequentists and Bayesian statistics on the extent to which they exhibit and quantify these properties in the formulation of their hypothesis test. Finding them both wanting, the former more than the latter, we proposed an extension of an existing

Bayesian hypothesis testing method using Bayes factors (Hoijtink, 2011; Klugkist et al., 2005). In particular, we explicated universality as range of observable data and precision as the width of a segment within that range that is consistent with the theory. A theories falsifiablity is quantified as the ratio of the range of theory-consistent data to the entire range of possible data (given that this is bounded and some other conditions are met; Chapter 6). The severity of a test, as the specificity of the hypothesis, was then opertationalized using Bayes factors where the evidence is quantified in terms of the ratio of goodness-of-fit of the data over the complexity of the model (i.e., the proportion of the likelihood that falls inside the hypothesized range of values over the proportion of the hypothesized range of values with respect to the entire parameter space that is considered; Hoijtink, 2011; Klugkist et al., 2005). We showed that the more precise an hypothesis is the more evidence the data provide if they supports the hypothesis or the less data are needed for the same amount of evidence. It must be noted that most theories outside the hard sciences of physics and chemistry are not advanced enough to offer the range predictions necessary to make the proposed specific hypothesis test useful. However, this approach offers the advantage of more evidential 'bang' for the researchers' data 'buck'. Thus, the availability of a specific hypothesis test such as ours might incentives researchers to develop more sophisticated theories, especially in areas where data is scarce or expensive.

7.2. Embedding of Results

This dissertation is written during a period of methodological upheaval. In 2015, the replication crisis in psychology started when the report was published on the low replication rate of seminal studies (Open Science Collaboration, 2015). This triggered great unrest in the community, a flurry of papers, development of new tools, and the birth of a new academic field; meta-science. Experimental philosophy as well was investigated on the replicability of its results (Cova et al., 2018), errors in reported statistics (Colombo et al., 2018), and the evidential value of these statistics (Stuart et al., 2019). Although, much has happened in the intervening years, there is still room for improvement. In particular, there is room for a closer collaboration of philosophy and science in this attempt at improvement.

The statistical method NHST and how it is used by scientists has received the brunt of the blame of the lack of replicability of previously assumed robust results (e.g., Colling & Szűcs, 2018). Calls have gone out to lower threshold of statistical significance (Benjamin et al., 2018), justifying the used error probability threshold

(Lakens, Adolfi, et al., 2018), and abandoning significance thresholds all together (McShane et al., 2019). Some philosophers got involved in this debate as well (on, for instance Benjamin et al., 2018), not necessarily choosing the same side. However, this has not resolved the misuse of NHST and its methodological deficiencies. In fact, one could say that these have been with us since the 1940s (e.g., Berkson, 1942) and previous attempts and change and improvement (e.g., J. Cohen, 1994) have failed. We might be at risk of repeating this history. At the vary least, a clear and thorough understanding of the practice of statistical inference is lacking, an understanding that seems to be necessary to advance and maybe resolve the debate on how to address deleterious problems. It is my opinion that many analysts studies, like the one reported in Chapter 4, will give us a better understanding of how methods of statistical inferences are actually used in the field, why certain shortcuts are taken, and maybe even how certain 'bad habits' get entrenched. Results of such studies can inform policy makers, educators, journal editors, and supervisors when quality assessment or improvement is required.

Furthermore, I expect that the systematic review of the literature on NHST (Chapter 3) and other similar reviews on research methods will prove useful in our endeavor to understand and improve scientific practice. Especially here, philosophers of science and statistics will have a substantial role to play. As is illustrated from the results of Chapter 3, such reviews can produce a staggering amount of arguments for or against a method, which are not necessarily mutually compatible or of the same quality. To make sense of such a varied collection of pros and cons, expertise with the philosophical foundations of the methods and their history might be required as well. Experience with logic, probability theory, rational choice theory, epistemology, and/or theories of inference could not hurt either. There are reasons why there is disagreement between the proposals on how to tackle the current problems with statistical inference and it is highly unlikely that they are all good proposals. In my opinion, evaluating these proposals and the offered arguments in light of the available literature, by both scientists and philosophers, will be of significant value and might even open new fruitful paths of literature research.

Philosophers cannot only contribute to but can also benefit from the investigation of science and scientists. Specifically, meta-science has been developing methods and instruments for the evaluation and improvement of scientific practice. For instance, several methods have been proposed to investigate the presence of publication bias and estimate the evidential value of the statistical results reported in a body of literature (e.g. z-curve, p-curve; respectively, Schimmack & Brunner, 2017; Simonsohn et

al., 2014). Although they have their defects (e.g., miss publication bias under certain conditions, Bishop & Thompson, 2016; Carter et al., 2019), they have proved useful on many occasions (e.g., Lakens, 2017b). Furthermore, instruments are now available to evaluate and assist with experimental design (e.g. Percie du Sert et al., 2017), academic writing (e.g. Barnes et al., 2015), and the transparency of the research (e.g., Aczel et al., 2020). The conception of objectivity reported in Chapter 5 can also be such a tool, a tool for the improvement of the objectivity of study design. Now, it is my opinion that experimental philosophy, like the research included in the meta-analysis of Chapter 2 could benefit from methodological assessment and development accordingly. Eventually, experimental philosophy should move towards assessing its methods; separately or in tandem with research to which the method is applied. I think experimental philosophy will greatly benefit from adequate experimental control, checks on data quality, and the assessment study designs and measurements in terms of their reliability and validity. I hope that experimental philosophers are taking heed of these current developments of research tools and that we can expect their first tentative implementations in the near future.

Finally, as mentioned before, currently science is beset by many problems, which have been attributed to many causes. Initially, the low rate of replicability of published results was attributed to misuse of statistical methods and publication bias (e.g., Open Science Collaboration, 2015). The proposed solution to this practice was preregistration and the reduction of publication bias by having journals review and accept a paper on its methodological quality prior to the data collection (i.e. registered reports; Chambers et al., 2015). However, according to some, this does not necessarily address the lack of measurement quality (Flake & Fried, 2019), validity (Hussey & Hughes, 2020), formal models (Szollosi et al., 2019), scientific theories (Oberauer & Lewandowsky, 2019), and the unwarranted generalization from particular results from a single experiment to a universal theoretical conclusion (Yarkoni, 2019). In these papers, allegations are levelled against the other papers for neglecting the actual cause of the problem and thus making their proposed solution either ineffective or harmful. However, there seems to be a common thread through all these papers. They all advocate the importance of methodological rigor. Specifically, they seem to propose that 1) theories need to be explicitly stated to 2) allow the specification of formal models and 3) the assessment of quality and validity of the measurement to be used; and 4) a rigorous derivation from the theory to what is to be tested in order to gain a supporting or refuting outcome. Finally, preregistration adds transparent control to this methodological

rigorousness. As some have observed (e.g. Lakens, 2019b; Yarkoni, 2019), this bears resemblance to what Popper called a severe test (Popper, 2002[1959]). Some or all of the proposed solutions can even be argued for through Popper's arguments for the necessity/value of severe tests of scientific theories. Unfortunately, Popper never explicated the concept of severity. Although, as these current developments appear to show, such an endeavour is underway, implicit in 'crisis' papers and explicit in Chapter 6, and such an explication could prove useful to the development of scientific practice. Specifically, we should not only look at the quality of research results, but at the quality of the methods that produced them. All these 'crisis' papers are concerned from various perspectives with how adequate tests are. Severity can be an overarching principle to explicate the quality of a test vis a vis a theory or hypothesis. It is too early to tell where this literature will take us, but it could be that those with philosophical expertise can make a meaningful contribution to the crises debate and the crystallization of severity.

7.3. General Limitations

Apart from the particular limitations mentioned in each chapter, there are some general limitations worth mentioning. First and foremost, all work presented in this dissertation is susceptible to human error. All of sciences is (at least at the moment) a human endeavour and thus error prone, but it many cases errors are mitigated by contingencies and checks. However, for instance, the choices made on how to include and analyse data for Chapters 2 and 3, are profoundly consequential, are mostly my own. Others are thus invited to attempt to reproduce and replicate the work presented in this dissertation.

Secondly, part of this dissertation (Chapters 5 and 6) rests mostly on conjecture. Even if the usefulness of the provided conception of objectivity and specific hypothesis tests is well argued for, empirical evidence to support is still lacking. The defects, weaknesses, and applicability is still to be investigated. Hopefully, the proposals are considered useful enough for others to put them into practice and carry out these investigations.

Finally, the dissertation is somewhat lacking in integration with general philosophy of science. The work presented here is geared towards the practice of sciences. Some chapters (Chapters 5 and 6) have their foundations in philosophy of science, but they are written with practice by scientists in mind. Other chapters (Chapters 3 and 4) are about scientific practice (i.e., statistical inference). Though these are not founded in philosophy of science these might be of interest for further

philosophical investigations.

7.4. Implications, Speculations, and Suggestions

Some general suggestions and speculations can be drawn from the work presented in this dissertation. First and foremost, I would like our work to be used and tested by others. Our negative notion of objectivity and the formulation of specific hypothesis tests are ready to be put into practice, under regular and extreme conditions, to investigate their applicability and robustness. The social sciences might be still far removed from formulating theories specific enough to include functional forms and range predictions on parameters, but some initial test could be performed on well-established effects and theories (e.g., memory retention; Wixted & Ebbesen, 1991). The negative notion of objectivity is ready to be implemented as a checklist and thus tested on reliability. In addition, validity can be tested on the many replication studies that are underway or will be carried out in the future. Furthermore, the results of systematic review on NHST literature could prove useful in future attempts at resolve the debate(s) on statistical inference or at least inform the participants on the existing arguments and solutions. It could also aid educators adapt their curriculum to address identified flaws and accommodate proposed solutions.

Secondly, if (future) researchers are to take away something from this dissertation, I hope it is the realization that we need adequate quality control. For instance, the meta-analysis presented in Chapter 2 teaches us at least that experimental philosophy needs methodological evaluation in terms of validity and reliability. Currently, it appears that a culture of methodological introspection beyond statistical power and replicability is lacking in the social sciences. Also, where methodological appraisal tools exist, they are not applied with regularity and might not keep pace with the development of research methods they are supposed to appraise. Much is to be gained in this area and I think philosophers have a role to play in their conception or development as they are at least partly founded on principles and logic besides empirical evidence.

Thirdly, it might be worthwhile to deviate from perfection and consider good-enough approaches. Instead of aiming for the unattainable (e.g., uncontroversial definitions or operationalisations), we should focus on those things that might work moderately well or slightly better than completely terrible, and try to make incremental but robust improvements. Instead of trying to invent new measures of evidence or trying to get scientists to adopt a method that is completely new to them, we should focus our attention on what can be improved on the conventional

methods and practices.

Fourthly, a pragmatic perspective on the frequentist versus Bayesian statistics debate could resolve some of the outstanding points of conflict. Ever since the birth of modern frequentist statistics[1] there has been a continuous disagreement between frequentists and Bayesians. For instance, Fisher (1922) argued against inverse probability, while Jeffreys (1939) staunchly defended it. One of the main arguments against frequentist statistics is its reliance on unobserved data. Jeffreys made this point succinctly when he wrote "[a] hypothesis that may be true may be rejected because it has not predicted ... results which have not occurred" (Jeffreys, 1939, p.385). One of the main arguments against Bayesian statistics is its subjectivity; the personal believes, expectations, and discretionary choices of the researcher that he/she adds to the analysis as prior. As Efron puts it, "[s]trict objectivity is one of the crucial factors separating scientific thinking from wishful thinking .... The high ground of scientific objectivity has been seized by the frequentists" (Efron, 1986, p.4). A noteworthy counter Bayesian argument is that Bayesian statistics is upfront about this subjectivity (Wagenmakers et al., 2008) while frequentist analyses suffers from subjectivity as well (e.g., choice of sampling plan), but this is locked away in the heads of the researchers (e.g., Berger & Wolpert, 1988). Furthermore, Bayesian statistics have recently been accused of lacking quantification or qualification of the probativeness of its tests, the stringency with which its hypotheses are tested (i.e., the severity of the falsification attempt; Mayo, 2018). At the time of writing, this debate is far from resolved and papers are still being published that promote the one and/or vilify the other.

However, some have wondered as to what extend is this debate academic instead of having real-life consequences for practicing scientists and the value of their research results. On the one hand, this question can be answered by empirical research on scientific practice and the statistical methods. Specifically, many analysts studies, such as the one presented in Chapter 4, provide inside into the effect of different statistical approaches on research results. Also, systematic reviews of methodological literature, such as the one presented in Chapter 3, can shed light on the extend of perceived methodological problems with respect to research practice. On the other hand, we can try to *ensure* that the debate is academic by developing tools and safeguards that address the identified problems. For instance, the lack of probativeness of Bayesian statistics can be addressed by proposed solutions such as the specific hypothesis test (Chapter 6). Also, both the hidden subjectivity of

[1] I think it is safe to say that modern frequentist statistics as we know it started with Fisher (1925).

frequentist statistics and the overt subjectivity in the Bayesian prior can be addressed by setting objectivity standards and safeguards on the application of statistical analyses, as is suggested in Chapter 5. It should be noted that, even if the problems discussed in the frequentist versus Bayesian statistics debate have no direct practical consequences for empirical science, the continuation of this conflict seems to make meaningful contribution to the advancement of statistics and research methodology.

Finally, effort could be allocated to further develop Popper's (2002[1959]) concept of severity. More than a few social scientists have expressed interest in the concept through Mayo's effort (1996; 2018). It is considered a useful concept that clarifies several aims and intentions in science (Lakens, 2019b). However, Mayo's (2018) formulation and the formulation presented in Chapter 6 are only operationalized for the test of statistical hypotheses. They ignore or leave vague everything that precedes the formulation of the statistical hypothesis and consider the data as given. In my opinion there is room for extending the operationalization of the concept to the design of experiments, development of measurements, and the formulation of substantive hypotheses and theories that inform these measurements and experiments.

7.5. Additional Musings

Some further concluding thoughts concern the methodological awareness in philosophy and science. Before I started this PhD project, I was under the impression that academics were at least aware of the methods they were using. Scientists and philosophers might disagree on their results, might even use different methods, but at least they know what they are doing and how they are doing it. Or so I thought. I came to realize that this is not actually the case or at least not completely. For instance, the experimental philosophers assumed that in philosophy one generally uses the *method of cases* (Machery, 2017) were thought experiments are used and their results are considered as evidence for against some theory, claim, or general principle. They attacked this method by showing that the results of thought experiments vary between peoples and between contexts irrelevant for the thought experiment (Sosa, 2007). However, according to others, this is not actually what (most) philosophers do. According to them, most if not all philosophers actually use argumentation as their method; each product is a collection of arguments and a logical conclusion that follows from them. In this case, the thought experiment is considered a rhetorical device and the problems identified by the experimental philosophers do not (have to) detract from the quality of the argumentation (Cappelen, 2012; Deutsch, 2015). This

puzzled me. I could not understand how such lack of methodological self-awareness was possible.[2]

My expectations were that things were different in the empirical sciences. There they might disagree on what to research and how meaningful results were, but at least there would be no confusion about, for instance, research design and statistical methods. Specifically, there would be no disagreement on whether or not someone was using a randomized control trial to test something and used an ANOVA to analyse the data. Although I might have been right in this assumption, it appears there is still room for disagreement and confusion about what scientists do and want. Namely, there seems to be confusion about philosophy and philosophical principles. According to some (e.g., Yarkoni, 2019) most scientists use induction to draw inferences from the statistical results of the sample to the population. This is contradicted by others (e.g., Lakens, 2020) who says that actually a hypothetico-deductive method is commonly used to test general hypotheses on a sample via statistical analysis. There also seems to be confusion about falsificationism. Some (Lemoine, 2019) argue that a naive and wrong version of Popper's principles (2002[1959]) is being used to the extent some physicist think that because their hypothesis about subatomic physics is falsifiable it is also scientific (Hossenfelder, 2017). This denied by others (Lakens, 2020) who state a more sophisticated version of falisificationism is being used that is well grounded in the statistical school of Neyman and Pearson. There also appears to be confusion about whether, as scientists, we aimed at prediction or explanation (Hossenfelder, 2020) or what the principles and values are on which preregistration is based (e.g., Lakens, 2019b; Szollosi et al., 2019). In brief, the situation in empirical science does not appear much different from that of philosophy.

There seems to be room for improvement. By bringing these conflicts and confusions into the limelight, methodological self-awareness can be fostered. The replication crisis triggered a movement of methodological development. The results of experimental philosophy also send shockwaves through the philosophical world, triggering debates, counter studies, and quality assessments. My suggestion is that some time from these endeavours should be relegated to reflection on the reasons and principles behind the methods we are using and trying to improve and explicate those in tandem with our methods. A more systematic approach to this self-reflection and improvement might be warranted.

[2]It needs to be noted that these musings are based on anecdotal opinions. These scientists and philosophers make general claims what their peer do or intent to do without substantive proof that this is actually the case. It could of course be the case that both sides of these debates are actually more-or-less right, because there is substantial methodological heterogeneity.

7.6. Concluding Remarks

In this dissertation, I have attempted to bring science and philosophy of science closer together. We seem to find ourselves in a virtuous cycle without end where we constantly evaluate and amend the method we employ during the endless process of investigation we call science. My only hope is that the research reported in this thesis will make some contribution to the angle with which this cycle is moving upwards.

A

APPENDIX TO CHAPTER 2

A.1. List of Excluded Studies

Table A.1 lists all studies in our original pool that were excluded from our meta-analysis. For each study, we specify why it did not meet the inclusion criteria. Since meta-analyses are based on effect-sizes of the results of all included studies and their standard error, the same kind of effect-size must be calculable for each study in a meta-analysis. The original study used two groups, one of "Westerners" and one of "East Asians," and binary-outcomes for each probe. Hence, the relevant kind of effect-size is the Relative Risk or Odds Ratio. Based on this kind of effect-size, we excluded studies that had outcomes with more than two values and studies or samples that did not fall in the East-Asian or Western group.

Table A.1: List of excluded study on semantic intuitions.

APA reference	Exclusion reasons
Nichols, S., Pinillos, N. Á., & Mallon, R. (2016). Ambiguous reference. *Mind*, 125(497), 145-175.	The design did not match. The studies in this paper did not use binary-response options for their vignettes where one was the causal-historical option and the other the descriptivist option.
Domaneschi, F., Vignolo, M., & Di Paola, S. (2017). Testing the causal theory of reference. *Cognition*, 161, 1-9.	The design did not match. The studies in this paper did not use binary-response options for their vignettes where one was the causal-historical option and the other the descriptivist option.
Genone, J., & Lombrozo, T. (2012). Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology*, 25(5), 717-742.	The design did not match. The studies in this paper did not use binary-response options for their vignettes where one was the causal-historical option and the other the descriptivist option.
Grau, C., & Pury, C. L. (2014). Attitudes towards reference and replaceability. *Review of Philosophy and Psychology*, 5(2), 155-168.	The ethnicity/cultural background of the participants was mixed. The data could not be separated into Western and East Asian samples.
Islam, F. (2017) *Logical semantics in a cross-cultural perspective*. Master's thesis in English Linguistics and Language Acquisition. Supervisor: Giosué Baggio. NTNU - Trondheim, May 2017	The author did not want to share data. The Asian participants were not East-Asians, but South-Asians.
Machery, E. (2012). Expertise and intuitions about reference. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, 27(1), 37-54.	The ethnicity/cultural background of the participants was mixed. The data could not be separated into Western and East Asian samples.

A.2. List of Experimental Design Deviations

Table A.2 below provides an overview of experiment design deviations from Machery et al. 2004. We focused on several design factors. For each factor, we identified deviations from the original design, based on the information reported in the method sections of all studies in our data set.

Table A.2: Overview of experimental design deviation and where they are present.

Factor	Values	Description	Studies that contain this deviation
Probe	Godel1, Godel2, Jonah1, Jonah2, Shake-speare1, Julius1, Julius2, Satyricon1, Satyricon2, Satyricon3	Different vignettes were used in the studies. The original study used the Gödel and Jonah vignettes. The Shakespeare, Julius, and Satyricon vignettes are about different characters and situations, though still concern reference of proper names.	Lam, 2009 (Shakespeare1, Julius1, Julius2); Machery, 2010 (Shakespeare1); Beebe, 2015 (Satyricon1, Satyricon2, Satyricon3)
Language probe Asia	English, Chinese, Japanese	In the original study, the probes were in English for the East-Asian sample. Some studies deviated by presenting the probes in Chinese or Japanese.	Lam, 2009 (Chinese); Machery, 2010 (Chinese); Machery, 2015 (Chinese); Sytsma, 2015 (Japanese); Kazaki, 2017 (Japanese); Izumi, 2018 (Japanese)
Language participant West	English, French, Dutch, Mixed	In the original study, the main language of the Western sample was English (USA based). Some studies deviated by using Westerners with a different place or origin, residence, and main language.	Machery, 2009 (French); Cova, 2018 (Dutch); Colombo, n.p. (Mixed)

Continued on next page

Table A.2: Continued from previous page

Factor	Values	Description	Studies that contain this deviation
Language participant asia	Chinese, Mongolian, Japanese, Mixed	In the original study, the (main) language of the East-Asian sample was Chinese (Hong Kong based). Some studies deviated by using East-Asians with a different place or origin, residence, and main language.	Machery, 2009 (Mongolian); Sytsma, 2015 (Japanese); Kazaki, 2017 (Japanese); Izumi, 2018 (Japanese); Colombo, n.p. (Mixed)
Moral valence	0=original, 1=good, 2=bad	The moral valence of the vignette was altered. E.g., in the original version Gödel stole Schmidt's incompleteness theorem, but in the altered version he took credit to assist Smith.	Beebe, 2015
Perspective probe answer	0=original, 1=protagonist's perspective, 2=narrator's perspective, 3=clarified narrator's perspective	The phrasing of the questions and the binary-response options are altered. Specification is added to guide the participant in taking the perspective of the one narrating the vignette (2, 3). The 'protagonist perspective' (1) is not supposed to test semantic intuitions. These probes were excluded from the analyses.	Sytsma, 2011; Beebe, 2015 (only 3, 'clarified narrator's perspective'); Machery, 2015 (only 3, 'clarified narrator's perspective'); Sytsma, 2015; Beebe, 2016 (only 3, 'clarified narrator's perspective')

Continued on next page

Table A.2: Continued from previous page

Factor	Values	Description	Studies that contain this deviation
Anchoring	0=unaltered, 1=changed anchoring in the probe.	The vignette is altered in its anchoring. Specifically, in the original the protagonists learns that, for instance Gödel proved the incompleteness theorem, but further in the vignette this is said to be incorrect. In the vignettes with changed anchoring the vignette is rearranged in such a way that it starts with that Schmidt proved the incompleteness theorem.	Beebe, 2016
Answer phrasing	0=original, 'with bare noun phrases', 1='without bare noun phrases', 2=added 'sono-demonstrative', 3=added 'sono-demonstrative' and 'anaphoric predicate', 4=added 'clarifying predicate'.	The answers phrasing is changed in particular ways for the purpose of decreasing ambiguities. Values 1, 2, and 3 concern particular in Japanese translations responses. Value 4 concerns added predicate(s) that are supposed to add clarity.	Machery, 2009; Kazaki, 2017; Izumi, 2018

Continued on next page

Table A.2: Continued from previous page

Factor	Values	Description	Studies that contain this deviation
Question target	0=reference ("talking about"), 1=truth values (sentence true?)	The type of binary-response options that were used. The original used a reference style where the protagonist of the vignette is talking about the one bearing the name (causal-historical option) or the one with the descriptive property (descriptivist option). The other type is a true-or-false question, where one is the causal-historical option and the other the descriptivist option.	Machery, 2009
Asian sample	0=students from the University of Hong Kong, 1=other populations	This variable was added later, because we realized that it might be possible that the students from the University of Hong Kong might be a highly specialized group that might respond differently than those from the other university in Hong Kong, mainland China, Japan, or other parts of East-Asia.	Lam, 2009; Machery, 2009; Machery, 2010; Machery, 2015; Sytsma, 2015; Beebe, 2016; Kazaki, 2017; Cova, 2018; Izumi, 2018; Colombo, n.p.

A.3. Detailed Quality Appraisal

In addition to the evaluation of the main results of the meta-analysis (see Section 2.3.4), we assessed the quality of the methods and measurements techniques of the included studies. Quality appraisal is a standard feature in systematic reviews (e.g., exhaustive meta-analysis of the literature on a particular topic or field of study; Petticrew & Roberts, 2006). In general, results of a methodological quality appraisal are used for deciding on study inclusion/exclusion (e.g., studies below a certain quality threshold are excluded), or as predictors for assessing the robustness of the results. Since no standards of quality appraisal have yet been formulated and tested for experimental philosophy research, we moved the quality appraisal to this appendix where we indicated presence or absence of methodological features of the studies without interpreting these results.

For the quality appraisal of the studies' methods, we used parts of a validated appraisal checklist for clinical non-RCT comparative studies (Deeks et al., 2003; Khan et al., 2009). In addition, we added three items concerning generalizability from the sample to the population suggested in Petticrew and Roberts 2006.

The checklist consisted of nine particular qualities, on which each study could be scored 'yes' if the information provided in the paper describing the study indicated it possessed this quality; 'no' if there was clear indication in the paper that the study did not possess this quality; 'unknown' if insufficient information was given in the paper about this quality; or 'NA' if the quality was not applicable to the study (e.g., group matching does not apply to a study that tests only a single group). These nine items were:

1. *Sample description*: adequacy and completeness of description of the relevant participants' characteristics. We restricted ourselves to age (mean and standard deviation) and sex (percentage of males/females) because the original study reported these characteristics and the literature does not identify other relevant characteristics.
2. *Matched groups*: similarity of participant groups in terms of characteristics that may affect the outcome (including socio-economic characteristics). In this case, the Western and East Asian participants would have to be matched on age and sex ratio.
3. *Equal group treatment*: identical treatment of participant groups. In this case, Westerners and East Asians should be tested under the same experimental conditions.
4. *Experimental description*: adequate and unambiguous description of the exper-

iment. In this case, the studies should contain information on what kind of design was used (within-subject or between-subject), how participants were sampled, and how the participants were tested.

5. *Adequate measure*: a measurement or test is adequate when its validity and reliability are reported. This can be ensured by reference to previous measurement assessments and/or argumentation supporting its validity and reliability.
6. *Relevant measure*: a measurement or test is relevant when it is considered appropriate for testing (answering) the reported hypothesis (research question). In this case, an explanation should be given why the probe and the binary question measurement are suitable for measuring intuitions about the reference of proper names (in comparison to other potential tests or measurements).
7. *No experimenter bias*: Experimenter bias is considered absent when the report indicates that the person(s) administrating the test could not have influenced the outcome. In our case, we consider absence of experimenter bias when the administrator(s) was (were) not in the same room as the participants during the experiment.
8. *Random sampling*: a sample is considered a random sample (of the population) when all members of the population had an equal probability of participating. This is a very demanding requirement. However, it should be noted that it is an assumption of all inferential statistical tests. In our case, at least clear non-random sampling practices should be absent (e.g., convenience sampling).
9. *Representative samples*: a sample is considered representative when there are clear indications that its composition/distribution of the relevant characteristics is similar to that of the population. In our case, the samples of Western and East Asian participants should be demographically similar to the respective populations.

The methodological quality assessment was performed by two of the authors [NvD and MC]. Two of us independently scored the included studies using the quality appraisal checklist. The results of the assessment are reported in Figure A.3 and A.4. None of the studies scored affirmatively on all of the methodological qualities, and in many cases there is not enough reported on the measurement and procedure of the studies to adequately infer absence or presence of a quality.[1]

Due to the low number of high-scoring studies, it was not possible to perform a sensitivity analysis where we compare the effect-size and heterogeneity of high-

[1] The original statistical analysis by Machery et al. (2004) is flawed on several levels, most egregiously in the aggregation of binary scores across probes and the use of the *t*-test for comparing the sample means across populations.

scoring vs. low-scoring studies. However, it was possible to run regression analyses with the qualities as predictors. Five of the nine qualities could not be included, because none scored affirmatively. The four predictors that could be included in the analyses explained between 14.0% and 34.30% of the variance between the probes (for code and results: see the additional materials file).

In an effort to improve the replicability, and cross-disciplinary comparability of this line of research on semantic intutions, we encourage researchers to consider quality assessment schemes from other disciplines (e.g., in medicine), in addition to other checks on research quality, like high-powered replication studies (e.g., Cova et al., 2018), checking for statistical reporting errors (e.g., Colombo et al., 2018), and meta-science tools for the uncovering publication bias and questionable research practices (e.g., Stuart et al., 2019).

Table A.3: Assessment of methods and measurement qualities 1 through 4.

First author	Year	Study No.	Sample description	Matched groups	Equal group treatment	Experimental description
Machery	2004	1	no	no	yes	yes
Lam	2009	1	no	no	yes	yes
		2	no	no	yes	yes
Machery	2009	1	yes	no	unknown	no
Machery	2010	1	no	unknown	unknown	no
		2	yes	no	NA	yes
Sytsma	2011	1	yes	NA	NA	yes
		2	yes	NA	NA	yes
		2	yes	NA	NA	yes
		3	yes	NA	NA	yes
Beebe	2015	1	yes	no	yes	yes
		2	yes	NA	NA	yes
		3	yes	NA	NA	yes
		4	yes	NA	NA	yes
Machery	2015	1	yes	no	unknown	yes
		2	yes	no	unknown	yes
Machery	2015	3	yes	NA	NA	yes
Sytsma	2015	1	yes	no	no	yes
Beebe	2016	1	yes	no	yes	yes
		2	yes	no	yes	yes
Izumi	2017	1	yes	NA	NA	yes
		2	yes	NA	NA	yes
Kazaki	2017	1	no	NA	NA	yes
Cova	2018	1	yes	no	no	yes
Colombo	n.p.	1	NA	unknown	yes	NA

Table A.4: Assessment of methods and measurement qualities 6 through 9.

First Author	Publication Year	Study Number	Adequate Measurement	Relevant Measurement	No experimenter bias	Random Sampling	Representative Sample(s)
Machery	2004	1	no	unknown	unknown	unknown	no
Lam	2009	1	no	unknown	unknown	unknown	no
		2	no	unknown	unknown	unknown	no
Machery	2009	1	no	unknown	unknown	unknown	unknown
Machery	2010	1	no	unknown	unknown	unknown	no
		2	no	unknown	unknown	unknown	no
Sytsma	2011	1	no	unknown	unknown	unknown	no
		2	no	unknown	yes	unknown	unknown
		3	no	unknown	yes	unknown	no
		4	no	unknown	no	no	unknown
Beebe	2015	1	no	unknown	yes	unknown	no
		2	no	unknown	unknown	unknown	no
		3	no	unknown	yes	unknown	no
		4	no	unknown	yes	unknown	no
Machery	2015	1	no	unknown	unknown	unknown	no
		2	no	unknown	unknown	unknown	no
Machery	2015	1	no	unknown	unknown	unknown	no
Sytsma	2015	1	no	unknown	unknown	unknown	no
Beebe	2016	1	no	unknown	yes	unknown	no
		2	no	unknown	yes	unknown	no
Izumi	2017	1	no	unknown	unknown	unknown	no
		2	no	unknown	unknown	unknown	no
Kazaki	2017	1	no	unknown	yes	unknown	unknown
Cova	2018	1	no	unknown	no	unknown	no
Colombo	n.p.	1	no	unknown	yes	unknown	no

B

APPENDIX TO CHAPTER 3

B.1. Search Strategy

As stated in the research question, the systematic review was limited to psychological science and methodological literature. We choose PsycINFO and Web of Science Core as databases to search. In general, we used the same search strategy for both databases. However, differences in technical details between both databases forced us to adopt two slightly different search strategies:

Search strategy PsycInfo (via EBSCOhost)

- searched in: 'title' and 'keyword'
- limited to: 'articles'
- in: 'English'
- for the following words connected by an OR operator:
    - null hypothesis
    - hypothesis test
    - statistical significance

- significance test
- p value AND[1] "statistic*
- p-value AND "statistic*

Search strategy for Web of Science Core (via EndNote X8)

- searched in: 'key terms' (title / abstract / keywords)
- limited to 'articles'
- in 'English'
- for the following words connected by an OR operator:
  - null hypothesis
  - hypothesis test
  - statistical significance
  - significance test
  - p-value
  - p value
- A second search in EndNote X8 on these entries was necessary to identify those with titles or keywords (excluding abstracts) that contained one or more of the following terms:
  - null hypothesis
  - hypothesis test
  - statistical significance
  - significance test
  - p-value
  - p value

Papers were excluded from the review if 1) they were about empirical studies; 2) they were published in a journal not pertaining to psychological science, statistics or research methodology; or 3) lacked a description of one or more problems with / quality of NHST, or NHST related concepts in the abstract.

Papers about the weakness and drawbacks of NHST are a subclass of papers about NHST in general. Terms that are related to or identify NHST are 'null hypothesis', 'hypothesis test', 'statistical significance', 'significance test', 'p-value', and 'p value.' However, papers that contain reports of quantitative research can include these terms (e.g., "the results were significant [p<.05]") without actually being about those concepts. We choose to limit our search to titles and keywords of paper, because it is conventional that the subject of the paper is mentioned there.

[1]The 'AND' operator between 'value' and 'statistic' was necessary, because EBSCOhost identifies 'p' as a stop word and ignores it in the search (which turns up very many results on emotion research).

B.2. Network Analysis

B.2.1. Network analysis methods

We performed network analyses (Borgatti, Everett, & Johnson, 2018) on the interconnection of the items (i.e., misconceptions, deficiencies, NHST defenses, solutions, conclusions, and the individual pieces of evidence that support them). We setup the network in two steps. First, we took the arguments, solutions, pieces of evidence, and conclusions as *nodes* (i.e., the basic units, points, that make up the network). Second, we determined the *co-occurrence* of arguments and of an argument and its corresponding conclusion in each article as indicator for an *edge* in the network. Co-occurrence means that the items in question appear in the same paper. An edge is a connecting line between two nodes in the network. Edges between pieces of evidence were specified if they co-occurred as evidence for that particular argument in that particular paper. For the analysis, this grand network was split into two subnetworks. In the first subnetwork, the pieces of evidence were disconnected, leaving only the arguments, solutions, and conclusions. The second subnetwork used only the extracted pieces of evidence, arguments, and solutions as nodes, disconnecting the conclusions.

These networks were analyzed in terms of the *centrality of the nodes* and *clustering based on their interconnectedness*. Centrality of the nodes can be defined as the prominence of their position in the network (Borgatti et al., 2018). For instance, how well connected a node is or how close it is to the other nodes. Clusters are identified as groups of nodes that have better connectivity with other nodes in the cluster than with nodes outside the cluster. We compared the scores on centrality measures against number of papers in which items occur. This procedure allows us to get additional information about how prominent arguments, solutions, conclusions, and pieces of evidence are besides their prevalence across papers. In addition, we identified *clusters* in the network in order to investigate the grouping of arguments, solutions, conclusions, and evidence based on how they are connected to each other across the papers. More specifically, we use an algorithm that identifies clusters that maximize the *modularity:* The ratio between the number of edges connecting members of the cluster and the number of edges connecting members of the clusters to nodes outside the cluster (Schuetz & Caflisch, 2008). This means that this algorithm detects groups that closely cohere in terms of in-group versus out-group connectivity (i.e., the number of connections between nodes of the cluster against the number of connections to nodes outside the cluster).[2]

[2]The code for this analysis is provided with the online supplementary materials at `https://osf`

B.2.2. Network Analysis results

A few things can be noted from the entire network. We obtain the following descriptive features of the interconnectivity of elements we extracted from the papers. The network contains 645 nodes, which have 4667 edges between them. This means that on average each node is connected to 7.24 other nodes. If every node were connected to every other node, the maximum connectivity, the network would have 2076904 edges (645*(645-1)/2, because the network is undirected), meaning that the network has 2.25% of all possible connections between the nodes. This is not surprising, because we did not allow edges between pieces of evidence and conclusions (evidence supports arguments and arguments support conclusions). The diameter - the longest geodesic distance (i.e., shortest path between two nodes) - of this network is 25. This indicates that any node can be reached from any other node in 25 steps or less. The shortest geodesic distance, the radius of the network, is 4. There is at least one node that can reach all other nodes in four steps, but not less. The network is too large to be informative as a static display on a page. An interactive version of the network can be found on with the online supplementary materials.

Arguments, solutions, and conclusions

If we exclude the evidence, we are left with a network that contains only the conclusions that were present in the publications and the arguments and solutions that were connected to them. This reduced network contained 137 nodes (11 conclusions and 124 arguments and solutions) and 1407 edges between them, which give the network an edge density of 0.151 (the network is undirected, so there are 137*(137-1)/2 = 9316 possible edges between the nodes). In this restricted network, all nodes are allowed to be connected to each other, which, in combination with the significantly lower number of nodes, explains the increase of the presence of 15.1% (from 2.25%) of all possible edges between the nodes. The network has a diameter of 5 and a radius of 3.

In general, the centrality measures nearly completely overlap (see Table B.1). Only betweenness seems to deviate considerably from the others. The centrality scores of each item have been added to the data set provided in the online supplementary materials.

Clusters were detected in the network with the modularity optimization algorithm (Schuetz & Caflisch, 2008). We detected 6 clusters, the largest of which

`.io/ayf56/`. Readers are invited to investigate other centrality measures and clustering algorithms.

Table B.1: Pearson correlation coefficients in between different measures of centrality.

	Degree	Betweenness	Closeness	Eigen value	Page rank
Degree	1	0.64	0.9	0.97	0.97
Betweenness	-	1	0.56	0.52	0.73
Closeness	-	-	1	0.83	0.81
Eigen value	-	-	-	1	0.93
Page.rank	-	-	-	-	1

consisted of 34 items and the smallest of 8 items. The cluster identification of each item has also been added to the available data set.

Arguments and evidence network

If we exclude the conclusions from the original network we are left with a subnetwork that contains the arguments, solutions, and the pieces of evidence that were used to support them in the papers. This subnetwork actually consists of three separate networks (i.e., they share no connections between them). The largest of the three contained 623 nodes (502 pieces of evidence and 121 arguments and solutions) and 4425 edges between them. This gives the network and edge density of 0.028. The two other networks contain 2 (1 piece of evidence, 1 argument) and 7 (6 pieces of evidence and 1 solution) nodes and 1 and 21 edges respectively. We excluded these small networks from the rest of the analysis.

In general, there is much overlap of the centrality measures. Only betweenness considerably deviates from the other measures and closeseness is only moderate correlated with page rank (see Table B.2). The centrality scores of each item have been added to the available data set.

Table B.2: Pearson correlation coefficients in between different measures of centrality.

	Degree	Betweenness	Closeness	Eigen value	Page rank
Degree	1.00	0.73	0.71	0.92	0.93
Betweenness	-	1.00	0.52	0.72	0.84
Closeness	-	-	1.00	0.72	0.62
Eigen value	-	-	-	1.00	0.91
Page.rank	-	-	-	0.91	1.00

C

APPENDIX TO CHAPTER 5

C.1. Setup of a Checklist for Objectivity Assessment

Given the negative nature of our notion, our checklist consists of questions assessing how susceptible the study is to suffer from the mentioned problematic practices. Concretely, a checklist consists of yes-no questions that indicate the presence or absence of features that prevent problematic practices.

Questions concerning the study being bias-resilient:[1]

1. Was (were) the outcome measure(s) directly related to the phenomenon of interest as stated in the research aim or research question? (e.g., 'death rate' to 'death by cardiac arrest')
2. Was (were) the intervention(s) clearly related to phenomenon of interest as stated in the research aim or research question? (e.g., 'cardiac arrest reducing medication' to 'death by cardiac arrest')
3. Was sampling procedure random?

[1] These questions pertain only to experimental research. However, the questions can be adapted for observational research.

4. Was the sampling procedure capable of producing a sample representative of the population? (i.e., do inclusion/exclusion criteria allow all member of the population)
5. When the subjects are volunteers, was the (non-)response rate similar across participant characteristics? (e.g., equal between men and women)
6. Was the allocation of the subjects to the experiment conditions random?
7. Were both the experimenter and the subjects blind to the experiment condition?
8. Was the drop-out rate of subjects similar across the experiment conditions?
9. If any answer to these question was 'no', were proper steps taken to ameliorate the potential bias that could have resulted from it?

Questions concerning the study being resilient against bias and false-positive rate inflation due to questionable research practices:

1. Was the study preregistered?
2. If so, was the following specified in the preregistration:
   (a) Management of missing and incomplete data.
   (b) Pre-processing of data (e.g., how to clean and normalize).
   (c) Data processing and dealing with violation of statistical assumptions.
   (d) Management of outliers.
   (e) Statistical analysis/model.
   (f) Dependent variable(s) of the model.
   (g) Predictors/covariates of the model.
   (h) Estimation method and computation of standard errors.
   (i) Inference criteria.
3. If preregistered, did the final report conform to this preregistration? Specifically, did the final report conform to the preregistration on:
   (a) Management of missing and incomplete data.
   (b) Pre-processing of data (e.g., how to clean and normalize).
   (c) Data processing and dealing with violation of statistical assumptions.
   (d) Management of outliers.
   (e) Statistical analysis/model.
   (f) Dependent variable(s) of the model.
   (g) Predictors/covariates of the model.
   (h) Estimation method and computation of standard errors.
   (i) Inference criteria.
4. If the final report did not completely conform to the preregistration, was the particular deviation handled by blinding data or blinded analysis?

5. If the final report did not completely conform to the preregistration, was the particular deviation handled by using a multiverse analysis and report all (theoretically) possible ways of handling this case?
6. If the study was not preregistered, was blinded data management and analysis used?
7. If blinded data management and analysis was used, was it used on the following:
   (a) Management of missing and incomplete data.
   (b) Pre-processing of data (e.g., how to clean and normalize).
   (c) Data processing and dealing with violation of statistical assumptions.
   (d) Management of outliers.
   (e) Statistical analysis/model.
   (f) Dependent variable(s) of the model.
   (g) Predictors/covariates of the model.
   (h) Estimation method and computation of standard errors.
   (i) Inference criteria.
8. If the study was not preregistered and the data management and analysis was not blinded, was a type of multiverse analysis performed?
9. If a type of multiverse analysis was performed, did it incorporate the following elements:
   (a) Management of missing and incomplete data.
   (b) Pre-processing of data (e.g., how to clean and normalize).
   (c) Data processing and dealing with violation of statistical assumptions.
   (d) Management of outliers.
   (e) Statistical analysis/model.
   (f) Dependent variable(s) of the model.
   (g) Predictors/covariates of the model.
   (h) Estimation method and computation of standard errors.
   (i) Inference criteria.

It needs to be noted that such a tool needs to be further developed, tested, and calibrated. To be useful, this objectivity checklist should of course be (to a large extent) reliable and valid. In the first case, inter-rater reliability and intra-rater reliability should be assessed. I.e., have different subjects use the tool and measure the agreement between their results (J. Cohen, 1960; Fleiss, 1971), and have subjects use the tool on two or more different occasions on the same material and measure the similarity of results between these occasions (Gwet, 2008). Close similarity between

scores indicate that users will in general give similar scores when using the checklist. If there are questions in the checklist that score low on inter-rater and/or intra-rater reliability, then they should be rephrased or dropped. The validity of the checklist can be assessed by empirically verifying if research that scores high on the checklist are less prone to produce problematic results (e.g., have a higher replication rate) in comparison to research that score low on the checklist (i.e., criterion validity). For now, we do not have a scoring system of the checklist. The easiest scoring system would be to use the proportion of 'yes' answers of the total number of questions that are relevant for the report that is evaluated. However, this scoring system should be further developed and tested. Also, scientific practice could be simulated to assess to what extent bias and false-positive rate inflation could still be introduced with varying levels of objectivity according to the checklist. Results from such simulation studies could also be used to calibrate the scoring system. Finally, scientists could be enlisted to perform mock research to attempt to bias outcomes and/or produce false-positive results with varying levels of objectivity safeguards in play. These mock studies might identify weaknesses and gaps in the tool (and objectivity conceptualization), which could be used when calibration the tool and supplementing/removing elements.

D

APPENDIX TO CHAPTER 6

D.1. Detailed Discussion of the Mayo's Severity

In this Appendix we give a more extensive illustration of the third problem we identified with Mayo's (2018 approach to severity. In particular that the application of the SEV function can lead to counter-intuitive or even problematic conclusions. For our demonstration we borrow a simple example from (Mayo, 2018, 142–144). She describes the scenario of an "accident at a water plant" (p. 142) where a leak in the cooling system is discharging water into the ecosystem. The cooling system is "meant to ensure that the mean temperature of discharged water stays below the temperature that threatens the ecosystem, perhaps not much beyond 150 degrees Fahrenheit."

The question is, is the temperature of the water high enough to constitute an ecological disaster that requires counter measures beyond mere repair of the cooling system? When the cooling system is working properly, the temperature of the water discharged into the ecosystem is around 150 degrees Fahrenheit and a temperature of 153 or higher is considered a "full-on emergency for the ecosystem" (p. 143). A

series of 100 water measurements are taken at random time periods and their sample mean $\overline{x}$ is computed. The standard deviation is known and when the cooling system is working properly, the distribution of means from samples of 100 measurements is captured by $\overline{X} \sim N(\mu = 150, 1)$. Mayo supposes that initially the test concerns "$\mathcal{H}_0 : \mu \leq 150$ vs. $\mathcal{H}_1 : \mu > 150$" (p. 142). She set the Type I error rate to 2.5% ($\alpha = 0.025$), thus rejects "$\mathcal{H}_0$ (infer there's an indication that $\mu > 150$) iff $\overline{X} \geq 152$" (p. 142).[1]

In this scenario, we observe a sample mean of 152 degrees Fahrenheit and thus reject $\mathcal{H}_0$, fulfilling (S-1) of the Severity Requirement (see Section 6.3.1). According to the severity rationale, we are warranted in concluding that $\mu > 150$, because "[w]ere the mean temperature no higher than 150, then over 97% of the time their method would have resulted in a lower mean temperature than observed" (p. 143). However, we are primarily concerned with the extent to which these data indicate that the water has reached devastating temperatures of 153 Fahrenheit or higher ($\mu \geq 150 + \delta;\ \delta = 3$). According to the Severity Requirement (S-2), we need the probability of observing our mean temperature of 152 or lower if the actual temperature is equal the critical threshold of 153 degrees. This can be calculated with the SEV function, which is $p(\overline{X} \leq 152; \mu = 153) \approx 0.16$. From these results, one can conclude that one does not have evidence for the claim that the water has reached temperatures of 153 degrees Fahrenheit or higher. Namely, the "severity principle blocks $\mu > 153$ because it is fairly probable (84% of the time) that the test would yield an even larger mean temperature than we got, if the water samples came from a body of water whose mean temperature is 153" (p. 144). Thus, the results do not meet requirement (S-2) and the hypothesis that the temperature of the water is 153 degrees or higher does not pass a severe test (see Figure D.1).

With this explication, the problem comes into view; the severity principle operationalized as the SEV function implies a curious, perhaps outright problematic concept of statistical evidence. In the water plant example, the observed mean temperature of 152 degrees Fahrenheit is not considered evidence that the water has reached the dangerous temperature of 153 degrees ($SEV(T^+, \overline{x}, \mu > 153) \approx 0.16$). However, this result is *independent of what is considered the normal or default state of affairs*. In her example, this is 150 degrees Fahrenheit, but this value is irrelevant for her SEV function as long as the observed mean is sufficiently high to reject the null hypothesis. Suppose that we had observed the same mean temperature of 152 degrees Fahrenheit, though the normal mean temperature is 100 degrees. Then one would again reject $\mathcal{H}_0$ and meet requirement (S-1); one would again not meet

[1] For convenience, Mayo rounds 151.96 up to 152.

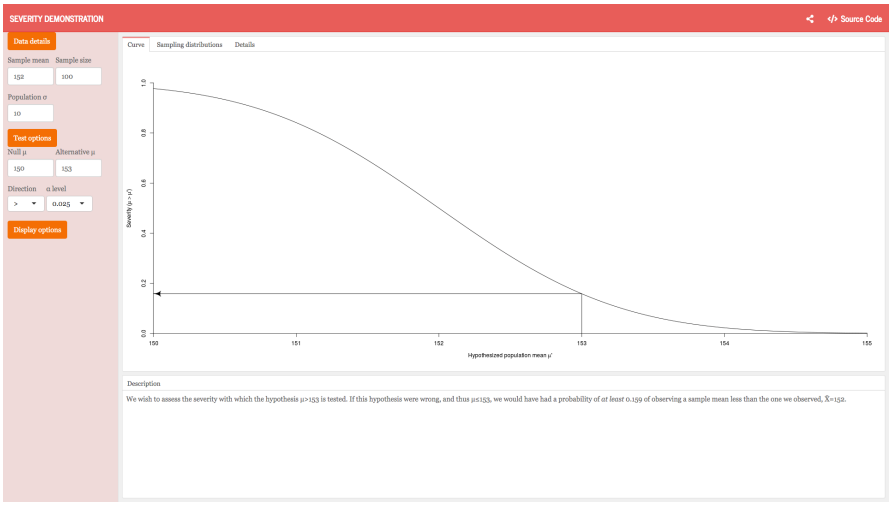


Figure D.1: *SEV* results for the original water plant example. In this case, $\mathcal{H}_0 : \mu = 150$, $\mathcal{H}_1 : \mu = 153$, $\overline{x} = 152$, and $SEV = 0.159$. This image is a screenshot from the *Severity Demonstration application*. This Shiny App was developed by Morey (2020) and can be accessed via `https://richarddmorey.shinyapps.io/severity/?mu0=150&mu1=153&sigma=10&n=100&xbar=152&xmin=150&xmax=155&alpha=0.025&dir=\%3E`

(S-2) and draw the same conclusion as before (see Figure D.2). However, when normal temperatures are around 100 degrees, you observe a mean temperature of 152 with a standard error of 1, and an "full-on emergency for the ecosystem" is imminent when the temperature is 153, it is clear that counter measures are acutely required. In short, this goes against any intuition one might have about the concept of severity, treatments of this concept by other philosophers and methodologists (e.g., Meehl, 1990b, 2005; Popper, 2002[1959]; Roberts & Pashler, 2000), and even against a reasonable interpretation of Mayo's strong severity principle (2018, p. 14).

In summary, Mayo's formalization of severity ignores context and leads to problematic conclusions when explicated.

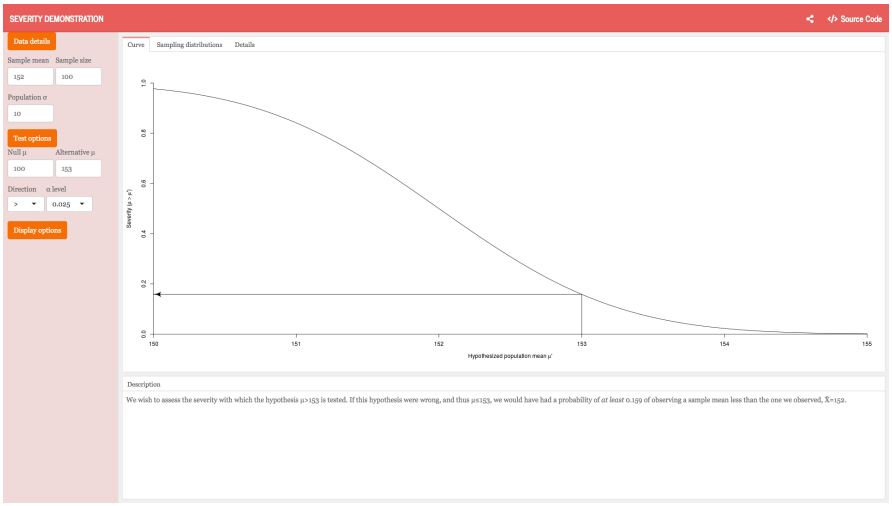


Figure D.2: *SEV* results for the original water plant example. In this case, $\mathcal{H}_0 : \mu = 100$, $\mathcal{H}_1 : \mu = 153$, $\overline{x} = 152$, and $SEV = 0.159$. This image is a screenshot from the *Severity Demonstration application*. This Shiny App was developed by Morey (2020) and can be accessed via `https://richarddmorey.shinyapps.io/severity/?mu0=100&mu1=153&sigma=10&n=100&xbar=152&xmin=150&xmax=155&alpha=0.025&dir=\%3E`

BIBLIOGRAPHY

Aczel, B., Palfi, B., & Szaszi, B. (2017). Estimating the evidential value of significant results in psychological science. *PLoS ONE*, *12*(8), e0182651.

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., ... others (2020). A consensus-based transparency checklist. *Nature human behaviour*, *4*(1), 4–6.

Aguinis, H., Cascio, W. F., & Ramani, R. S. (2017). Science's reproducibility and replicability crisis: International business is not immune. *Journal of International Business Studies*, *48*, 653–663.

Ahn, W.-Y., Gu, H., Shen, Y., Haines, N., Hahn, H. A., Teater, J. E., ... Pitt, M. A. (2020). Rapid, precise, and reliable measurement of delay discounting using a Bayesian learning algorithm. *Scientific Reports*, *10*, 12091.

Alford, R. D. (1988). *Naming and identity: A cross-cultural study of personal naming practices*. New Haven: HRA Flex Books.

Alger, B. (2019). *Defense of the scientific hypothesis: From reproducibility crisis to big data*. Oxford University Press.

Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., ... Camerer, C. (2019). Predicting the replicability of social science lab experiments. *PloS one*, *14*(12), e0225826.

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. *American Statistician*, *73*(sup1), 262–270.

Anvari, F., & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology*, *3*(3), 266-286.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437.

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*(26), 353–366.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Baril, G. L., & Cannon, J. T. (1995). What is the probability that null hypothesis testing is meaningless? *American psychologist*.

Barnes, C., Boutron, I., Giraudeau, B., Porcher, R., Altman, D. G., & Ravaud, P. (2015). Impact of an online writing aid tool for writing a randomized trial report: The COBWEB (Consort-based WEB tool) randomized controlled trial. *BMC Medicine*, *13*(1), 221.

Bartoš, F., & Schimmack, U. (2020). Z-curve. 2.0: Estimating replication rates and discovery rates. *PsyArXiv*. (Accessed on 26 June 2020)

Beebe, J. R., & Undercoffer, R. (2016). Individual and cross-cultural differences in semantic intuitions: New experimental findings. *Journal of Cognition and Culture*, *16*(3-4), 322–357.

Begley, C. G., & Ellis, L. M. (2012, mar). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533.

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of personality and social psychology*, *100*(3), 407.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... others (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10.

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2011). Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: An argument for proper multiple comparisons correction. *Journal of Serendipitous and Unexpected Results*, *1*, 1–5.

Bennett, J. H. (Ed.). (1990). *Statistical inference and analysis: Selected correspondence of R. A. Fisher*. Oxford: Clarendon Press.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, *18*, 1–32.

Berger, J. O., & Wolpert, R. L. (1984). *The Likelihood Principle*. Hayward, Calif.: Institute of Mathematical Statistics.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle*. Institute of Mathematical Statistics.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, *37*(219), 325–335.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian Theory*. New York: Wiley.

Betz, G. (2013). In defence of the value free ideal. *European Journal for Philosophy of Science*, *3*(2), 207–220.

Biddle, J. (2007). Lessons from the vioxx debacle: What the privatization of science can teach us about social epistemology. *Social Epistemology*, *21*(1), 21–39.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, *53*, 259–326.

Bishop, D. V., & Thompson, P. A. (2016). Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ*, *4*, e1715.

Booth, A. (2006). "brimful of starlite": toward standards for reporting literature searches. *Journal of the Medical Library Association*, *94*(4), 421.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2018). *Analyzing social networks*. Sage.

Borsboom, D., van der Maas, H., Dalege, J., Kievit, R., & Haig, B. (2020). Theory construction methodology: A practical framework for theory formation in psychology. *PsyArXiv*. (Accessed on 24 June 2020)

Boutron, I., Dutton, S., Ravaud, P., & Altman, D. G. (2010). Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *Jama*, *303*(20), 2058–2064.

Bright, W. (2003). What is a name? Reflections on onomastics. *Language and Linguistics*, *4*(4), 669–681.

Brown, M. (2013). The source and status of values for socially responsible science. *Philosophical Studies*, *163*, 67–76.

Brown, N. J. L., & Heathers, J. A. (2017). The grim test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8*(4), 363–369.

Bueter, A. (2015). The irreducibility of value-freedom to theory assessment. *Studies in History and Philosophy of Science Part A*, *49*, 18–26.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, *14*(5), 365–376.

Cain, D., Loewenstein, G., & Moore, D. (2005). The dirt on coming clean: Perverse effects of disclosing conflicts of interest. *The Journal of Legal Studies*, *34*(1), 1–25.

Cambrosio, A., Keating, P., Schlich, T., & Weisz, G. (2006). Regulatory objectivity and the generation and management of evidence in medicine. *Social Science & Medicine*, *63*(1), 189 - 199.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., …

others (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... others (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Cappelen, H. (2012). *Philosophy without intuitions*. Oxford: Oxford University Press.

Carnap, R. (1950). *Logical foundations of probability*. Chicago: The University of Chicago Press.

Carney, D. (2016). *My position on "power poses".* Retrieved from `http://faculty.haas.berkeley.edu/dana_carney/pdf_my%20position%20on%20power%20poses.pdf` (accessed April 15, 2020)

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32.

Carroll, C., Booth, A., & Lloyd-Jones, M. (2012). Should we exclude inadequately reported studies from qualitative systematic reviews? an evaluation of sensitivity analyses in two case study reviews. *Qualitative Health Research*, *22*(10), 1425–1434.

Carroll, H. A., Toumpakari, Z., Johnson, L., & Betts, J. A. (2017, 10). The perceived feasibility of methods to reduce publication bias. *PLOS ONE*, *12*(10), 1-19.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural computation*, *22*(4), 887–905.

Center for Open Science. (2020). *Registered reports: Peer review before results are known to align scientific values and practices.* Retrieved from `https://www.cos.io/our-services/registered-reports` (Accessed on 13 August 2020)

CERN. (2013). *New results indicate that new particle is a higgs boson.* Retrieved from `https://home.cern/news/news/physics/new-results-indicate-new-particle-higgs-boson` (Accessed on 2 June 2020)

Chambers, C. D. (2013). Registered reports: a new publishing initiative at cortex. *Cortex*, *49*(3), 609–610.

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015).

Registered reports: realigning incentives in scientific publishing. *Cortex*, *66*, A1–A2.

Chen, J. J., Tsong, Y., & Kang, S.-H. (2000). Tests for equivalence or noninferiority between two proportions. *Drug Information Journal*, *26*, 569–578.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Cohen, J. (1994). The earth is round (p <.05). *American psychologist*, *49*(12), 997.

Cohen, J. (1995). The earth is round (p<. 05): Rejoinder. *American psychologist*.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, *24*, 385–396.

Colling, L. J., & Szűcs, D. (2018). Statistical inference and the replication crisis. *Review of Philosophy and Psychology*, 1–27.

Colombo, M., Duev, G., Nuijten, M. B., & Sprenger, J. (2018). Statistical reporting inconsistencies in experimental philosophy. *PloS one*, *13*(4).

Cooke, A., Smith, D., & Booth, A. (2012). Beyond pico: the spider tool for qualitative evidence synthesis. *Qualitative health research*, *22*(10), 1435–1443.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in psychology*, *1*, 26.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., ... others (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 1–36.

Cumming, G. (2013a). The new statistics: A how-to guide. *Australian Psychologist*, *48*(3), 161–170.

Cumming, G. (2013b). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7–29.

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the american psychological association publication manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology*, *64*(3), 138–146.

Dancygier, B. (2009). Genitives and proper names in constructional blends. In V. Evans & S. Pourcel (Eds.), *New directions in cognitive linguistics* (pp. 161–184). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Dancygier, B., & Vandelanotte, L. (2017). Viewpoint phenomena in multimodal

communication. *Cognitive Linguistics*, *28*(3), 371–380.

Daston, L., & Galison, P. (1992). The image of objectivity. *Representations*, *40*, 81–128.

Daston, L., & Galison, P. (2007). *Objectivity*. New York: Zone Books.

Deeks, J., Dinnes, J., D'amico, R., Sowden, A., Sakarovitch, C., Song, F., ... Altman, D. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment*, *7*(27), iii–x, 1–173.

Deutsch, M. E. (2015). *The myth of the intuitive: Experimental philosophy and philosophical method*. Cambridge, MA: MIT Press.

Devitt, M., & Porot, N. (2018). The reference of proper names: Testing usage and intuitions. *Cognitive Science*, *42*(5), 1552-1585.

Diamond, G. A., & Kaul, S. (2004). Prior convictions: Bayesian approaches to the analysis and interpretation of clinical megatrials. *Journal of the American College of Cardiology*, *43*, 1929–1939.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York: Palgrave MacMillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psycholology*, *5:781*.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psycholology*, *72*, 78–89.

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, *2*, 364–377.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, *67*(4), 559–579. doi: 10.1086/392855

Douglas, H. (2004). The irreducible complexity of objectivity. *Synthese*, *138*(3), 453–473.

Douglas, H. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.

Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen-Huitink, J. (in press). Conditionals and inferential connections: Toward a new semantics. *Thinking and Reasoning*, 1–41.

Downs, S., & Black, N. (1998a). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of epidemiology and community health*, *52*, 377-84.

Downs, S., & Black, N. (1998b). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of epidemiology and community health*, *52*, 377–384. doi: 10.1136/jech.52.6.377

Duyx, B., Urlings, M. J., Swaen, G. M., Bouter, L. M., & Zeegers, M. P. (2017). Scientific citations favor positive results: a systematic review and meta-analysis. *Journal of Clinical Epidemiology*, *88*, 92–101.

Dyson, F. W., Eddington, A. S., & Davidson, C. (1920). A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of may 29, 1919. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *220*(571-581), 291–333.

Edwards, M. A., & Roy, S. (2017). Academic research in the 21st century: maintaining scientific integrity in a climate of perverse incentives and hypercompetition (vol 34, pg 51, 2017). *Environmental Engineering Science*, *34*(8), 616–616.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Edwards, W., Lindman, H., & Savage, L. J. (1992). Bayesian statistical inference for psychological research. In *Breakthroughs in statistics* (pp. 531–578). Springer.

Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, *40*, 1–5.

Elliott, K. C., & McKaughan, D. J. (2009). How values in scientific discovery and pursuit alter theory appraisal. *Philosophy of Science*, *76*(5), 598–611.

Engman, A. (2013). Is there life after p< 0.05? statistical significance and quantitative sociology. *Quality & quantity*, *47*(1), 257–270.

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*, *1*, 281–295.

Evans, M. (2015). *Measuring statistical evidence using relative belief*. Boca Raton, FL: CRC Press.

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, *114*(14), 3714–3719. Retrieved from `https://www.pnas.org/content/114/14/3714`

Farnham, A., Kurz, C., Öztürk, M. A., Solbiati, M., Myllyntaus, O., Meekes, J., ... others (2017). Early career researchers want open science. *Genome biology*, *18*(1), 1–4.

Feltz, A., & Cokely, E. T. (2009). Do judgments about freedom and responsibility de-

pend on who you are? personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition*, *18*(1), 342–350.

Firestein, S. (2015). *Failure: Why science is so successful*. Oxford University Press.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594-604), 309–368.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.

Fisher, R. A. (1935a). *The design of experiments*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935b). The logic of inductive inference (with discussion). *Journal of the Royal Statistical Society*, *98*, 39–82.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. New York: Hafner.

Flake, J. K., & Fried, E. I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *PsyArXiv*. (Accessed on 23 June 2020)

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Francis, G. (2014). The frequency of excess success for articles in psychological science. *Psychonomic bulletin & review*, *21*.

Frankenhuis, W. E., & Nettle, D. (2018). Open science is liberating and can foster creativity. *Perspectives on Psychological Science*, *13*(4), 439–447.

Freese, J., & King, M. M. (2018). Institutionalizing transparency. *Socius*, *4*, 2378023117739216.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, *100*, 25–50.

Frick, R. W. (1995). A problem with confidence intervals. *American psychologist*.

Galak, J., Leboeuf, R., Nelson, L., & Simmons, J. (2012, 08). Correcting the past: Failures to replicate psi. *Journal of personality and social psychology*, *103*.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.

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 was posited ahead of time. *PsyArXiv*. (Accessed on 26 June 2020)

Genone, J., & Lombrozo, T. (2012). Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology*, *25*(5), 717–742.

Gervais, W. (2017). *Post publication peer review.* Retrieved from `http://willgervais`

.com/blog/2017/3/2/post-publication-peer-review (Accessed April 15, 2020)

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199–200.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606.

Goldacre, B. (2014). *Bad pharma: how drug companies mislead doctors and harm patients*. Macmillan.

Good, I. J. (1979). Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika*, *66*, 393–396.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian *t*-tests. *The American Statistician*, *74*, 137–143.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, *92*.

Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*, 123–138.

Gwet, K. (2008). Intrarater reliability. *Methods and Applications of Statistics in Clinical Trials*, *2*, 473–485.

Hacking, I. (2015). Let not talk about objectivity. In J. Y. Tsou, A. Richardson, & F. Padovani (Eds.), *Objectivity in science.* Springer Verlag.

Haidich, A. (2010). Meta-analysis in medical research. *Hippokratia*, *14*(Suppl 1), 29—37.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. MIT press.

Harding, S. (2015). Objectivity for sciences from below. In J. Y. Tsou, A. Richardson, & F. Padovani (Eds.), *Objectivity in science.* Springer Verlag.

Harris, R. (2017). *Rigor mortis: How sloppy science creates worthless cures, crushes hope, and wastes billions*. New York: Basic Books.

Hawkins, C. B., & Nosek, B. A. (2012). Motivated independence? implicit party identity predicts political judgments among self-proclaimed independents. *Personality and Social Psychology Bulletin*, *38*(11), 1437–1452.

Heathers, J. A. (2020). *Preprints aren't the problem — we are the problem.* Retrieved from https://medium.com/@jamesheathers/preprints-arent-the-problem-we-are-the-problem-75d29a317625 (Accessed on 26 June 2020)

Heathers, J. A., Anaya, J., van der Zee, T., & Brown, N. J. L. (2018). Recovering data from summary statistics: Sample parameter reconstruction via iterative techniques (sprite). *PeerJ Preprints*. (Accessed on 26 June 2020)

Hicks, D. J. (2014). A new direction for science and values. *Synthese*, *191*(14), 3271–95.

Higgins, J. P., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (Vol. 4). John Wiley & Sons.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, *327*(7414), 557.

Higgins, J. P., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 137–159.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. CRC Press.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses*. New York: Springer.

Horwich, P. (1982). *Probability and evidence*. Cambridge: Cambridge University Press.

Hossenfelder, S. (2017). *How popper killed particle physics.* Retrieved from `http://backreaction.blogspot.com/2017/11/how-popper-killed-particle-physics.html` (Accessed on 24 June 2020)

Hossenfelder, S. (2020). *Predictions are overrated.* Retrieved from `http://backreaction.blogspot.com/2020/05/predictions-are-overrated.html` (Accessed on 22 July 2020)

Howard, D. (1997). A peek behind the veil of maya: Einstein, schopenhauer, and the historical background of the conception of space as a ground for the individuation of physical systems. In E. John & J. D. Norton (Eds.), *The cosmos of science: Essays of exploration* (pp. 87–150). University of Pittsburgh Press.

Howie, D. (2002). *Interpreting probability: Controversies and developments in the early twentieth century*. Cambridge: Cambridge University Press.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach (3rd. ed.)*. Chicago: Open Court.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician*, *57*, 171–182.

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used

measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, 2515245919882903.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Ioannidis, J. P. A., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in cognitive sciences*, *18*(5), 235–241.

Ioannidis, J. P. A., & Nicholson, J. P. A. (2012). Research grants: Conform and be funded. *Nature*, *492*, 34–36.

Ioannidis, J. P. A., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*, *335*(7626), 914.

Izumi, Y., Kasaki, M., Zhou, Y., & Oda, S. (2018). Definite descriptions and the alleged east–west variation in judgments about reference. *Philosophical Studies*, *175*(5), 1183–1205.

Jadad, A. R., & O'Grady, L. (2008). How should health be defined? *BMJ: British Medical Journal (Online)*, *337*.

JASP Team. (2018). *JASP (Version 0.9)[Computer software].* Retrieved from `https://jasp-stats.org/`

Jeffrey, R. C. (1971). *The logic of decision*. Chicago and London: University of Chicago Press. (Second edition 1983)

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, UK: Oxford University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Jeffreys, H. (1973). *Scientific inference* (3rd ed.). Cambridge, UK: Cambridge University Press.

Joanna Briggs Institute. (2017). *Checklist of text and opinion.* Retrieved from `https://joannabriggs.org/sites/default/files/2020-08/Checklist_for_Text_and_Opinion.pdf` (Accessed on 3 October 2020)

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.

Johnson, V. E. (2019, mar). Evidence From Marginally Significant t Statistics. *The American Statistician*, *73*(sup1), 129–134.

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, *50*(1), 59–99.

Joshua, K. (2019). Philosophical intuitions are surprisingly robust across demographic differences. *Epistemology & Philosophy of Science*, *56*(2).

Jukola, S. (2017). On ideals of objectivity, judgments, and bias in medical research – a comment on stegenga. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *62*, 35–41. doi: https://doi.org/10.1016/j.shpsc.2017.02.001

Kasaki, M. (2017). Problems of translation for cross-cultural experimental philosophy. *Journal of Indian Council of Philosophical Research*, *34*(3), 481–500.

Kass, R. E., Caffo, B. S., Davidian, M., Meng, X. L., Yu, B., & Reid, N. (2016). Ten Simple Rules for Effective Statistical Practice. *PLoS Computational Biology*, *12*(6), e1004961.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kass, R. E., & Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society, Series B*, *54*, 129–144.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Khan, K. S., Ter Riet, G., Glanville, J., Sowden, A. J., Kleijnen, J., et al. (2009). *Undertaking systematic reviews of research on effectiveness: Crd's guidance for carrying out or commissioning reviews* (Tech. Rep.). Center for Reviews and Dissemination, University of York. Retrieved from `https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf`

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., ... Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490.

Klugkist, I., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological Methods*, *10*(4), 477-493.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*(1), 57–69.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710.

Knobe, J. (2007). Experimental philosophy. *Philosophy Compass*, *2*(1), 81–92.

Knobe, J. (2016). Experimental philosophy is cognitive science. *A companion to experimental philosophy*, 37–52.

Knobe, J., Buckwalter, W., Nichols, S., Robbins, P., Sarkissian, H., & Sommers, T. (2012). Experimental philosophy. *Annual review of psychology*, *63*, 81–99.

Knobe, J., & Nichols, S. (2013). *Experimental philosophy*. Oxford University Press.

Koskinen, I. (2018). *Defending a risk account of scientific objectivity.*

Kripke, S. (1980). *Naming and necessity*. Cambridge/MA: Harvard University Press.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta–analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*, 178–206.

Ladyman, J. (2002). *Understanding philosophy of science*. Psychology Press.

Lakatos, I. (1970). alsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (p. 91-196). Cambridge, UK: Cambridge University Press.

Lakatos, I. (1980). *The methodology of scientific research programmes: Volume 1: Philosophical papers* (Vol. 1). Cambridge university press.

Lakens, D. (2017a). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*, 355–362.

Lakens, D. (2017b). Professors are not elderly: Evaluating the evidential value of two social priming effects through p-curve analyses. *PsyArXiv*. (Accessed on 22 July 2020)

Lakens, D. (2019a). *Does your philosophy of science matter in practice?* Retrieved from `http://daniellakens.blogspot.com/search?q=philosophy+of+science` (Accessed on 2 June 2020)

Lakens, D. (2019b). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, *62*(3), 221–230.

Lakens, D. (2020). *Review of "the generalizability crisis" by tal yarkoni.* Retrieved from `http://daniellakens.blogspot.com/2020/01/review-of-generalizability-crisis-by.html` (Accessed on 24 June 2020)

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., ... others (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171.

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*, 259–269.

Lam, B. (2010). Are cantonese-speakers really descriptivists? revisiting cross-cultural semantics. *Cognition*, *115*(2), 320–329.

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., ... others (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*(3-4),

141–153.

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*, 114–127.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge: Cambridge University Press.

Lehrer, J. (2010). *Feeling the future: Is precognition possible?* WIRED. Retrieved from `https://www.wired.com/2010/11/feeling-the-future-is-precognition-possible/` (Accessed on 11 August 2020)

Lemoine, P. (2019). *Why falsificationism is false.* Retrieved from `https://necpluribusimpar.net/why-falsificationism-is-false/` (Accessed on 24 June 2020)

Leuschner, A. (2012). Pluralism and objectivity: Exposing and breaking a circle. *Studies in History and Philosophy of Science Part A*, *43*(1), 191–198.

Lilienfeld, S. O. (2012). Public skepticism of psychology: why many people perceive the study of human behavior as unscientific. *American Psychologist*, *67*(2), 111.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*, 986–1005.

Lindley, D. V. (1986). Comment on "Why isn't everyone a Bayesian?" by Bradley Efron. *The American Statistician*, *40*, 6–7.

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, *49*, 293–337.

Lindley, D. V. (2006). *Understanding uncertainty*. Hoboken: Wiley.

Lindner, M. D., Torralba, K. D., & Khan, N. A. (2018). Scientific productivity: An exploratory study of metrics and incentives. *PloS one*, *13*(4), e0195321.

Lindsay, D. S. (2015). *Replication in psychological science.* Sage Publications Sage CA: Los Angeles, CA.

Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016). Research preregistration 101. *APS Observer*, *29*(10).

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

Longino, H. E. (1996). Cognitive and non-cognitive values in science: Rethinking the dichotomy. In L. H. Nelson & J. Nelson (Eds.), *Feminism, science, and the philosophy of science* (pp. 39–58). Dordrecht: Springer Netherlands.

Longino, H. E. (2004). How values can be good for science. In P. K. Machamer & G. Wolters (Eds.), *Science, values, and objectivity* (pp. 127–142). University of Pittsburgh Press.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2019). Replication Bayes factors

from evidence updating. *Behavior Research Methods*, *51*, 2498–2508.

Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*, *72*, 4–13.

MacCoun, R. (1998). Biases in the interpretation and use of research results. *Annual review of psychology*, *49*, 259-87.

MacCoun, R., & Perlmutter, S. (2015). Blind analysis: hide results to seek the truth. *Nature News*, *526*(7572), 187.

Machery, E. (2017). *Philosophy within its proper bounds*. Oxford University Press.

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition*, *92*(3), B1–B12.

Machery, E., Olivola, C. Y., & LeBlanc, M. (2009). Linguistic and metalinguistic intuitions in the philosophy of language. *Analysis*, *69*, 689–694.

Machery, E., & Stich, S. (2019). *Demographic differences in philosophical intuition: A reply to joshua knobe.* Retrieved from `https://leiterreports.typepad.com/files/reply-to-knobe----demographic-difference-in-philosophical-intuition---11-3-2019.docx` (Accessed June 2nd, 2020)

Machery, E., Sytsma, J., & Deutsch, M. (2015). Speaker's reference and cross-cultural semantics. In A. Bianchi (Ed.), *On reference* (p. 62-76). Oxford: Oxford University Press.

Malički, M., & Marušić, A. (2014). Is there a solution to publication bias? researchers call for changes in dissemination of clinical research results. *Journal of Clinical Epidemiology*, *67*(10), 1103 - 1110.

Marks-anglin, A., & Chen, Y. (2020). *A Historical Review of Publication Bias.*

Marmaridou, S. S. A. (2000). *Pragmatic meaning and cognition*. John Benjamins Publishing.

Martin, R., & Liu, C. (2014). A note on p-values interpreted as plausibilities. *Statistica Sinica*, 1703–1716.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, D. G. (2010). An Error in the Argument from Conditionality and Sufficiency to the Likelihood Principle. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability and the objectivity and rationality of science* (pp. 305–314). Cambridge: Cambridge University Press.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the science wars*. Cambridge: Cambridge University Press.

McGraw, K. O. (1995). Determining false alarm rates in null hypothesis testing

research. *American psychologist*.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (in press). Abandon statistical significance. *The American Statistician*.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 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*(1), 806–834.

Meehl, P. E. (1986). What Social Scientists Don ' t Understand. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science: Pluralisms and subjectivities* (pp. 315–338). Chicago: University of Chicago Press.

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.

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological reports*, *66*(1), 195–244.

Meehl, P. E. (1992). Cliometric metatheory: The actuarial approach to empirical, history-based philosophy of science. *Psychological reports*, *71*, 339–339.

Meehl, P. E. (2005). Cliometric metatheory ii: Criteria scientists use in theory appraisal and why it is rational to do so. *Psychological Reports*, *91*(6), 339–404.

Megill, A. (1994). Four senses of objectivity. In *Rethinking objectivity.*

Monson, C. M., Schnurr, P. P., Resick, P. A., Friedman, M. J., Young-Xu, Y., & Stevens, S. P. (2006). Cognitive processing therapy for veterans with military-related posttraumatic stress disorder. *Journal of Consulting and clinical Psychology*, *74*(5), 898.

Morey, R. D. (2020). *Severity demonstration.* Retrieved from `\url{https://richarddmorey.shinyapps.io/severity/}` (Accessed on 27 August 2020)

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The Fallacy of Placing Confidence in Confidence Intervals. *Psychonomic Bulletin & Review*, *23*, 103-123.

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on cumming. *Psychological science*, *25*(6), 1289–1290.

Moyé, L. A. (2008). Bayesians in Clinical Trials: Asleep at the Switch. *Statistics in Medicine*, *27*, 469–482.

Müller, H. M. (2010). Neurolinguistic findings on the language lexicon: the special role of proper names. *Chinese Journal of Physiology*, *53*(6), 351–358.

Munafò, M. (2016). Open science and research reproducibility. *ecancermedicalscience*, *10*.

Munafo, M. R., Clark, T. G., & Flint, J. (2004). Assessing publication bias in genetic association studies: evidence from a recent meta-analysis. *Psychiatry research*, *129*(1), 39–44.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229.

Mutz, D. C. (2011). *Population-based survey experiments*. Princeton University Press.

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, *57*(3-4), 53–67.

Myung, J. I., & Pitt, M. A. (2009). Optimal Experimental Design for Model Discrimination. *Psychological Review*, *116*(3), 499–518.

Nelson, J. A. (2014). The power of stereotyping and confirmation bias to overwhelm accurate assessment: The case of economics, gender, and risk aversion. *Journal of Economic Methodology*, *21*(3), 211–231.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psycholgy*, *69*, 511–534.

Neurath, O. (1973). Anti-spengler. In M. Neurath & R. S. Cohen (Eds.), *Empiricism and sociology* (pp. 158–213). Springer.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*, 97–131.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A*, *231*, 289–337.

Neyman, J., & Pearson, E. S. (1967). *Joint Statistical Papers*. Berkeley, Calif.: University of California Press.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and Systems of Thought: Holistic Versus Analytic Cognition. *Psychological Review*, *108*(2), 291–310.

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *Elife*, *5*, e21451.

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.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods*, *48*(4), 1205–1226.

Nuijten, M. B., van Assen, M. A., Hartgerink, C. H., Epskamp, S., & Wicherts, J. (2017). The validity of the tool "statcheck" in discovering statistical reporting inconsistencies. *PsyArXiv*. (Accessed on 22 July 2020)

Oaksford, M., & Chater, N. (1994). A Rational Analysis of the Selection Task as Optimal Data Selection. *Psychological Review*, *101*(4), 608–631.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic bulletin & review*, *26*(5), 1596–1618.

O'Mara, A. J. (2008). Methodological and substantive applications of meta-analysis: Multilevel modelling, simulation, and the construct validation of self-concept. *Unpublished doctoral dissertation). Oxford University*.

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

Organization, W. H., et al. (1950). *The preamble of the constitution of the world health organization*.

Parker, S. (1995). The" difference of means" may not be the" effect size". *Amperican Psychologist*, *50*, 1101–1102.

Pashler, H., & Wagenmakers, E. (2012, 11). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528-530.

Patsopoulos, N. A., Evangelou, E., & Ioannidis, J. P. (2008). Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. *International journal of epidemiology*, *37*(5), 1148–1157.

Percie du Sert, N., Bamsey, I., Bate, S. T., Berdoy, M., Clark, R. A., Cuthill, I., ... others (2017). The experimental design assistant. *PLoS biology*, *15*(9), e2003779.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences. a practical guide*. Blackwell.

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353.

Pohl, R. F. (2004). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., ... Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature reviews neuroscience*, *18*(2), 115.

Popper, K. R. (1979). *Objective Knowledge: An Evolutionary Approach*. Oxford:

Clarendon Press.

Popper, K. R. (2002[1959]). *The Logic of Scientific Discovery*. London: Routledge. (Reprint of the revised English 1959 edition. Originally published in German in 1934 as "Logik der Forschung")

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, *10*(9), 712–712.

Proverbio, A. M., Mariani, S., Zani, A., & Adorni, R. (2009). How are 'Barack Obama' and 'President Elect' differentially stored in the brain? An ERP investigation on the processing of proper and common noun pairs. *PLoS ONE*, *4*(9), e7126.

R Development Core Team. (2004). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org` (ISBN 3–900051–00–3)

Raudenbush, S. W. (2009). Analzing effect sizes: Random effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295–315). New York: Russell Sage Foundation.

Reiss, J., & Sprenger, J. (2014). *Scientific objectivity.* Retrieved from `https://plato.stanford.edu/entries/scientific-objectivity/` (Accessed on 16 April 2020)

Reitsma, J., Rutjes, A., Whiting, P., Vlassov, V., Leeflang, M., Deeks, J., et al. (2009). Assessing methodological quality. *Cochrane handbook for systematic reviews of diagnostic test accuracy version*, *1*(0), 1–28.

Riley, R. D., Higgins, J. P., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ*, *342*, d549.

Ritchie, S., Wiseman, R., & French, C. (2012). Failing the future: Three unsuccessful attempts to replicate bem's 'retroactive facilitation of recall' effect. *PloS one*, *7*, e33423.

Roberts, S., & Pashler, H. (2000). How Persuasive is a Good Fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Robinson, G. K. (2019). What properties might statistical inferences reasonably be expected to have?—crisis and resolution in statistical inference. *The American Statistician*, *73*, 243–252.

Röhmel, J. (2001). Statistical considerations of FDA and CPMP rules for the investigation of new anti–bacterial products. *Statistics in Medicine*, *20*, 2561–2571.

Rohrer, J. M., DeBruine, L., Heyman, T., Jones, B. C., Schmukle, S., Silberzahn, R., ... others (2018). Putting the self in self-correction. *PsyArXiv*. (accessed April 15, 2020)

Romero, F. (2016). Can the behavioral sciences self-correct? a social epistemic study. *Studies in History and Philosophy of Science Part A*, *60*, 55–69.

Romero, F. (2018). Who should do replication labor? *Advances in Methods and Practices in Psychological Science*, *1*(4), 516–537.

Rothwell, P. M. (2005). External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet*, *365*(9453), 82–93.

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.

Rudner, R. (1953). The Scientist *qua* Scientist Makes Value Judgments. *Philosophy of Science*, *20*, 1-6. Retrieved from `http://www.jstor.org/stable/185617`

Russell, B. (1905). On Denoting. *Mind*, *14*, 479–493.

Safer, D. J. (2002). Design and reporting modifications in industry-sponsored comparative psychopharmacology trials. *The Journal of nervous and mental disease*, *190*(9), 583–592.

Sansone, C., Morf, C. C., & Panter, A. T. (2003). *The sage handbook of methods in social psychology*. Sage Publications.

Savage, L. J. (1972). *The Foundations of Statistics* (2nd ed.). New York: Wiley. (Originally published in 1954)

Schafer, A. (2004). Biomedical conflicts of interest: a defence of the sequestration thesis—learning from the cases of nancy olivieri and david healy. *Journal of Medical Ethics*, *30*(1), 8–24.

Scharp, K. (2013). *Replacing truth*. Oxford University Press UK.

Scheel, A. M., Schijen, M., & Lakens, D. (2020). An excess of positive results: Comparing the standard psychology literature with registered reports. *PsyArXiv*. (Accessed on 14 August 2020.)

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological methods*, *17*.

Schimmack, U. (2020). *Why the journal of personality and social psychology should retract article doi: 10.1037/a0021524 "feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect" by daryl j. bem.* Retrieved from `https://replicationindex.com/2018/01/05/bem-retraction/` (Accessed on 16 April 2020)

Schimmack, U., & Brunner, J. (2017). Z-curve. *OSF Preprints*. (Accessed on 22 July 2020)

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for Compelling Evidence. *Psychonomic Bulletin & Review*, *25*, 128–142.

Schuetz, P., & Caflisch, A. (2008). Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. *Physical Review E*, *77*(4), 046112.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, *27*(2), 135–153.

Searle, J. R. (1958). Proper names. *Mind*(266), 166–173.

Searle, J. R. (1975). A taxonomy of illocutionary acts. In K. Gunderson (Ed.), *Language, mind and knowledge* (p. 344-369). University of Minnesota Press.

Semenza, C. (2006). Retrieval pathways for common and proper names. *Cortex*, *42*(6), 884–891.

Senn, S. (2011). You May Believe You Are a Bayesian but You Are Probably Wrong. *Rationality, Markets and Morals*, *2*, 48–66.

Sessa, B. (2017). MDMA and PTSD treatment: "PTSD: from novel pathophysiology to innovative therapeutics". *Neuroscience letters*, *649*, 176–180.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (in press). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science*.

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.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80.

Simonsohn, U. (2011). Spurious? Name similarity effects (implicit egotism) in marriage, job and moving decisions. *Journal of Personality and Social Psychology*, *101*, 1–24.

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.

Sindhu, F., Carpenter, L., & Seers, K. (1997). Dovelopment of a tool to rate the quality assessment of randomized controlled trials using a delphi technique. *Journal of Advanced Nursing*, *25*(6), 1262-1268.

Sober, E. (2007). Evidence and value freedom..

Sophia, A., & Marmaridou, S. (1989). Proper names in communication. *Journal of Linguistics*, *25*(2), 355–372.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical studies*, *132*(1), 99–107.

Sprenger, J., & Hartmann, S. (2019). *Bayesian philosophy of science*. Oxford: Oxford University Press.

Stan Development Team. (2016). *rstan: The R interface to Stan.* Retrieved from `http://mc-stan.org/` (R package version 2.14.1)

Stapel, D. A. (2012). *Ontsporing*. Prometheus Amsterdam.

Stark, P. B., & Saltelli, A. (2018). Cargo-cult statistics and scientific crisis. *Significance*, *15*(4), 40–43.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

Steel, D. (2010). Epistemic values and the argument from inductive risk. *Philosophy of Science*, *77*(1), 14–34.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on bayes factor design analysis using an informed prior. *Behavior research methods*, *51*(3), 1042–1058.

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*(4), 497–507.

Stegenga, J. (2018). *Medical nihilism*. Oxford University Press.

Stelfox, H., Chua, G., O'Rourke, K., & Detsky, A. (1998). Conflict of interest in the debate over calcium-channel antagonists. *The New England journal of medicine*, *338*(2), 101—106.

Sterne, J. A., Sutton, A. J., Ioannidis, J. P. A., et al. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*, d4002.

Stich, S., & Tobia, K. (2016). Experimental philosophy and the philosophical tradition. In J. Sytsma & W. Buckwalter (Eds.), *A companion to experimental philosophy* (pp. 5–21). John Wiley & Sons.

Stuart, M. T., Colaço, D., & Machery, E. (2019). P-curving x-phi: Does experimental philosophy have evidential value? *Analysis*, *79*(4), 669–684.

Sutton, A. J. (2009). Publication bias. In H. Copper, L. Hedges, & L. Valtentine (Eds.), *The handbook of research synthesis and meta-analysis.* Russell Sage Foundation.

Sytsma, J., & Buckwalter, W. (2016). *A companion to experimental philosophy*. John Wiley & Sons.

Sytsma, J., & Livengood, J. (2011). A new perspective concerning experiments on semantic intuitions. *Australasian Journal of Philosophy*, *89*(2), 315–332.

Sytsma, J., Livengood, J., Sato, R., & Oguchi, M. (2015). Reference in the land of the rising sun: A cross-cultural study on the reference of proper names. *Review of Philosophy and Psychology*, *6*(2), 213–230.

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2019). Is preregistration worthwhile. *Trends in Cognitive Sciences*, *24*(2), 94–95.

Szucs, D. (2016). A Tutorial on Hunting Statistical Significance by ChasingN. *Frontiers in psychology*, *7*, 1444.

Szucs, D., & Ioannidis, J. (2017). When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in human neuroscience*, *11*, 390.

Tarski, A. (1936). The concept of truth in formalized languages. In A. Tarski (Ed.), *Logic, semantics, metamathematics* (pp. 152–278). Oxford University Press.

Tawakol, A., Ishai, A., Takx, R. A. P., Figueroa, A. L., Ali, A., Kaiser, Y., … Pitman, R. K. (2017). Relation between resting amygdalar activity and cardiovascular events: a longitudinal and cohort study. *The Lancet*, *389*, 834–845.

Thomas, J., Graziosi, S., Brunton, Z., J.and Ghouze, O'Driscoll, P., & Bond, M. (2020). *Eppi-reviewer: Advanced software for systematic reviews, maps and evidence synthesis.* UCL Social Research Institute.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC medical research methodology*, *8*(1), 45.

Tong, A., Flemming, K., McInnes, E., Oliver, S., & Craig, J. (2012). Enhancing transparency in reporting the synthesis of qualitative research: Entreq. *BMC medical research methodology*, *12*(1), 181.

Trafimow, D. (2013). To disconfirm or not to disconfirm: a null prediction vs. no prediction. *Frontiers in psychology*, *4*, 733.

Travers, J., Marsh, S., Williams, M., Weatherall, M., Caldwell, B., Shirtcliffe, P., … Beasley, R. (2007). External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? *Thorax*, *62*(3), 219–223.

Tukey, J. W. (1995). Controlling the proportion of false discoveries for multiple comparisons: Future directions. In V. S. L. Williams, L. V. Jones, & I. Olkin (Eds.), *Perspectives on statistics for educational research: Proceedings of a workshop*

(pp. 6–9). Research Triangle Park, NC: National Institute of Statistical Sciences.

Tunç, D. U., & Tunç, M. N. (2020). A falsificationist treatment of auxiliary hypotheses in social and behavioral sciences: Systematic replications framework. *PsyArXiv*. (Accessed on 16 April 2020)

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, *76*(2), 105.

US Department of Health and Human Services. (2001). *National toxicology program's report of the endocrine disruptors low-dose*.

Vaesen, K., Peterson, M., & Van Bezooijen, B. (2013). The reliability of armchair intuitions. *Metaphilosophy*, *44*(5), 559–578.

Valentine, T., Brennen, T., & Brédart, S. (1996). *On the importance of being ernest: the cognitive psychology of proper names*. London: Routledge.

van 't Veer, A. E., & Giner–Sorolla, R. (2016). Pre-registration in social psychology – a discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12.

van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (Eds.). (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4.

van Dongen, N. N. N. (2020). *Investigations into art appreciation: An interdisciplinary approach* (Unpublished doctoral dissertation). Erasmus University, Rotterdam. (Defended on 4 September 2020)

van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology*, *3*(1).

Veldkamp, C. L., Hartgerink, C. H., van Assen, M. A., & Wicherts, J. M. (2017). Who believes in the storybook image of the scientist? *Accountability in research*, *24*(3), 127–151.

Veldkamp, C. L., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PloS one*, *9*(12), e114876.

Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A., ..., ... Albarracín, D. (in press). A multi–site preregistered paradigmatic test of the ego depletion effect. *Psychological Science*.

vom Bruck, G., & Bodenhorn, B. (Eds.). (2006). *An anthropology of names and naming*. Cambridge: Cambridge University Press.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *14*(5), 779–804.

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian Versus Frequentist Inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). Springer.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... others (2018). Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review*, *25*(1), 35–57.

Wagenmakers, E.-J., Verhagen, A., J.and Ly, Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 123–138). John Wiley & Sons.

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. *Journal of Personality and Social Psychology*, *100*, 426–432.

Wang, L., Verdonschot, R. G., & Yang, Y. (2016). The processing difference between person names and common nouns in sentence contexts: an ERP study. *Psychological Research*, *80*(1), 94–108.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p–values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.

Weber-Schoendorfer, C., & Schaefer, C. (2008). The safety of cetirizine during pregnancy: A prospective observational cohort study. *Reproductive Toxicology*, *26*, 19–23.

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.

Wikipedia. (2019). Objectivity (science). Retrieved from `https://en.wikipedia.org/wiki/Objectivity_(science)` (Accessed: 3/18/2019)

Wilholt, T. (2008). Bias and values in scientific research. *Studies in History and Philosophy of Science Part A*, *40*(1), 92–101.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological science*, *2*(6), 409–415.

Wootton, D. (2015). *The invention of science: A new history of the scientific revolution*. Penguin UK.

Wright, J. (2018). Rescuing objectivity: A contextualist proposal. *Philosophy of the Social Sciences*, *48*(4), 385-406.

Yarkoni, T. (2019). The generalizability crisis. *PsyArXiv*. (Accessed on 23 June 2020)

Yarkoni, T. (2020). *Induction is not optional (if you're using inferential statistics): reply to lakens.* Retrieved from `https://www.talyarkoni.org/blog/2020/05/06/induction-is-not-optional-if-youre-using-inferential-statistics-reply-to-lakens/` (Accessed on 24 June 2020)

Yen, H.-L. (2006). *Processing of Proper Names in Mandarin Chinese: A Behavioral and Neuroimaging Study* (Unpublished doctoral dissertation). University of Bielefeld.

Ziliak, S. T., & McCloskey, D. N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor, Mich.: University of Michigan Press.

Ziman, J. (1996, 08). Is science losing its objectivity? *Nature*, *382*, 751-754.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*.

PUBLICATIONS

Chapter 2 published as:
van Dongen, N. N. N., & Colombo, M., Romero, F., Sprenger, J. (2020). Intuitions about the reference of proper names: A meta-analysis. *Review of Philosophy and Psychology*,1-30.

Chapter 3 has been submitted as:
van Dongen, **N. N. N.**, & van Grootel, L. (under review). A systematic review of essays on the problems and solutions of null hypothesis significance testing. *Psychological Methods*.

Chapter 4 has been published as:
van Dongen, N. N. N., van Doorn, J. B., Gronau, Q. F., van Ravenzwaaij, D., Hoekstra, R., Haucke, M. N., Lakens, D., Hennig, C., Morey, R. D., Homer, S., Gelman, A., Sprenger J., &, Wagenmakers, E-. J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician*, 73(sup1), 328–339.

Chapter 5 has been submitted as:
van Dongen, N. N. N., & Sikorski, M. (under review). Objectivity for the Research worker. *European Journal of Philosophy of Science*.

Chapter 6 has been submitted as:
van Dongen, N. N. N., Sprenger, J., & Wagenmakers, E.-J. (under review). A Bayesian perspective on severity: Risky predictions and specific hypotheses. *Psychonomic Bulletin and Review*.

Other publications:

van Dongen, N. N. N., & Zijlmans, J. (2017). The science of art: The universality of the law of contrast. *American Journal of Psychology*, 130(3), 283-294.

van Dongen, N. N. N., Van Strien, J. W., & Dijkstra, K. (2016). Implicit emotion regulation in the context of viewing artworks: ERP evidence in response to pleasant and unpleasant pictures. B*rain and Cognition*, 107, 48-54.

Dijkstra, K., & **van Dongen, N. N. N.** (2017). Moderate contrast in the evaluation of paintings is liked more but remembered less than high contrast. *Frontiers in psychology*, 8, 1507.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., ..., **van Dongen, N. N. N.**, ..., & Cushman, F. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 1-36.

van Dongen, N. N. N. (2020). The scientific hypothesis is here to stay. *Metascience*, 29, 391-394.

Short CV:

Noah N. N. van Dongen was born in Rotterdam on the 24th of August 1986. After completing secondary education, he studied Fine Arts at the Willem de Kooning Art Academy in Rotterdam, followed by a Masters in Arts and Culture Studies at the Erasmus University Rotterdam. He graduated cum laude in 2014. Subsequently, he worked as an academic teacher at the Department of Arts & Culture Studies at the Erasmus University Rotterdam and collaborated with Prof. van Eijck, Prof. van Strien, and Prof. Dijkstra on several studies into art appreciation. He continued this research as an external PhD, which he completed in September 2020.

In 2016 he started a PhD in philosophy of science under the supervision of Prof. Sprenger, first at Tilburg University and later at the University of Turin in Italy. During this period, he co-founded the Platform for Young Meta-Scientists and co-organized the yearly conference Perspectives on Scientific Error. In the third year of his PhD, he visited Prof. Wagenmakers at the University of Amsterdam for six months.

ACKNOWLEDGEMENTS

There are a few people I would like to pay tribute to. First of all, my supervisor, Jan. I thank you for taking a chance on hiring a non-philosopher for a philosophy project. Before I successfully applied for this job, I was working part-time as a graphic designer and doing research on art appreciation in my free time. I was trying and very much failing to get funding for my research.[2] I had started to look for other PhD opportunities, with little success (only one unsuccessful interview out of several applications). When I found the job offer for your project, I was already making plans for pursuing a career outside academia. Your project appealed to me and I enjoyed writing my application letter and project proposal, but I did not have high hopes. I was thus amazed and relieved when you offered me the job. I don't know where I would have ended up without it, but it is almost certain that it is something completely different from what I am doing now and probably less academic. The months leading up to the start of the job felt unreal and I kept expecting a cancellation of some sort. When my contract started, I was uncertain at first and wasn't sure I would fit in. You reassured me through your advice, supervision, and collaboration on several projects. You gave me all the freedom I wanted to pursue what I thought was interesting and relevant. You encouraged me to educate myself in philosophy, mathematics, and research methods. Although I was free to do what I wanted (or at least it felt like that), you were available for help and guidance when I needed it. I have learned a lot, from you directly and from the freedom to pursue my own educational interests. The move to Italy was an interesting one. I enjoyed my time in Turin and have fond memories of the food and wine, lunches and dinners, presentations and discussions, but not the bureaucracy and lack of air-conditioning. I am glad that I said 'yes' when asked to move with the

[2] See my previous dissertation for more details (van Dongen, 2020, p. 1–2, 137).

project to Italy.

Leonie, my co-supervisor, I am glad that through quite a chain of people I found you for the supervision and collaboration of the most time consuming project of this dissertation. The systematic review of the NHST literature was a daunting task, both by its magnitude and the absence of proper tools and guidelines. I learned a lot. I would like to especially thank you for inviting me along with your visit to the EPPI-Centre. Meeting those people, and working with you in such a quaint but cozy building in the center of London felt more like a holiday than actual work.

My fellow team members, Michal, Felipe, Mattia, Michele, Federico, Stefano, thank you for the many interesting conversations, assistance, and collaborations. Some of you were there from the beginning, one of you went to greener pastures halfway through (but stayed in touch), and the rest joined near the end. It was nice to share an office with you, for however brief a period. Some specific 'thank you's: Felipe, thank you for the Perspectives on Scientific Error and almost living together in Turin. Mattia, thank you for your knowledge of wine, going to the organic wine bar in Geneva was great and this hobby of yours is one I might emulate. Federico and Stefano, due to the pandemic we did not see much of each other, but at least the lunches we did have were great. Michele, thank you for all your help with the intricacies of Italian internet-contracts and the bureaucracy of the university. I enjoyed our conversations about religion and having you as a desk neighbor when Michal left. Michal, ex-roommate, it was great working with you. We had interesting conversations about Star Wars, sci-fi novels, research methods, and the usefulness of philosophy of science. Renting an apartment together (due to Felipe) was an interesting and pleasant experience. Maybe next time you'll win the sushi-eating contest.

For a brief period, I worked at the Tilburg University. I would like to thank all my ex-colleagues there. Out of fear of forgetting someone, I am not providing names and just assume that you know who you are. We had interesting conversations during our lunch breaks at the terrible university mensa. Also, thank you for the useful advice and feedback during the in-house presentations of my work that was very much in progress. Matteo, I thank you personally, because of our collaboration on the meta-analysis, which was very enjoyable, and our conversations about religion, which were challenging. People at the University of Amsterdam, thank you. In 2019 I visited you for half a year under the invitation of EJ. It was nice to get to know you and have lunch with you on occasion. You are a significant part of why I enjoyed my stay at the UvA. Riet and Fabian, I thank you especially. Our reading group meetings

were enjoyable and educational. Although we could not make much of the book we were reading, the discussions about our ideas, translations, and interpretations contributed greatly to the chapter on severity.

Eric-Jan, EJ, thank you. You played a big part in determining the direction of my PhD and the content of this dissertation. When I started and needed to figure out what I wanted to do, I contacted several people I thought would be interesting to talk to. My hope was that this would help me get a sense of the lay of the land and some inspiration for what I was going to do for the next four years. You were one of those people and responded promptly to my request for a meeting. We had a conversation that went longer than planned, though felt much too short. The beginnings of the third chapter of this dissertation were born during this first meeting. A six-month visit to the UvA followed, which resulted in the sixth chapter. During this visit, I also got to know Denny Borsboom, which eventually led to my new postdoc position. So, you are (somewhat) responsible for that as well.

Red team, Daniel, Anne, Peder, and Leo, and Red Team satellites (Farid, Andrea, Patrick, etc.), thank you. Having morning meetings with you during the 2020 lockdown was enjoyable and sanity preserving. Also the journal club meetings, presentations, and feedback were a valuable part of my PhD. Daniel, I want to thank you especially. Although our walks might not affect my emotional state, they are cognitively stimulating, thought provoking, and inspirational. Thank you for making me feel welcome in your group and having such wonderful PhD candidates and postdoc.

Metaphanters, especially Anne, Olmo, and Sarahanne, thank you. Seeing your work and working with you was and still is inspirational and motivating. We don't meet that often, but when we do it is impressive to see what you are capable of and working on. I am happy that I can still count myself as an 'early career researcher' and thus remain a part of PYMS/Metaphant for a little while longer.

My friends, I want to thank you as well. Since I already thanked you in my previous dissertation (van Dongen, 2020, p. 138–139), I'll just quote it here:

> Oude vrienden - Frans, Gari, Spencer, Tijl, Emiel, Freek, Shea - jullie maakten het leven leuk van voor de basisschool tot aan de universiteit. Van hutten bouwen en busje trap spelen tot uitgaan en drugs doen, jullie zorgden dat het nooit saai was in weekenden en vakanties. Jullie komen voor in de leukste herinneringen uit die tijd. Ook al zien we elkaar niet meer of maar zelden, ik zal jullie altijd als mijn vrienden blijven zien.
>
> Rolex (Rozanna en Alex), met jullie is het altijd leuk. Jullie zijn zo raar op jullie eigen manier dat het nooit saai is en ik me vaak normaal voel. Met z'n drieën

> koffie drinken en sporten toen we nog allemaal in Rotterdam woonden en met z'n drieën bellen toen Ro naar Edinburgh verhuisde en ik in Turijn ging werken. Het verbaast me elke keer weer hoeveel we tegen elkaar te vertellen hebben. Jullie blijven leuk en ik ben blij dat jullie het nog steeds met me uithouden.
>
> Josjan, het is niet eenvoudig om te omvatten wat je voor mij betekent en in het kort te omschrijven waar ik je dankbaar voor ben (zou moeten zijn). De intellectuele uitdaging die jij constant biedt, de hoeveelheid boeken die we te bespreken hebben, het samen eten (en veel te veel wijn drinken), gamen, en zo voort en zo voort. Ons horizontale-verticale pact geeft houvast nu deze PhD is afgelopen, de volgende bijna, en ik nog geen zekerheid heb over de volgende stap in mijn (wetenschappelijke) carrière. Ik ben blij dat jij de leuke wil zijn bij mijn slimme, want met maar één PhD is er van het omgekeerde niet echt te spreken. Tot slot ben ik je dankbaar voor dit proefschrift. Jij stelde voor, tijdens een witbiertje op mijn balkon op de Suze Groeneweglaan, om samen onderzoek te doen en The Tell-Tale Brain van Ramachandran te lezen. Hoofdstuk 2 en 3 zijn hiervan het resultaat. Ik ga ervan uit dat we voorlopig niet ophouden met deze samenwerking.
>
> D&D gang - Arjan, Daan, Max (en weer Alex, Josjan, en Yorick) - jullie zorgen voor structuur, een wekelijkse activiteit die we nu al negen jaar volhouden. Ook al maken we zelden een campaign af, beginnen we steevast later dan gepland, en is er altijd wel wat te zeuren, blijft het leuk en gezellig. Ik hoop dat we dit voor altijd blijven doen.
>
> Het weeshuis - Ari, Anne, Casper, Charlotte, Els, Floor, Gerwin, Joanne, Jowi, Jur, Mat, Tommie, (en weer Arjan, Alex, Daan, Josjan, Max, en Yorick) – omvat al meer dan tien jaar mijn sociale leven; van uitgaan en Sinterklaas tot etentjes en vakanties. Jullie zijn een zootje ongeregeld, maar ik zou absoluut niet zonder jullie willen.

A few things I want to add. 1) A personal thanks to Gerwin. Going to the gym with you over the past year was great fun and motivating. Getting to know you better was also enjoyable. I hope we will continue this trend. 2) Rolex (Rozanna and Alex), in addition to the previous thanks, thank you for still being my friends and being my unofficial paranymphs and making my defense well visited and festive. Josjan, now that this PhD is finished as well and my academic career seems more secure, I am glad we are seeing each other more often and I look forward to our future endeavors, of which I am optimistic.

My family, of course I want to thank you as well. Since I also already thanked you in my previous dissertation (van Dongen, 2020, p. 138–139), I'll quote those 'thank you's here as well:

Jacqueline, Yorick en Sarah, bij jullie voel ik me altijd thuis. Het is altijd gezellig en fijn als we samen zijn. Bedankt voor jullie plagen, grapjes, en commentaar. Jacqueline, mama, al vind ik het nog steeds jammer dat ik niet de vrijheid kreeg om Terminator of Robocop actiefiguurtjes te kopen, ben ik je erg dankbaar voor de vrijheid en alle ondersteuningen die ik van jongs af aan heb gekregen om m'n eigen weg te kiezen. Ook Fedor ben ik dankbaar, hij is er dan wel al enkele jaren niet meer, het bovenstaande was het grootste deel van mijn leven met en door hem. Daarbij komt de frustratie van tegen hem Triviant spelen en zijn encyclopedische kennis. Mijn betweterigheid heb ik hoogstwaarschijnlijk aan hem te danken.

Familie – zonder uitzondering, iedereen van familie Loos en Van Dongen - jullie zijn de beste. Na iets meer dan dertig jaar heb ik redelijk wat families gezien; van veraf en dichtbij. Tot nu toe zijn jullie de allerleukste. Het is altijd gezellig, de gesprekken zijn interessant, en we hebben nooit ruzie (hoeveel families kunnen dat nou zeggen), verjaardagen en feestjes zijn altijd leuk, en er is altijd lekker eten en genoeg te drinken. Ik ben heel blij met jullie.

Ivanka, thank you once again. This is what I said in my previous dissertation (van Dongen, 2020, p. 139):

Ivanka, als aller laatst. Zowel persoonlijk en professioneel is het moeilijk voor te stellen wie ik zou zijn zonder je. Door jou ben ik het onderzoek in gegaan, jouw statistiekboeken hebben mijn leven verrijkt en zonder jouw kritische blik en geduldige correcties was mijn schrijfstijl nog steeds zo afgrijselijk. Zonder jou was ik vast een arme kunstenaar geweest. Ik ben je dankbaar voor het plagen, aanspreken op mijn gebreken, en de hulp bij het navigeren van de sociale wateren. Dat ik menselijk, attent en bescheiden blijf heb ik vooral aan jou te danken. Door jou kook ik risotto en ga ik op vakantie naar plaatsen waar ik anders niet zou komen. Het is geweldig om iemand te hebben die (bijna) net zoveel houdt van lekker eten, het niet erg vindt om een klein vermogen uit te geven in een restaurant, me naar harte lust laat experimenteren in de keuken, en een gedoogbeleid voert wat betreft te aanschaf van kookgerei en apparatuur. Je geeft me de ruimte voor al mijn rare hobby's en bent er altijd om voor me te zorgen, al vind je het wel verdacht fijn om dat te doen wanneer ik een keertje ziek ben (en ik doe alsof het niet zo is). Kortom, je bent uniek in het zolang uithouden met mij en ik vind het altijd fijn om thuis bij jou te zijn.

All of that is still true. I want to add that I am still grateful to have you and that I look forward to the next phases of our life together. I love you.