FACULTY OF PSYCHOLOGY
AND EDUCATIONAL SCIENCES

ON MENDING THE IRREPRODUCIBILITY OF PSYCHOLOGICAL SCIENCE

THE INSUFFICIENCY OF INCREASING SAMPLE SIZE, AND THE NEED FOR THEORY-INFORMED POWER ANALYSIS

Word count: 48,516

Nathan Laroy
Student number: 01810892

Supervisor(s): Prof. Dr. Beatrijs Moerkerke, Prof. Dr. Tom Loeys

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Psychology – main subject Theoretical and Experimental Psychology

Academic year: 2022 - 2023

GHENT
UNIVERSITY

Table of Contents

ABSTRACT

Efforts to address the replication crisis in psychological science have often overlooked a crucial factor: statistical power. Fortunately, this trend is changing, with publishers and funders starting to mandate statistical power analyses, the dissemination of primers growing, and recent reports starting to show statistical power in published research is actually increasing. Despite this progress, a systematic review presented here highlights that while the number of power analyses has increased between 2016 and 2021 across three domains and six journals, it remains unacceptably low overall. Causes and consequences are discussed, aiming to move beyond the stagnation which has been surrounding this issue during the last sixty years. The proposition of simply increasing sample sizes to enhance replicability and statistical power is scrutinized, revealing ethical and practical concerns that render it an unsuitable solution. Specifically, this approach risks perpetuating ritualistic scientific practices in psychology, thus deepening the actual underlying crisis of theory. To truly transform statistical power analysis from ritual to informative practice, research psychologists must confront their incessant adherence to ill-conceived rules of thumb, mindless statistical proceduralism, and ambiguous verbal theorization. Embracing formalism is posited as the essential path forward, enabling statistical power analyses to hold genuine scientific value.

NEDERLANDSE SAMENVATTING

Inspanningen om de replicatiecrisis in de psychologische wetenschappen aan te pakken, missen al te vaak een cruciale factor: statistische power. Deze trend is gelukkig aan het veranderen: uitgevers en financiers beginnen stilaan statistische poweranalyses verplicht te stellen, de verspreiding van powerprimers neemt toe en recente rapporten tonen hoe statistische power in gepubliceerd onderzoek daadwerkelijk lijkt toe te nemen. Ondanks deze vooruitgang blijkt uit een hier gepresenteerde systematische review dat, niettegenstaande het feit dat het aantal poweranalyses tussen 2016 en 2021 weliswaar is toegenomen over drie domeinen en zes tijdschriften heen, de specifieke aantallen over het geheel genomen onaanvaardbaar laag blijven. Oorzaken en gevolgen worden besproken, met als doel de stagnatie omtrent dit onderwerp van de afgelopen zestig jaar te doorbreken. Het voorstel om simpelweg de steekproefgrootte van onderzoek te vergroten om de repliceerbaarheid en statistische power van onderzoek te verbeteren, wordt onder de loep genomen. Hierbij komen ethische en praktische bezwaren naar voren, die dit tot een ongeschikte oplossing maken. Specifiek dreigt zo'n aanpak ritualistische wetenschappelijke praktijken in de psychologische wetenschappen te zullen gaan bestendigen, waardoor de onderliggende theoriecrisis nog dieper wordt. Om statistische poweranalyse werkelijk te kunnen transformeren van wat dreigt tot ritueel te verworden naar een informatieve gebruikspraktijk, moeten onderzoekpsychologen hun onophoudelijke vasthouden aan ondoordachte vuistregels, achteloos statistisch proceduralisme en dubbelzinnige verbale theorievorming onder ogen zien. Formalisme omarmen wordt aangevoerd als de essentiële weg voorwaarts, waardoor statistische poweranalyses echte wetenschappelijke waarde zullen krijgen.

ACKNOWLEDGMENT

Contrary to what the length of the current thesis may suggest, I generally do not like to overstay my welcome. For this reason, I would like to keep this acknowledgment section rather short. I thank my promotors for giving me the chance to delve into a topic of my desire, and approach it in a way that I see fit. I thank my supervisor, Lara, for making time for me and trudging through my long messages and questions. I would like to thank professor Paul E. Meehl, whom I have never had the pleasure to meet, but whose knowledge and wit have sparked within me a thirst for science, through his publications and the recordings of his lectures on philosophy of psychology, which are freely available online. And, finally, I would like to thank any brave reader who chooses to read my ramblings herein put together under the guise of a Master's thesis. I hope you find it interesting, perhaps even amusing. It has been a work, not so much of love, but of mere stubbornness.

INTRODUCTION

> *"Science is difficult, and anyone who believes that it is easy to gather good scientific data in a discipline like experimental psychology is probably doing it wrong." – G. Francis (2012, p. 15).*

Ever since Bem (2011) published his account on *psi*, a theoretical entity encompassing cognitive functions such as precognition and premonition, a significant rise in publications on the topic of what is now known as *the replication crisis* has taken place. The claims published in *Journal of Personality and Social Psychology* (JPSP) on some kind of (sub)conscious cognitive and/or affective divination were met with incredible scepticism at the time, which stimulated researchers active in the 2010s to conduct a wide array of replication studies. The main goal of this enterprise was to unambiguously refute the implausible claims reported in Bem's (2011) paper (e.g., Galak et al., 2012; Ritchie et al., 2012; Rouder & Morey, 2011; Wagenmakers, Wetzels, et al., 2011).[1] To wit, the claims made by the author are impossible *a priori*, because they violate the laws of thermodynamics (Reber & Alcock, 2020). It is therefore highly probable that Bem's (2011) claims—and, by extension, those made by other parapsychological researchers—are the result of employing improper methodology, or perhaps even the result of outright scientific malpractice. For example, Schwarzkopf (2013) enumerates several problems in parapsychological literature (taking a meta-analysis as case study), including misinterpretation of analytical artifacts and a deep-rooted misconception of the nature of scientific method and statistical inference. Surprisingly, the original publisher, *Journal of Personality and Social Psychology* (JPSP), a leading empirical outlet of social psychology (Finkel & Baumeister, 2019), initially refused to publish any direct replication studies, because they were deemed unimportant (Aldhous, 2011).

2011 is considered by some to be a pivotal year, during and after which it has become increasingly clear that the knowledge base generated by years of psychological science is built—at least in part—on shoddy methodology and poor statistics. In a seminal paper, Wiggins and Christopherson (2019) summarize that, at the time, four crucial elements played a role in exposing the apparently inherent unreliability of psychological research: 1) the distress caused by the fact that Bem's (2011) paper had been published in a major outlet, 2) the initial obstinacy displayed by JPSP in dismissing the importance of scientific self-correction via replication, 3) the recent discovery of D. Stapel's egregious research fraud (see Callaway, 2011), and 4) a general increase in awareness of questionable research practices (QRP; Simmons et al., 2011). These elements are

[1] The account of *psi* and other parapsychological 'findings' predate Bem's (2011) publication (see, e.g., Bem & Honorton, 1994). In fact, parapsychology as a laboratory 'science' has long been around (e.g., Rhine, 1934) and has been the subject of heavy criticism ever since its conception. It has received notable scrutiny in terms of the applied statistics (Heinlein & Heinlein, 1938; Utts, 1991), the apparent lack of scientific standards of experimentation (Girden, 1962; Moss & Butler, 1978), and its ontological foundations (Reber & Alcock, 2019).

proposed to have set the stage for a movement of increasingly distrustful academics and researchers attempting to replicate long-standing findings (or ‘effects’), most of which were considered canon in the psychological literature. As it turns out, several of these ‘effects’ were unable to stand their ground against this sudden wave of scrutiny (Klein, 2014). Examples include *social priming* (Bargh et al., 1996; e.g., Doyen et al., 2012; Pashler et al., 2012), *the facial feedback hypothesis* (Strack et al., 1998; e.g., Wagenmakers, Beek, et al., 2016), and *ego depletion* (Baumeister et al., 1998; e.g., Hagger et al., 2016). The notion of *crisis* was enforced by methodologists and scientists in general in subsequent years. Specifically, a growing body of literature included publications on the exact nature of the replication crisis, the flaws of scientific methodology thus far employed by psychological science professionals, and the consequences with respect to the validity of past publications (Shrout & Rodgers, 2019). Most notably, the publication of the Open Science Collaboration project, which generated a reproducibility estimate for psychological science at large (Open Science Collaboration, 2015) has arguably solidified the notion of *crisis*. The seminal work concludes, among other things, that less than 40 % of their replication efforts have yielded statistically significant results. Formerly statistically significant data ceased to be so when novel data was introduced to the original dataset, replicated effect sizes (ES) were only half as large as those originally published, and a failure to replicate was virtually always related to the original study’s methodological design.

Throughout the years, it has become clear that psychological sciences suffer, as a whole, from this so-called replication crisis. Social psychology is viewed by many as *ground zero*, the epicentre of the replication crisis (Finkel & Baumeister, 2019), but—as will become clear in this thesis—, it is certainly not the sole culprit. Literature on the causes of this apparent replication crisis is vast and rich, and a sizeable amount of it is occupied by the topic of QRPs. These questionable practices—that is, methodological and statistical practices, carried out in a systematic, though not necessarily deliberate fashion—are said to inject large numbers of false positive results in the literature; false *positives*, since numerous attempts at replication *fail to yield* similar outcomes. Thus, curative efforts tend, historically, to focus on minimizing the probability of finding a false positive—or, using statistical jargon, of committing a type I error (Banks et al., 2016). However, this historical preoccupation with minimizing false positives has long drawn attention away from an equally deleterious practice: a consistent failure to adequately take into account the role of statistical power and informed sample size determination in published research, both by individuals performing the research, as well as by their peers, whose role it is to evaluate and critically interpret published findings with due incredulity and care. This negligence is especially worrying, given how low statistical power can also yield increased type I error rates in published literature. Fortunately, this is changing. Statistical power has been gaining more attention, for example, as more and more funding bodies and publishers explicitly demand formal *a priori* power analyses to be conducted and reported. Equally promising is the steady increase in the publication of so-called power primers, the goal of which is to provide extensive instructions for researchers on how to perform power analyses in specific analytical circumstances. Prominently at the forefront of the scientific reform movement, an arguably simple solution to the crisis is often proposed: to minimize the probability of reporting false positives and, at the same time, ensure decent statistical power, sample sizes must increase.

The current thesis aims to do two things. Firstly, the following central research question is tackled: has statistical power actually gained a more prominent role in recent years, as evidenced by the inclusion of *a priori* power analyses in published research? A comparison between research articles published in 2016 and 2021 will show that, overall, reporting of *a priori* power analyses has increased, but the absolute numbers remain remarkably small and there are notable differences between subdisciplines. Secondly, in the discussion section, the current thesis aims to contest the proposition that merely increasing sample sizes is a good solution to the posed problems. Specifically, it is argued that increasing sample sizes—which would increase statistical power overall, and reduce type I error rates—constitutes a mere continuation of ritualistic scientific practice, sustained by uncritical use of statistics, poor methodology, a publish-or-perish academic culture, unsuccessful dissemination of statistical know-how, poor theorization, and the actual difficulty of performing proper power analyses in practice (especially in terms of determining a minimal ES of interest). Instead, an argument is made to go beyond ritual and attack the problem at a more fundamental level. Specifically, it is argued that power analysis and sample size determination must be grounded in theory instead of rules of thumb. Finally, a reappraisal of slow science, stepping away from result-centric scientific practice and an improved dissemination of insights from philosophy of science are proposed to guide and focalize curative efforts away from the periphery of the problem, and toward the source of the replication crisis.

To carry out all of the above in a reasoned manner, it is deemed beneficial to first take a closer look at the underlying problem being addressed: the replication crisis. What is it? What are its origins? How is it sustained? In what sense is failing to replicate a published finding exactly problematic? How expansive is this crisis, such that it has become worthy of the notion *crisis*? What is the epistemic purpose of replication in the first place? In the coming paragraphs, these are the questions which will be addressed—some more comprehensively than others—, in order to set the stage for the central topics of the current thesis: power analyses and sample size justification in psychological research.

THE REPLICATION CRISIS

The replication crisis is a challenging subject for a multitude of reasons. The main problem is that the epistemic function of a replication—and, in a broader sense, replicability—is not very clear. This renders a clear qualification of the notion of *crisis* challenging to achieve. In fact, the existing typology of different kinds of replication is diffuse, which makes it even more challenging to ascertain its epistemic function(s). One serious deficit of psychological and meta-scientific literature at large is the absence of uniform definitions of the terminology which is often employed whilst discussing replication-related topics. Some authors employ concepts such as *replicable*, *reproducible* and *repeatable* interchangeably, whilst others adhere to definitions with clear, though perhaps idiosyncratic delineations. Furthermore, the typological diffusion of different notions of replication has led some authors to propose terminological simplifications (e.g., *direct* vs *conceptual* replications; see S. Schmidt, 2009), thus denying a sense of nuance *a priori* and causing an inadvertent disregard for valid replication efforts which are not recognized as such (Haig, 2021). These challenges notwithstanding, the current thesis cannot advance without a clear idea of what replication is and ought to achieve—even if only in a rudimentary sense. Only then can a wieldy qualification of the *crisis* be formulated, and the role of statistical power in it be addressed.

THE MANY FACES OF REPLICATION

Replication is commonly viewed as the tenet of scientific knowledge accumulation. In a rudimentary sense, replication serves as a kind of confirmation criterion for the results of experiments. It helps us guarantee that research results are representative of an external reality; they give us more confidence that we are not fooling ourselves with specious data (Popper, 1959). At the same time, unsuccessful attempts at replication are believed to cast doubts on original findings. As such, unsuccessful replicability may lead researchers to believe that the empirical information they had previously gathered was nonrepresentative, the unfortunate consequence of a noisy process. A lack of replication is believed to result in the amassment of negligible factoids. Consider, for example, Fishman and Neigher (1982), who decried that contemporary psychological literature was rife with “ecologically irrelevant, single-study experiments with data that are unreplicated, underaggregated, and biased” (p. 542). This, according to the authors, has resulted in the publication of a lot of scattered information, which can hardly constitute a basis for systematic knowledge accumulation. In a similar vein, Lykken (1968) has argued that replicability tends to be a characteristic trait of ‘good’ research—implying that irreplicable research tends to be ‘bad’. In summary, replication studies are believed to help scientists separate the grain from the chaff (Schlosberg, 1951). But what actually is *replication*? And in what sense is it a prerequisite to scientific knowledge accumulation? Answering this question is unexpectedly challenging, especially given how explicit definitions of replication are historically rare—a state of affairs which Schweizer (1989; as cited in S. Schmidt, 2009) argues, exists because its meaning seems *obvious; quod non*.

In the early days of psychological science, replication has often been implicitly equated to *repeatability*. Specifically, the idea that repeatability of experimental outcomes is not just desirable, but necessary to accumulate scientific knowledge was heavily underscored by several prominent figures of the discipline. For example, in a paper on methods of scientific psychology, Wundt (1907) argued that one of the four *Grundregeln* or basic principles of proper methodology is that *"Jede Beobachtung muß zum Zweck der Sicherung der Ergebnisse unter den Gleichen Umständen mehrmals wiederholt werden können"* [for the purpose of securing results, each observation must be able to be repeated several times under the same circumstances] (p. 308). He also mentions that circumstances which keep one from re-observing a phenomenon necessarily restrict the *Sicherheit* or reliability of an experiment's results. That is, reliability of outcome measurements by virtue of the repeatability of the procedure which yields them is what supposedly makes the scientific enterprise different from a pseudoscientific one—or, at least, that is *one* such factor of differentiation (see Hansson, 2021). Other authors have put forward starker notions of the importance of repeatability, where the notion is promulgated in association with truth claims, or even 'proofs' of hypotheses. For example, in a methods paper, Dunlap K. (1925) puts forward the notion that "proof established by [a hypothesis] test must have a specific form, namely, repeatability" and that "[n]othing is accepted as proof, in psychology or in any other science, which does not conform to this requirement" (p. 503). A practical example of this idea can be found already in the late twenties of the twentieth century, where repeatability of experiments as fundamental to scientific 'proof' was used as an argument to dismantle telepathy and clairvoyance literature; e.g., "proof in science is repeatability, and every time we attempt to repeat these experiments [on telepathy] in the laboratory we have a dismal failure" (Estabrooks, 1929, p. 211). Marquis (1948) drives the supposed necessity of repeating experiments to a further extreme, stating that any phenomenon subject to inquiry "must exist in replication" and that "[i]f we are interested in what appears to be a unique situation [...], it is necessary to reformulate the problem in terms of those aspects of the situation which can be identified in several instances" (p. 433; also Stevens, 1939). Taken together, these early accounts indicate the historically entrenched nature of the belief that repeated observation is a necessary element of proper psychological science practice (for early applied examples in psychological research, see Barr, 1932; Mead, 1917; Peters, 1938; Reed, 1917).

Similarly, meta-scientific literature of the time has traditionally ascribed specific importance to repeatability as well. An early example of this can be found in Whewell (1858), who states that "the hypotheses which we accept ought to [...] *foretel* (sic) phenomena which have not yet been observed; at least all phenomena of the same kind as those the hypothesis was invented to explain" (Book II, Chapter V, Section III, Article 10). In this sense, replication via repetition is a process of iterative corroboration, i.e., its goal is to ascertain the robustness of research findings, by virtue of analysing the invariance of the outcomes of independent investigative processes (see also Wimsatt, 1981). Thus, repeatability is a requirement for accepting hypotheses, because repeatability implies predictability, and the latter is a core attribute of substantive *testable* hypotheses. Note that this assertion presupposes that a theoretical claim from which a hypothesis is derived is sufficiently formal or mechanistic, to the degree that it actually provides a specific expected outcome measurement; for example, a specified amount of water displacement from a reservoir when an object of specified mass is put in it. In a similar vein, but from a falsificationist viewpoint, Popper (1959) has defined scientifically significant effects as those which may be *regularly reproduced* by means of a prescribed methodology, and states that "we shall take [a theory] as falsified only if we discover a *reproducible* effect which refutes the theory" (p. 203; reproducible is here meant as repeating a procedure with new data to assess the reliability of its outcome). He further stated that so-called *occult effects*—i.e., effects for which there exists no method for recreating the initial conditions which have yielded them—are simply not scientifically significant. In fact, the requirement of experimental repeatability is traditionally a central feature of all empirical, so-called *hard sciences*, and a significant portion of psychological science's history is exemplified by the adamant wish to emulate the *hardness* of the natural sciences (Farrell, 1978; James, 1892; Sterrett, 1909; Watson, 1913)[2]—perhaps to its detriment (Hughes, 1930). For example, Symonds (1928) writes that "the development of the natural sciences depended on the development of […] exact measurements, and the development of psychology as a science likewise depends on the perfection of its measuring instruments" and, without hesitation, adds that "[m]uch of the recent work in the development of tests, particularly in the measurement of personality, is practically worthless because the tests do not tell a consistent story" (p. 73). Lindsay and Ehrenberg (1993) carry out a similar attitude in saying: "It is hardly worth asking why something occurs, or how to apply it in practice, if we are not sure whether it can be observed at all, let alone routinely" (p. 218). This attitude is further reflected in the early development of techniques which allow researchers to quantify the reliability of a series of measurements (e.g., Dunlap J. W., 1933; Symonds, 1928; Remmers & Whisler, 1938; Spearman, 1913). Finally, it has even been argued that repeatability ought to be a criterion for publication. For example, Lubin (1957) has stated that under a system which rewards proven repeatability of findings, "the quality of replication (and generalization) designs would improve, and a great deal of over-elaborate statistical analysis will disappear", and that "replicability is a *sine qua non*" (p. 520;

[2] Psychology as a field has struggled with its scientific status all throughout its existence; during its nascence (e.g., James, 1892; Ladd, 1892) and during the twentieth century (Bowlby, 1984; Copeland, 1930; Piaget & Kamii, 1978; Watson, 1913). A significant portion of this debate can be attributed to the uncertainty about what exactly is the object of psychology (psyche, mind, or behaviour; Ardila, 2007). However, a more fundamental reason is likely the simple fact that psychology as a discipline had grown out of physiology and anatomy, a blooming field and the educational background of Wilhelm Wundt, grandfather of psychological science (Danziger, 1990).

emphasis in original; see also Furchtgott, 1984; though see Lykken, 1968; Pereboom; 1971, for dissent on practical grounds).

However, repeatability is not enough. Specifically, mere repeatability is not a sufficient criterion for making valid epistemic claims, i.e., for inferring to the truth value of the hypothesis or theory for which an experimental test was devised. Repeatability has its uses, specifically for those objects of scientific inquiry which, indeed, occur recurrently, or for which the initial conditions can be realized so as to manually elicit its recurrence. This use is mostly, if not uniquely, in the form of guaranteeing a level of precision of measurement—i.e., that the measurement was not a false positive (Zwaan et al., 2018)—granted that the measurement technique itself is valid. It is "particularly suitable early in a program of research to establish quickly and relatively easily and cheaply whether a new result can be repeated at all" (Lindsay & Ehrenberg, 1993, p. 221). Also, repeatability as a tool for reducing uncertainty has a distinct translation in Fisherian statistics, a forebearer of what is currently known as *null hypothesis significance testing* (NHST). Specifically, Fisher (1958) argues that repeated, independent measurements play a central role in guaranteeing a level of precision regarding the results of an experiment, by "diminishing the error to which they are subject", especially since it is "the only means of estimation of such error" (p. 153; see also Fisher, 1933; 1935). Similarly, it was argued that the establishment of repeated instances "is fundamental for the validity of […] tests of significance" (Baxter, 1940, p. 497) in their application to said instances. Interestingly, however, in Fisher's time—grossly speaking—the statistical function of repeatability was mostly devised by common sense, while formal theory on such error estimation from the results themselves was mostly lacking (spare for the simplest case of a comparison of two treatments; Yates, 1964; see Student, 1908). Regardless, repeated measurements of the same outcome variable using the same method do little to inform us about the *validity of theoretical claims*; it tells us only that something *is repeatable*, and reliably so (Irvine, 2021). Equally, it is not because something *cannot* be replicated, that it is therefore necessarily false (Buzbas et al., 2023; Devezer et al., 2019). Consistency over repetition indicates law-likeness, but not all objects of scientific inquiry, especially in the social sciences, behave according to empirical laws. In fact, the singular exception to this statement from psychology is probably Weber's law, which arguably borders more on the knowledge domain of physics or physiology than psychology. Note that repeatability is also not necessarily indicative of *measurement validity* either (i.e., the measurement tool accurately measures what it is proposed to measure). The reader is reminded of the following aphorism, which many a psychology student has surely heard in psychometrics 101: reliability does not imply validity (F. Schmidt et al., 2000); a measurement instrument may be reasonably reliable, in that it consistently measures something with acceptable accuracy, but the thing being measured is therefore not necessarily guaranteed to be the intended construct. Furthermore, limiting the coerciveness of scientific evidence to the repeatability of its constituents necessarily implies that rare and unmanipulable phenomena are lesser or cannot be subjects of scientific inquiry, which is obviously false (Quay, 1974). If it were true, this would render a large portion of psychology unfit for scientific inquiry, for example, single-case studies, field research, or intensive longitudinal studies. In fact, it would render large portions of those natural sciences which have historically been so ardently pedestaled by psychology researchers unfit for scientific inquiry; for example, cosmology, whose domain of knowledge constitutes by definition the study of unique, unrepeatable events (Rees, 1980).

One may assume that early twentieth-century psychology researchers were at least somewhat aware of the difference between mere repetition of empirical observations and the act of ascertaining the validity of the hypothesized construct under investigation. Nonetheless, explications of this duality are rare to find in most of twentieth century psychology literature, at least not until the 70s and 80s. This is curious. As argued, there seems to be historical consensus on the idea that replicability as repeatability is a classic tenet of scientific knowledge accumulation and abductive inference, in what one may nowadays call a belief system of *reliabilist justification* (Haig, 2021; see e.g. Goldman 1988). However, reliability is hardly a criterion for truth. Replication must involve more than repeatability, in that repeatability is only one of many tools in a scientist's possession for truth-finding. But the broader notion has historically received little systematic or formal scrutiny in specialized literature (Schickore, 2011; Steinle, 2016). That is, the role of replication is presented by many as self-evidentiary, at the cost of the development of a detailed notion of *what* replication ought to achieve (i.e., what exactly is its function in the making of valid truth claims), and *by which mechanism* said achievement ought to be realized. To be clear, this statement does not imply that philosophers of science and statisticians of the twentieth century—say, pre-1970s—did not have at least some implicit understanding or specify to some degree *what* replication accomplishes (see Steinle, 2016, for some historical examples)—if pressed on the issue, many would likely have understood the gravity of the question; however, any formal conceptualization of the notion of replication was simply absent from standard literature (i.e., via formal mathematical, logical, or computational models explicating how replicating experiments ought to aid in scientists' uncovering of 'truth'). Schickore (2011) argues that at least one reason for this state of affairs is twentieth-century philosophers of science's preoccupation with "[m]ethodological discussions [centred] on inferential relations between evidence and theory, but rarely on the question of *how* empirical evidence is validated" (p. 529; emphasis added). It is only recently that the field of meta-science has started to look at this particular question concerning replication and what it's role exactly is in terms of producing valid inferences to truth claims (e.g., Baumgaertner et al., 2019; Buzbas et al., 2023; Devezer & Buzbas, 2022; Devezer et al., 2019; Haig, 2021; Irvine, 2021; Ulrich & Miller, 2020; Witt, 2019).

Preliminary steps to this effect were undertaken by the gradual development of a typology of replication. One of the first accounts was presented by Lykken (1968), who differentiated between three kinds of replication: *literal replication*, which entails the exact duplication of an original experimental procedure; *operational replication*, which entails an attempt to redo an experiment simply by following the provided steps in a published paper; and *constructive replication*, which entails avoiding mere repetition in favour of devising a new way of testing the truth claim proposed by another researcher. Literal replication according to Lykken's (1968) definition, represents what has been discussed so far as repeatability. Operational replicability resembles more an exercise in scientific communication, but still serves as a reliability check; this kind of replication shows that "the same findings can be obtained in any other place by any other researcher" (S. Schmidt, 2009, p. 90) and, as such, provides evidence that scientists are entitled to the belief that this particular finding exists independently of themselves. Constructive replication is nowadays better known as *conceptual replication*, which is ideally a form of methodological triangulation (Haig, 2021): employing different methods—procedures, operationalizations—of probing a particular variable of interest, in hopes of arriving at similar or equivalent outcomes, such that a consistency of results is indicative of the veracity of a theoretical claim, whereas an inconsistency rules against it. In this sense, conceptual replicability of a theoretical claim's predictions is a constitutive property of its validity. Lykken (1968) emphasises the need for constructive replication through the following aphorism: "We are interested in the *construct* […], not in the *datum*" (p. 156). A similar typology was put forward by Radder (1992), who spoke of kinds of *reproduction*. Again, a distinct type is suggested to encompass mere repetition, i.e., "the same actions are performed and the same experimental situations produced from the point of view of the daily language description of the material realization of the experiment" (p. 65). Note how the emphasis on reproduction via a verbal description of procedure is reminiscent of Lykken's (1968) notion of *operational replication*. Note also how, if a verbal description is nonspecific, the extent of this lacking specificity causes the attempted replication to fall within the semantic confounds of a constructive replication, for it necessarily deviates from the original. Additionally, Radder (1992) defined two more types of reproduction, which are focused on a theoretical interpretation, or an experimental result. The idea of reproducing an experiment given a fixed theoretical interpretation is reminiscent of Lykken's (1968) notion of *constructive replication*—to quote the latter: "To obtain an ideal constructive replication, one would provide […] *nothing more than* a clear statement of the empirical 'fact' […] and then let the replicator formulate his own methods of sampling, measurement, and data analysis" (p. 156; emphasis in original). For example, S. Schmidt (2009) puts forward the case of Einstein's hypothesis that there is an upper bound on the speed of light (a proposed empirical fact), a claim which can be tested using a number of experimental set-ups, but the outcome variable of which is identical across experiments.

This is opposed to reproduction of an experimental result *per se*, for which different theoretical interpretations/descriptions exist; for instance, Radder (1992) provides the example of the assessment of Avogadro's number, which was conducted by reproducing the same essential conclusion using a set of distinctly different theoretical approaches, each of which may be represented by a distinct composite description "of all kinds of premises that are necessary for drawing the conclusion that *q* [the intended experimental result] is the result of the overall experimental process" (p. 64). The difference between replication of a theoretical interpretation and replication of an experimental outcome *per se* deserves emphasis, because the latter is a far stronger approach to theory adjudication than the former. Let's retain the example of Avogadro's number: the determination of this constant was so compelling, because it was done using "a variety of *very different* experimental situations involving *very different procedures*, [...] which require both independent skills and independent assumptions" (Cartwright, 1991, p. 149). The difference in inductive strength lies thus in the fact that replicating an experiment based on the same theoretical assumptions implies that the outcome may still be reliably illusory within the confounds of those assumptions, while a replication based on entirely different theoretical assumptions is able to move beyond the inductive restrictions posed by one approach and allows multiple approaches to coincide on an empirical fact (see Salmon, 1984, for an elaborate discussion; see Hudson, 2023, for an example from psychological literature).

Replications which are not mere repetitions are said to allow a systematic approach to corroborate hypotheses in a broader context, i.e., one that is not restrained by particular methodological operationalizations (and, ideally, not by a particular set of theoretical assumptions; see example of Avogadro's constant). Furthermore, these kinds of replications address the main shortcoming of mere repetition which was explicated above: there is no way to derive from a(n) (un)successful experiment repetition whether or not the measurement itself was valid. To overcome this catastrophic weakness, replications of the sort described by Lykken (1968; *constructive replication*) and Radder (1992; *replication within and without a particular set of theoretical assumptions*) are paramount. What makes such a replication informative with respect to theory construction/adjudication, is the fact that, ideally, the employed measurement procedure presupposes a causal pathway "to access the value of the target [outcome] by, for example, using different instrumentation and/or different ways of experimentally intervening on the target" (Irvine, 2021, p. 2). It allows to "bolster and extend a theory" (Derksen & Morawski, 2022, p. 1491), and exists on a spectrum, defined by how causally independent different studies are (Irvine, 2021; Radder, 1992). The philosophy behind this mode of reasoning about evidence is broadly *coherentist* (Haig, 2021), in that belief in one object is deemed valid by its consistency with other constituents of said belief, and to other beliefs, by extension. Or, as S. Schmidt (2009) puts it, "[w]ith every difference that is introduced, the confirmatory power of the replication increases"

(p. 93). However, importantly, the valid use of a replication of the aforementioned kind rests entirely on the specificity of the underlying theory being tested. If a theory is substantially nonspecific, then it doesn't allow for what Meehl (1990) famously coined a "derivation chain" from theory to empirical fact, i.e., "a conjunction of theoretical and auxiliary premises that are necessary to predict observable outcomes" (Scheel, Tiokhin, et al., 2020, p. 3). The point is best described by Meehl (1990) himself: "To the extent that the derivation chain from the theory and its auxiliaries [i.e., supporting hypotheses] to the predicted factual relation is loose [i.e., not "deductively tight"], a falsified prediction cannot constitute a strict, strong, definitive falsifier of the substantive theory" (p. 200). It follows, then, that the same must hold for the adjudication of a theory or between theories; performing a replication that is causally independent in Irvine's (2021) terms requires that there is a clear and unambiguous formal understanding of how a theory (or set of theories, if one aims to perform a replication of the third sort described by Radder, 1992) mechanistically elicits the outcome of interest. If these relations are not deductively tight, and the auxiliary hypotheses (e.g., concerning instrumentation) are not explicated, than it is impossible to derive a valid conclusion from such a replication effort to the theory or a valid adjudication between theories.

A number of related typologies besides those proposed by Lykken (1968) and Radder (1992) have been put forward throughout the years, and enumerating all of them here would be unwieldy (but see S. Schmidt, 2009; Hudson, 2023; Hüffmeier et al., 2016; Tsang & Kwan, 1999). Throughout the years, several of these have been homogenized into two distinct groups, labelled *direct* and *conceptual* replication (S. Schmidt, 2009). Notably, this subdivision is made purely in reference to methodological procedure; i.e., a direct replication constitutes a duplication of methodological procedure as a whole (from data acquisition to analysis), while conceptual replication constitutes a deviation from procedure. As such, tiny or larger deviations from procedure, in terms of target population, instrumentation, or analysis, are often grouped into the same overarching *conceptual* type. This dyadic typology is by far the one most employed in contemporary replication literature. These types of replication are likely known by the majority of researchers across the sciences, but more nuanced subdivisions to complement these larger groups have been proposed as well, including, for example, *internal* and *micro* replication. An internal replication is closely related to Lykken's (1968) notion of *literal replication*, in that it adopts "the same methods, sample sizes, [and] data analyses" (Haig, 2021, see 'varieties of replication'), but it is different in the sense that internal replications occur exclusively within one study or research program. In a sense, it is a form of locally standardized methodological practice amounting to self-replication, where consecutive such replications build on the previous and are bundled in a study or program (Bamberger, 2019). Micro-replications are small replications of previous work, the goal of which is to "pick out one aspect that is crucial in guiding the experiment at hand and make it part of the current set-up" (Guttinger, 2019, p. 467). They function as validators of new designs, as positive controls that an experimental object of interest behaves as it had been reported in previous literature—and, as such, they are constantly being replicated—, after which they can be used to generate new knowledge (Guttinger, 2018, 2019; see also Devezer & Buzbas, 2022). Importantly, these types of replications are rarely declared as such, yet they are essentially replication-related. The plurality of subdivisions and proposed typologies is further accentuated by authors like Radder (1992), who introduces further categorizations based on who the replicator is (from original researchers to scientifically untrained laypeople), which results in his proposal of *twelve different kinds* of replication. But, as noted before, these smaller subdivisions are essentially ignored by the majority of working psychology scientists. Generally speaking, the discussion occurs almost always in reference to S. Schmidt's (2009) proposed dyad of *direct* and *conceptual* replications.

There are advantages and disadvantages to this reductionist typology of replication studies. Haig (2021) argues that, although there are many forms of replication which correspond to either of the two presented kinds more in terms of degrees than in a categorical sense, still, most of contemporary psychology's replication efforts are, in fact, roughly divisible along those types. As such, the rough division allows for relatively fruitful debate on the topic of replication, specifically in those terms which most researchers already understand. On the other hand, the conceptual paucity of this distinction in two types has caused several misunderstandings. First of all, the distinction has caused the unfortunate rise of discussions on which kind of replication is preferrable *overall* (e.g., Crandall & Sherman; 2016; Lynch et al., 2015; Stroebe & Strack; 2014; see Hudson, 2023); but, of course, both types are valid in their own right and applicable in certain situations, and they cannot exist without each other. For example, it has been argued that without prior checking of the reliability of an effect via direct replications, successful conceptual replications might actually be unwittingly false (Chambers, 2017). Secondly, the term *direct* has itself caused trifling debate at the semantic margins of what a replication can or cannot be. It is a trivial matter that *direct* replications in the sense of *identically mimicking an original* cannot truly be carried out, simply because one cannot account for all possible variables which have made up the circumstances from which the original came about. Furthermore, Fabrigar and Wegener (2016) point out that in psychological science, original instrumentation and operationalizations are often created with a specific population in mind, so a 'direct' replication on a sample drawn from an ever so slightly different population might as well not regenerate the psychological phenomenon of interest. In fact, any undisclosed variable or circumstance of an original might technically prohibit third party researchers from carrying out *identical* replications unless they were part of the original research team—but this statement itself presupposes that the original authors are aware of all consequential context variables, which they are likely not. As such, Irvine (2021) concludes that any replication exists on a spectrum of conceptual replications, where some replications are *more direct* (sometimes coined *close* replications) and others are *more conceptual*. Steiner et al. (2019) show how one can exploit this property by adhering to a "prospective replication approach" (p. 285), where attempts at replication are performed by systematically violating assumptions that are implicitly made when no replication yet exists; that is, the interpretability of any single study rests on a set of assumptions which imply *a priori* that it is replicable and the findings generalizable (e.g., no hidden variation in treatment variables, identical distributions of characteristics of different populations, unbiased estimation of causal estimands, *et cetera*; see Steiner et al., 2019), but these assumptions need to be checked by systematically varying them across replications, that is, by systematically constructing increasingly conceptual replications. Indeed, this may be especially true for psychological science, where empirical effects are usually sensitive to context and unstable over time (Gergen, 1973). Degree of directness is then defined in terms of causal independence between studies—i.e., independent along the lines of what Radder (1992) understands as a replication approach based on different theoretical assumptions, auxiliary hypotheses, procedure, *et cetera*. The goal of a closer replication is to focus on psychometric invariance (Fabrigar & Wegener, 2016), while the goal of more conceptual replication is generalization, extension of findings to different circumstances, and a broadly coherentist approach to theory justification (Thagard, 2007). The informativeness of replicability thus lies not in any single replication, but in the aggregate of a multitude of reports. Scientific knowledge is gained by examining more precisely the discrepancies and similarities between an original study and attempts at closer and more conceptual replication, to separate signal from artifact and noise, to explore what is the underlying, supposedly stable phenomenon at play, and which variables affect our measurement of it in a way which merits iterative adjustments to theory and experimental procedure (Buzbas et al., 2023; Chang, 2004; Irvine, 2021).

The current thesis will adhere to the notion of replications existing on a spectrum rather than being dichotomously categorizable. Nonetheless, for simplicity's sake, frequent usage will be made of the terms *close* and *conceptual* replication from this point onwards, to refer to replications which are, respectively, either closer to or more deviating from an original study in terms of instrumentation, operationalization of variables, sampled population, theoretical assumptions, *et cetera*. In summary, the above overview has explicated the historically entrenched preoccupation with repeatability in psychological science. It was explained how the main epistemic function of repeatability is how it enables researchers to assess reliability of experimental outcomes—a goal which may be summarized under the mantra "Trust but verify" (Simons, 2014). The nuance was added that systematic incremental deviations from original research allow to perform internal replications which nonetheless extend knowledge on the circumstances which are or are not conducive to eliciting the effect of interest. However, it was also addressed that this form of repeatability is itself an insufficient criterion for truth claims, and that what is now known as conceptual replications is needed to corroborate theories and adjudicate between them; especially conceptual replications which are based on different theoretical assumptions can serve this goal. In tandem, a gross overview of replication typologies was provided. The epistemic functions of replication are now deemed to have been outlined to a degree which should enable us to propose a proper qualification of the alleged *replication crisis*. Doing so will allow the central topic of this paper—i.e., the role of statistical power analyses and sample sizes in the sustainment of the crisis—to be addressed relative to an unambiguous referent.

A WORKING QUALIFICATION OF THE REPLICATION CRISIS

Qualifying the notion of *replication crisis* in order to extract from it an unambiguous, focussed referent of *the problem* is exceedingly difficult—perhaps even impossible. First of all, there is no agreement on what constitutes *a crisis* in a scientific field, and from the way it is used in psychological literature, one can only conclude that "it is fair to accuse […] authors of being imprecise in their word choice" (Goertzen, 2008, p. 830). Secondly, even though the current crisis is denoted as a *replication crisis*, this does little to clarify what is meant by it, because—as was argued in the previous section—'replication' is itself an umbrella term encompassing a whole spectrum of activities. Thirdly, even if, in a hypothetical scenario, a clearer definition of the crisis is provided, it is unlikely that it can be reduced to any one cause. In fact, a decade of crisis literature has not unveiled any single factor—or single set of factors—, and it is not within the current thesis' scope to try and succeed where more eminent writers have failed. To obtain a reasonable qualification of the alleged replication crisis, the current thesis will therefore omit discussion on the applicability of the term *crisis* altogether; semantic disputes can be interesting from a historiographical, sociopolitical and broadly philosophical perspective (see, e.g., Goertzen, 2008; Morawski, 2019, 2020; Reinero et al., 2020; Romero, 2019; Steinle, 2016), but are generally unproductive in light of what the current thesis aims to achieve. Furthermore, since the replication crisis is still fully ongoing, we lack the clarity of hindsight. This unfortunately disables us from taking on a holistic third person perspective of the replication crisis. To move forward in a constructive manner, the problem is pragmatically dealt with by focussing on two specific aspects of the crisis: what exactly is not being replicated to the extent that it deserves the connotation of a crisis, and what has subsequently become the main focus of the reform movement in terms of concrete proposals to ameliorate said crisis? In the coming paragraphs, it will be argued that the replication crisis is, in fact, *not* one which encompasses the entirety of the replication spectrum, but only a small fraction of it, focussing mostly on close rather than conceptual replications. Based on this constraint, it will be argued that the reform movement has subsequently mainly centred on ridding the field of what is now widely known as *questionable research practices*. As such, the current replication crisis and the reform movement can be qualified as being mainly concentrated on the epistemic functions of repeatability as described in the previous section.

The idea that psychology is amid a replication crisis is exemplified by a great number of publications on the topic in the last decade (e.g., Asendorpf et al., 2013; Colling & Szűcs, 2021; Morawski, 2019, 2020; Open Science Collaboration, 2015; Pashler & Harris, 2012; Shrout & Rodgers, 2018; Stroebe & Strack, 2014; Świątkowski & Dompnier, 2017; Wiggins & Christopherson, 2019). Several journals have dedicated entire special issues to the topic, and several largescale advisory reports have been published (Peels, 2019). There have also been discussions on whether and how psychology students should be taught on the subject of the replication crisis (Chopik et al., 2018), and it has even reached mainstream media (e.g., Frith & Frith, 2014; Resnick, 2016; Tucker, 2016; Yong, 2013, 2016). However, the claim of there being a replication crisis has been denounced as well (e.g., Flis, 2019; Stroebe & Strack, 2014). Finally, some have used the widespread declaration of crisis to ostensibly argue that the notion of replication is overrated and that its value should be reappraised altogether (e.g., Feest, 2019). But where did it all start?

Opinions differ, but there seems to be a general consensus that 2011 was a decisive year that ‘kicked off’ the replication crisis. To briefly recapitulate from the introduction: Bem (2011) had published a paper in JPSP, in which he had supposedly shown, using a series of experiments, that the existence of extra-sensory perception—a kind of precognition or ability to ‘feel the future’—could be empirically corroborated. However, it was not necessarily the absurdity of the claims which was troubling; in fact, the ‘discipline’ of parapsychology had been around since the 1930s (e.g., Rhine, 1934). The real problems were that 1) the proposed ‘evidence’ was obtained using scientific practice that was standard for psychological research (LeBel & Peters, 2011), and 2) the original outlet, JPSP, refused to publish both successful and unsuccessful replication studies of the original, stating that the outlet was not supposed to become the ‘Journal of Bem Replication’ (Eddy, 2011). That same year, the massive fraud committed by Diederik Stapel, a prominent social psychologist, became known to the world (Callaway, 2011), and resulted in 58 retractions by the end of 2015 (Palus, 2015). Furthermore, these two major events were preluded by two preceding decades characterized by a steady flow of publications pointing out systematic problems in the output of the social sciences (e.g., Cohen, 1990, 1994; Coyne, 2009; Fanelli, 2010; Gigerenzer, 1998; Kerr, 1998; Meehl, 1990; see also Ioannidis, 2005), including some reviews pointing at highly irregular experimental outcomes (e.g., Vul et al., 2009) and—with the rise of meta-analyses—a prevalent publication bias in specific psychological subdisciplines (e.g., Ferguson, 2007; McDaniel et al., 2006; see also Hubbard & Armstrong, 1997).

It may be argued that these circumstances created a ‘perfect storm’, causing researchers to question the validity of some longstanding high-profile papers and research programs. At the centre of it all was social/goal priming research. Social/goal priming is an alleged phenomenon where people who incidentally or briefly encounter specific stimuli of a broadly social or goal-related nature (e.g., a word, a picture, *et cetera*), will be influenced by this mere exposure, in that they will tend to behave in line with the conceptual meaning that is associated with the encountered stimulus, relative to a control (Pashler et al., 2012). For example, reading words related to old age is said to instigate participants in an experiment to walk slower afterwards compared to controls who have not been exposed in this manner (Bargh et al., 1996). After what happened in 2011, several research groups attempted to replicate famous findings of this kind, but were generally unsuccessful (e.g., Doyen et al., 2012; Nieuwenstein & van Rijn, 2012; Pashler et al., 2012). Common denominators of these studies are that they often used larger samples than the original—for example, Doyen et al. (2012) doubled the sample size from 60 to 120 participants—to increase statistical power, they virtually always constituted a close replication—i.e., they made explicit efforts to deviate as little as possible from original designs, sometimes by contacting the original authors for specific information on proper procedure—, and when replications failed, subsequent optimizations in the research design in an attempt to gather more nuanced corroboratory evidence in favour of the original outcomes almost always yielded findings that were either improbable under the presumption that the original was actually true, or worse, entirely opposed to the substantiating theory behind the original findings. A telling example is Nieuwenstein and van Rijn (2012), who designed their replication based on what a meta-analysis had shown were the ideal experimental circumstances to elicit so-called 'unconscious thought effects' (Strick et al., 2011), yet were still unable to replicate the original even in these supposedly optimized circumstances. The irony of the publication year of that particular meta-analysis notwithstanding, Nieuwenstein and van Rijn's (2012) paper illustrated that not only original studies, but also meta-analyses had questionable contents; a conclusion which is especially problematic in light of the fact that meta-analytical procedures were generally thought of as yielding trustworthy aggregate summaries of particular research programs—or, at least, that is a purpose for which they were designed. Nieuwenstein and van Rijn's (2012) paper thus unveiled that social priming research is likely a victim of publication bias, permeating to and magnified by meta-analytical summaries. Apart from social priming research, there were also other domains which were starting to show cracks in their knowledge base by virtue—or rather, by vice—of failed replications. Most prominently, Bem's (2011) contentious paper was unsuccessfully replicated several times (e.g., Galak et al., 2012; Ritchie et al., 2012)—and, in fact, is still being unsuccessfully replicated to this day (e.g., a largescale multi-lab effort by Kekecs et al., 2023). The severity of the issue started to become even more pronounced when coordinated efforts to replicate several studies at once were generally unable to do so convincingly. Three telling examples are Nosek & Lakens (2014), Eerland et al. (2016) and Camerer et al. (2018). The former organized fourteen replication studies of different published effects, of which nine *did not* replicate and five were interpreted to be *'partial'* replications (i.e., replications which find evidence that is only mildly corroboratory of an original finding, or which replicates only part of it; Maxwell et al., 2015). In a similar vein, Eerland et al. (2016) published findings from a multi-lab registered replication report, this time of one particular effect (see experiment 3, Hart & Albarracín, 2011). Eleven labs participated, and most found "effects […] that were close to zero" and several "found effects that were numerically in the opposite direction of the effect reported in the original study" (Eerland et al., 2016, p. 166). As with the individual replication studies presented above, these coordinated efforts again were committed to staying as close as possible to the original design, and sought explicit criticism from the original author(s) in designing the replication studies' protocols. Camerer et al. (2018) focussed specifically on a set of behavioural social science reports published in *Science* and *Nature* between 2010 and 2015. Again, original authors were consulted and statistical power was analysed *a priori*, so as to be able to detect small effects. The replication rate was around 60 %, which is slightly more promising compared to its forebearers. However, arguably the most influential multi-lab coordinated attempt at replicating psychological science was provided by Aarts et al. (Open Science Collaboration [OSC], 2015). In this huge undertaking, one hundred investigations were systematically replicated across several years by 270 contributing authors. Notably, again, "[t]hrough consultation with original authors, obtaining original materials, and internal review, replications maintained high fidelity to the original designs" (OSC, 2015, p. 944). Except, instead of merely attempting to replicate the original findings, an attempt was made to estimate *the reproducibility of the entire field of psychological science*. This ended up being a meagre 40 %, although one must, in jotting down this arguably depressing statistic, immediately add that OSC (2015) entertained a very particular definition of replicability, namely, one in terms of statistical significance. It has been argued that this choice is problematic due to publication bias in favour of statistically significant results, which implies that these findings are at the tails of a distribution of potential statistics for studying a particular effect, causing it to be highly probable that close replications are not statistically significant due to statistical regression to the mean; as such, a replication crisis becomes almost a "mathematical inevitability" (Trafimow, 2018, p. 1190).[3] However, OSC (2015) also made comparisons between original and replicated effect sizes to complement the reproducibility assessment based on statistical significance tests. Specifically, they assessed whether replicated effect sizes were part of the originally reported 95 % confidence intervals (CI), but this was the case for only 30 out of 73

[3] Note that OSC (2015) was likely aware of this issue, as the conclusion of the report explicitly explains that failures to replicate are a necessary consequence of scientific practice. It reads: "[…] how many of the effects have we established are true? Zero. And how many of the effects have we established are false? Zero. Is this a limitation of the project design? No. It is the reality of doing science, even if it is not appreciated in daily practice" (p. 7). Singular studies never provide conclusive evidence, and neither do singular replications.

studies, which was itself only a subsection that contained the necessary information to allow for calculating a CI in the first place.[4]

The above examples unveil an important facet of the current replication crisis: influential replication efforts were largely, if not exclusively, carried out in the format of *close replication studies*; i.e., care was taken to make sure that the methodologies employed in replication studies were as close as possible to the original study, likely in an attempt to ensure commensurability between the original and the replication. To wit, the main takeaways from OSC (2015) are that 1) "[t]he claim that "we already know this" belies the uncertainty of scientific evidence", 2) "[r]eplication can increase certainty when findings are reproduced and promote innovation when they are not", and 3) "there is room to improve *reproducibility* in psychology" (p. 7; emphasis added). The third conclusion has been repeatedly reaffirmed, for example, by follow-up Many Labs studies, which across three iterations have produced a replication rate of only 57 % (29 out of 51; see Ebersole et al., 2016; Klein et al., 2018). Recently, Boyce et al. (2023) uploaded a preprint to the PsyArXiv repository, in which replication efforts conducted by students across an eleven year timespan (2011 – 2022) were systematically assessed. Of the 176 analysed original-replication pairs, only 49 % were subjectively interpreted as having yielded a successful replication. Of the 146 originals which had provided enough information to calculate prediction intervals, 46 % of replications produced a point estimate that lies within said prediction interval. The authors' conclusion is hard, but fair: "Our results indicate that [...] the robustness of the psychology literature is low enough to limit cumulative progress by student investigators" (Boyce et al., 2023, abstract).

A critical reader may observe that this conclusion (that replication efforts were mostly of a *close* nature) is based on a preselection of salient replication efforts, and may therefore be biased. Although it is true that the above examples are pooled from a larger literature and may therefore, theoretically speaking, represent it in a biased fashion, several arguments speak against this: firstly, as argued earlier in this section, the replication crisis was preceded by several years/decades of increasing scepticism about the output of the social sciences, especially in terms of publication bias. For example, Rosenthal's (1979) notion of the file drawer problem (i.e., the tendency of researchers to leave null results unpublished) was well known by then, and had become the subject of many discussions throughout the 2000s (see earlier in this section). To such an extent, even, that investigators were designing tools which could be used to identify publication bias as indicated by a pervasive dearth of null results in the literature (e.g., Stanley,

[4] During the writing process of this manuscript, a Special Issue was published in *Journal of Memory and Language*, a major outlet of its niche. Four out of seven attempts to replicate influential findings in that field were not successfully replicated (see Rastle et al., 2023). Notably, samples were, on average, quadrupled in size relative to original reports, but methodologies were allowed to deviate from the original, causing the replications to be slightly less *close*, though by no means outrightly *conceptual* in nature.

2005). Hubbard and Ryan (2000) note on this issue that widespread publication bias is at least in part due to frequent misunderstandings surrounding the evidential value gained from NHST and *p*-values:

> "Those unfamiliar with the deficiencies of [NHST] falsely equate it with methodological rigor and routinely incorporate it in their accounts. They persist in investing such tests with capabilities they simply do not possess. […] rejection of the null hypothesis is erroneously believed to yield the probability of the null […] being true as well as the probability that a research outcome will replicate" (p. 672).

In fact, NHST, being as pervasive a practice as it is, "has attained the status of a methodological imprimatur" (Hubbard & Ryan, 2000, p. 673; see also Bakan, 1966; Cohen, 1990, 1994; Gigerenzer, 1998; Meehl, 1990). Fidler (2005) enumerates at least eight such ways in which *p*-values and NHST have been frequently misinterpreted in the literature at large. These misinterpretations were also likely implicitly enforced by outlets requiring studies contain significant findings, since null results are frequently deemed uninteresting. For example, Angell (1989) summarizes that research studies remain unpublished mainly because researchers simply fail to write them up, do not submit them for publication, and because there exists the (tacit) assumption that editors are not interested in negative or null results; though see Kupfersmid (1988) for a summary of not so tacit, rather disgruntled disaffections in extant literature at the time, showing the historical pervasiveness of the issue. Hubbard and Ryan's (2000) subsequent recommendation was to implore researchers to conduct replication studies in the format of 'replicate and extend', i.e., what Haig (2021) refers to as *internal replication*, and which is, by definition, a version of *close replication*. Against this background of worries surrounding NHST misinterpretations and publication bias against null results, it makes sense that the reform movement would focus heavily on filling this void, which necessarily involves redoing experiments as they were done originally, to find out in which file drawers the null results are 'hiding'. On this issue, OSC (2015) emphasizes: "Humans desire certainty, and science infrequently provides it. […] a single study almost never provides definitive resolution for or against an effect and its explanation. […] In some cases, the replications increase confidence in the reliability of original results; in other cases, [they] suggest that more investigation is needed" (p. 7).

Secondly, Daniel Kahneman, a researcher of behavioural economics, wrote an open email in which he plead for an urgent restoration of the field's credibility—specifically, of social priming research—by systematically replicating colleagues' findings. The impact of this open email should not be underestimated, for Kahneman was and still is a respected individual of the scientific community, or, as B. Nosek put it, "a hard man to ignore" (as cited in Yong, 2012). Kahneman recommended social psychologists commit themselves to setting up a "daisy chain" of replicatory efforts, where labs systematically perform each other's investigations anew. Importantly, to do so successfully, it was recommended that “parties would record every detail of the methods, commit beforehand to publish the results, and make all data openly available”, because “priming effects are subtle, and could be undermined by small experimental changes” (Yong, 2012, see ‘Chain of replication’). Again, emphasis is explicitly put on replicability as repeatability, i.e., *close replication*, by staying as close as possible to the original design under the pretence that the effects being studied are so fragile that minor changes in protocol could undermine their being reproduced.

Thirdly, and likely the strongest counterargument to potential scepticism surrounding the reality of the reform movement’s focus on close replications, is the fact that psychological theories are 1) almost always underspecified and 2) consist often merely of a verbal structure which allows internal logical inconsistencies to remain unnoticed. It follows that replications of a strongly conceptual nature are simply impossible to carry out informatively. First of all, conceptual replications are, in and of themselves, difficult to interpret, especially when the results are negative. Maxwell et al. (2015) formulate it as follows: “[…] if the replication study fails to find an effect previously reported in a published study, any discrepancy in results may simply be due to procedural differences in the two studies. For this reason, there has been increased emphasis […] on exact […] replications” (p. 488). In a Consensus Study Report published by National Academies of Sciences, Engineering, and Medicine (2019), it is summarized that “[a] failure to replicate previous results can be due to any number of factors, including the discovery of an unknown effect, inherent variability in the system, inability to control complex variables, substandard research practices, and, quite simply, chance” (p. 72). Therefore, it makes sense to want to stay as close as possible to an original, for discrepancies in outcome are virtually unattributable when theories are underspecified and potentially logically inconsistent. More problematically, however, is the fact that most psychological fields of study lack *formalized* theories. Remember that for a replication to be conceptual *and informative*, the goal must be 1) to challenge an explicitly stated auxiliary hypothesis of the central theory, 2) to challenge an aspect of its formalization via a derivation chain from the logically abstract to an empirical postulate, or 3) to corroborate or weaken a theory by establishing an empirical fact which is itself rooted in an entirely different but well-established theoretical approach consisting of a distinct composite description of theoretical assumptions, auxiliary hypotheses and logical postulates leading to testable empirical predictions (see earlier discussion on replication typologies in ‘The many faces of replication’; also Meehl, 1990; Radder, 1992). Numerous authors have noted the absence of any such formalization in psychological science for many years, a characteristic which prohibits systematic cumulation of knowledge beyond mere statistical effects (van Rooij & Baggio, 2021). Furthermore, complete descriptions of experiments and assumptions are frequently missing, a factor which undoubtedly contributes to the non-replicability of psychological science (Hensel, 2020). Given these two factors (informality and descriptive incompleteness), what more can one reasonably utter upon repeatedly completing an unsuccessful attempt at replication than a sighed "oh well…"? Meehl (1978) famously decried that it "is simply a sad fact that in soft psychology theories rise and decline, come and go, more as a function of baffled boredom than anything else; and the enterprise shows disturbing absence of that *cumulative* character that is so impressive in disciplines like astronomy, molecular biology, and genetics" (p. 807; emphasis original). A speaking example is the ego-depletion theory (Baumeister et al., 1998), which, after having been heavily criticized during the early years of the replication crisis, seems to have simply been lost to memory in the absence of definitive evidence pro or contra (Friese et al., 2018). The truth is that Meehl's (1978) quote might as well have been written up in modern times. Eronen and Bringmann (2021) note that, apart from being generally underspecified and informal, even those theories which allow some form of—arguably weak—testing/falsification and have subsequently been found deficient are often retained because, to paraphrase, they 'used to do well' (see also Meehl, 1990)—an example being the Rescorla-Wagner model of classical conditioning (Miller et al., 1995). This crucial deficit in psychological science has been addressed many times, and is starting to take attention away from mere replication to a seemingly more fundamental problem of 'theory crisis' (e.g., Devezer & Buzbas, 2022; Devezer et al., 2019; Oberauer & Lewandowsky, 2019; van Rooij & Baggio, 2021). This positive evolution notwithstanding, the fact of the matter remains that most current psychological theories are underspecified and under-formalized, which basically renders most, if not all, conceptual replication studies virtually uninformative, because it is unclear what aspect of the tested theory a(n) (un)successful conceptual replication is supposed to address. In fact, it could be argued that in the absence of properly specified theory, conceptual replications cannot really exist, for there is no way of specifying the degree of closeness between a replication and an original—neither strictly formally, nor verbally in a logically consistent way. As such, close replications are the only option left, and the option which dominates current reform efforts.

Furthermore, in the absence of formalized theory, psychological science in practice mostly focusses on establishing statistical *effects*. Cummins (2000) explains that "a substantial proportion of research effort in experimental psychology isn't expended directly in the explanation business; it is expended in the business of discovering and confirming effects" (p. 6). This is not necessarily problematic *a priori*, because, as Cummins (2000) explains, the primary explananda of psychology are *capacities* (see also van Rooij & Baggio, 2021) which need not be discovered (because we are already aware of them; e.g., the capacity to see depth, to learn a language, to detect patterns, *et cetera*), and many of the uncovered effects are "incidental to the exercise of some capacity of interest" (p. 9). That is, uncovering effects, their law-likeness, and the circumstances in which they are elicitable, serves a purpose in a specific set of circumstances (for example, to adjudicate between two theories which both explain a capacity of interest, but of which the first explains incidental empirical effects and the second does not). However, capacities must be properly specified for empirical effects to have bearing on an epistemic matter, for if they are not, what is left is a series of effects which describe, but do not explain the data subsumed under them. For example, the McGurk effect is simply an effect, *not an explanation of the effect* (Cummins, 2000). In other terms, effects are *explananda*—i.e., things to be explained—, not *explanantia*—i.e., the things which serve as explanation. On this, van Rooij and Baggio (2021) critically assess that "methodological reform so far seems to follow the tradition of focussing on establishing statistical effects, and, arguably, the reform has even been entrenching this bias" (p. 683). To wit, lacking theory creates an environment where one "[tries] to write novels by collecting sentences from randomly generated letter strings" and hopes that the right sentences—i.e., informative effects—present themselves by chance (van Rooij & Baggio, 2021, p. 683). As such, the reform movement has mainly focussed on devising methods which ensure effects are replicable (e.g., data sharing, preregistration, registered reports to change publication incentives, *et cetera*; see Renkewitz & Heene, 2019), and has focussed explicitly on the quality of individual studies—a practice which was already criticized by Danziger (1985) as being a problematic tendency of psychological science practitioners throughout its modern existence when he wrote: "The tenacious hold which inductivist mythology acquired over research practice of psychologists led to the delusion that the question of methodological bias need be addressed only in the context of the individual research study. As long as each study was well designed their piling up […] would somehow result in a scientific discipline" (p. 2).

In summary, it seems fair to conclude that 'the replication crisis' is somewhat of a misnomer, given how both the discipline at large and the reform movement seem to focus mainly on replicability in the close sense—i.e., *repeatability* and *internal replicability*—while the problem is probably more fundamentally located at the level of theory construction (see Devezer et al., 2019; Eronen & Bringmann, 2021; Guest & Martin, 2021; Klein, 2014; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Smaldino, 2019; van Rooij & Baggio, 2021; van Rooij & Blokpoel, 2020). A final, but telling example of this focus on cumulating small effects is the existence of the 'short report'. Ledgerwood and Sherman (2012) show how the rise of the 'short report', in which a single study is presented without much elaboration or complex theoretical nuance, might have unintentionally inundated the field with false positives that are being chased around in what from outside seems like "fruitful interaction among researchers" (p. 62; the authors use the term "goose chase"). Furthermore, it floods the field with "apparently novel, disconnected findings" (p. 62), and the lack of elaboration inhibits proper historical contextualization of research. It also promotes newsworthiness over basic findings, a problem which has been coined as fundamental in the shaken field of social priming (Yong, 2012). Again, these problematic consequences associated with short reports mainly exacerbate the prevalence of false positives, and the reform movement seems consequently to have pinned itself on ridding the field of false positives as best as it can.

Thus, the replication crisis is best qualified as a *close replication* or *repeatability crisis* in practice, and a *theory crisis* in reality. The immediate result of this seems to be that the reform movement largely focusses on ameliorating methods of individual studies and changing publication standards, so as to guarantee reliability of outcomes and 'solve' publication bias on the whole. A large portion of these amelioratory efforts share a specific common goal: to change practices often engaged in by researchers—not necessarily in a deliberate fashion—that cause an excess of false positives in the extant literature, hence causing its low replicability rate. To the reform movement's credit, their efforts are, in fact, slowly taking hold in the unwieldy institution that is scientific research. It must be emphasized that this is a good thing. Unfortunately, however, one incredibly important and technical dyadic factor seems to have been largely forgotten, namely, statistical power analyses and sample size determination/justification. Dyadic, because power analysis and sample size determination/justification are intimately linked, but distinct features of research; and forgotten, or rather, ignored, as power analyses have never really taken any strong footing in psychological science in practice, despite its centrality to those statistical philosophies which are mostly employed by psychological researchers. It is only in recent years that statistical power has become a relatively central topic of discussion in the reform movement.

However, before all things 'power analysis' can be discussed, a brief interlude is necessary to familiarize the reader with a particular branch of literature and its contents, which has arisen as a direct consequence of the reform movement's preoccupation on reducing false positives and increasing close replicability of single studies. Practices which facilitate the generation of false positives are oftentimes grouped under the acronym 'QRPs', or *questionable research practices*. In the interluding section, a brief discussion is provided on what QRPs exactly are and what their alleged role is in the sustainment of the replication crisis. If this is done, the stage will be fully set to start tackling the main research topic of the current thesis.

Interlude – Questionable Research Practices

In the previous segment, it was extensively argued that the current replication crisis is one which may best be characterized as being concerned with close replications. It was shown how the focus of the reform movement has therefore largely directed at issues pertaining to repeatability and reliability of any single study's outcomes. This has given rise to a vast literature on QRPs, or practices which negatively affect these desiderata.

One can distinguish between two crudely defined categories of behaviour that seem to generate—at least in part—the observed lack of successful replications (Scheel, Schijen, & Lakens, 2021). The first category is, broadly speaking, of a cultural nature. It consists of two closely related elements: a widespread and historically grown pressure to publish (De Rond & Miller, 2005; McNemar, 1960), and a pervasive publication bias in favour of novel and 'exciting' data, accompanied by a near complete disregard for negative and null findings (Francis, 2012; Greenwald, 1975; McNemar, 1960). Both entail obvious detriments to the advancement of scientific knowledge. Firstly, a pressure to publish—perhaps best known under its much loathed dictum 'publish or perish' (Case, 1928)—produces an academic climate that encourages quantity over quality by making an academic's job security or tenure, but also their funding and social prestige a function of their publishing rate (De Rond & Miller, 2005). Angell (1986) describes how this climate creates a "fragmented and repetitive literature", "[a]n almost irresistible incentive to cut corners", and "an erosion of the integrity of the scientific enterprise" (p. 261). Essentially, the disconnect between the scientific standard and existing economic incentives invites academics to perform substandard research. In recent survey research respondents have freely admitted that one of the commonest causes for self-assessed research misconduct is this mythical pressure to publish (Pupovac et al., 2017). Combine this with a publishing culture which actively suppresses the dissemination of null findings (as is often shown in meta-analyses and sometimes even in overt statements by reviewers and publishers; see Hubbard & Armstrong, 1997; Kepes, Banks, and Oh, 2014; Thornton & Lee, 2000), and you have a recipe for disaster: exciting findings get published, null findings never see the light of day from the file drawers where they are kept and forgotten, and the published literature receives little of its due scrutiny, because failed replications end up in that same dismal place where, historically, most null findings go to die. An absence of null findings in the literature threatens the validity of meta-analytical reviews (e.g., Kepes, Banks, McDaniel, and Whetzel, 2012), by producing—*a priori*—a skewedness in the published data, in favour of a proposed effect. This, in turn, leads researchers to investigate hypotheses and theories which are based on misleading literature—all inevitably leading to naught.[5] Consequently, numerous calls for revisions of this biased, quantity-driven publication culture have been and are still being made (Wiggins & Christopherson, 2019).

A second, more insidious category of problematic research practices constitutes a spectrum of behaviour going from persistent methodological lack of caution to plain laxity, which at

[5] Recent publications on nudging and choice architecture (Thaler & Sunstein, 2021) are exemplary in this regard. It has been frequently reported that meta-analytically derived effect sizes are often implausibly sizeable (Cohen's d > .4) due to publication bias. If the publication bias is accounted for, evidence in favour of the existence of nudging effects seem to dissipate almost completely (Maier et al., 2022). This does not imply that nudging effects *per se* do not exist on any level, but it does show that the published literature presents a severely skewed picture. This not only hinders researchers' understanding of the theoretical constructs and mechanisms under investigation, but also leads to misinformed policymaking.

times verges on outright malpractice (Hartgerink & Wicherts, 2016). Such practices are commonly referred to as QRPs, or “research behaviour that makes evidence in favour of a certain conclusion look stronger than it is” (Scheel, Schijen, & Lakens, 2021, p. 2). It is important to distinguish QRPs from sheer fraudulent practices like data falsification and plagiarism. Though it may be rather trivial that falsifying data cannot be justified under any circumstances, and that plagiarism is—quite rightly so—simply illegal, QRPs are more subtly deceitful. Suter (2020) defines QRPs in an excusably derogatory manner by paraphrasing Harris (2017): “[QRPs include] sloppy science, dodgy methods, cutting corners, taking shortcuts, sketchy procedures, and the like” (p. 1). However, pejoratives aside, this is exactly right: QRPs are ‘questionable’ rather than categorically unethical, and they are more easily and self-deceptively justified by the researcher who employs them (John et al., 2012). Cutting corners and taking shortcuts can always somehow be justified by ad hoc excuses, i.e., brushed off as inconsequential in the grand scheme of things, but that is exactly the point: they are not and should not be viewed as inconsequential little missteps. They include *p*-hacking (*p* refers here to the omnipresent *p*-value as an indicator of statistical significance in NHST; e.g., Brodeur et al., 2022; Friese & Frankenbach, 2020; Head et al., 2015; Neher, 1967), *HARKing* (hypothesizing after results are known; Kerr, 1998; though see Rubin, 2022, for a series of objections), *selective reporting* of studies that ‘worked’ (e.g., John et al., 2012; Neher, 1967), *disregard for statistical power* complemented by an overly persistent focus on statistical significance (e.g., Cohen, 1962; see also Rubin, 2023, for a recent discussion on questionable metascience practices, i.e., practices which are problematic in researching questionable practices) and many others. Of course, this list is not exhaustive by any means, and extensive overviews may be found elsewhere (see Artino et al., 2019; Banks et al., 2016; Suter, 2020). QRPs are dangerous because they tend to introduce large amounts of false positives to the extant literature (though not always, see Lakens, 2019), thus creating a body of literature that is largely irreproducible (in the sense of close replications; Linder & Farahbakhsh, 2020).

To illustrate how exactly QRPs lead to inflated false positive rates, consider the following example: *p*-hacking is an umbrella term for broadly “any measure that a researcher applies to render a previously non-significant *p*-value significant” (Stefan & Schönbrodt, 2023, p. 2). Subsumed under it are activities like *selective reporting* of statistical tests with dependent variables which have yielded significant *p*-values (i.e., $p < 0.05$), *optional stopping* or ‘data peeking’ where significance tests are iteratively conducted on growing data samples until a statistically significant result pops up and is subsequently reported on its own, and *selective data trimming* of results such that statistically significant results are obtained (see Stefan & Schönbrodt, 2023, for a compendium of *p*-hacking strategies and simulations visualizing their effects). *P*-hacking leads to increased false positive rates in published literature, because null or negative findings are systematically obfuscated or tortured until they yield ‘publishable’ results. And the reason this is done, is because—as explained previously—academics operate in a climate which actively disfavours negative or null findings. Another example, HARKing, is closely related to *p*-hacking, but distinctly different in practice. *Hypothesizing after results are known* basically entails that researchers alter their hypotheses after having performed statistical analyses, such that results are always confirmatory of hypotheses which, once in print, have all the allure of having been proposed *a priori*. Whereas *p*-hacking distorts statistics to fit the hypothesis, HARKing distorts the hypothesis to fit the statistics. Either way, the result is the same: the literature is saturated with confirmatory results, so false positive rates are necessarily elevated.

For the purpose of the current thesis, the comprehensive, though perhaps too vaguely defined second category encompassing all QRPs is further divided. In wanting to remedy the prevalence of QRP engagement,[6] it might be useful to distinguish QRPs based on *when* they are engaged in during the whole research cycle (starting from theory construction and hypothesis formation to writing down a report). That is, some QRPs do not result in an alteration of the data itself, but merely allow to alter what is shown and what is not, *a posteriori*. Which output one can decide upon to show or not show to the reader is limited to an *a priori* range, which itself is determined by those experimental design variables that dictate the nature of the data to be gathered, and, as such, how it may possibly be manipulated. What may or may not be selectively or distortedly reported depends on the experiment. Hence, it might be useful to make to instrumental distinction between *presentational* and *antecedental* QRPs. The former encompasses practices which researchers may engage in to present an idealized, distorted or selective version of their findings, *a posteriori*. The latter refers to those variable properties of a research design and inference procedure which may be tweaked prior to actual data gathering, in order to elicit favourable outcomes in terms of publication value. Examples of such variables include the choice of significance criterion (for NHST; e.g., one may choose to adhere to a standard 0.05 alpha cut-off level, or one may choose to be more restrictive than that; see, e.g., Benjamin et al., 2018; Cesana, 2018; Manderscheid, 1965), sample sizes (cf., *p*-hacking), choosing a between-subject or within-subject design (see, e.g., Charness et al., 2012), opting for fixed, random or mixed effects designs (see, e.g., Firebaugh et al., 2013; Hedges & Vevea, 1998), longitudinal, cohort studies or cross-sectional studies, *et cetera*.

The main purpose of dichotomizing QRPs into presentational and antecedental variants is to facilitate discussions and implementation of remediation efforts. There lies a strength in

[6] For estimates based on surveys, see Artino et al. (2019), Fanelli (2009), Fiedler & Schwarz (2016), Fraser et al. (2018), Gopalakrishna et al. (2022), John et al. (2012), Kaiser et al. (2022), Krishna & Peter (2018), Martinson et al. (2005), Moran et al. (2022), Pupovac et al. (2017), Rajah-Kanagasabai & Roberts (2015), and Tijdink et al. (2014). For estimates using non-survey techniques, see Bakker and Wicherts (2014), Banks et al. (2016), Hartgerink et al. (2016), and Nuijten et al. (2016). For further reading, see Francis (2014), Fox et al. (2018), Ioannidis and Trikalinos (2007), Motyl et al. (2017), and Renkewitz and Keiner (2019).

focussing on remedying antecedental rather than presentational QRPs, for they are chosen beforehand and their uses are logically justifiable (i.e., their impact on the probability of finding real effects may be assessed *a priori*). Theoretically speaking, the chosen specifications for these variables are confined only by practical limitations and one's creativity in designing solid research. If the foundations of an investigation are properly restricted, substantiated and communicated, there will remain far less room for subsequent attempts at presenting data through a distorted lens.

The main QRP that will be discussed later—namely, pervasive neglect for informed power analysis and sample size determination/justification—is a prime example of an antecedental component which, if handled correctly, can provide adequate restrictions on research *a priori*. For instance, Brodeur et al. (2022) recently found that not preregistration *per se*, but the inclusion of a pre-analysis plan is associated with reduced evidence for *p*-hacking. Furthermore, preregistrations that specifically involve a discussion on power analysis and sample size determination appear less prone to *p*-hacking as well. These findings may be preliminary, but they provide a first glance into the potential effectiveness of focussing remediation efforts on antecedental aspects of a study, rather than its presentational aspects.

To conclude this interlude, consider the following: Banks et al. (2016) have stated that "QRPs are occurring at rates that far surpass what should be considered acceptable" (p. 328; see also Stricker & Günther, 2017). One could argue that the fact that QRP engagement seems to occur at any level should already incentivize academics to revise their conduct; the order of magnitude at which QRPs are reportedly occurring warrants largescale changes, irrespective of the exact prevalence, for there is no acceptable rate of such engagement. These practices can damage the field's credibility (Anvari & Lakens, 2018; though see Mede et al., 2020), but most importantly, they have and are still inhibiting the advancement of our understanding of psychological phenomena. Specifically, QRPs allow for unacceptably high levels of researcher degrees of freedom, which, in turn, allow for virtually anything to be presented as statistically significant. This leads to a literature that is saturated with false positives (Simmons et al., 2011), to the extent that psychological sciences are currently held captive by a self-sustaining cycle of publication crisis, QRPs and an inability to perform successful close replications *en masse*.

Statistical Power in Psychological Science

The image of the replication crisis presented thus far reflects a field struggling to preserve its credibility. To recapitulate, the current thesis started out with an extensive outline of the epistemic functions associated with close and conceptual replications in science; whereas the former serves primarily to establish reliability of experimental procedures, apparatus and the like, and fulfils a distinctive role in reliabilist systems of belief justification, the latter serves to substantiate epistemic claims in terms of their validity and generalizability, following a broadly coherentist system of belief justification. A short history of the replication crisis was presented, and several arguments were put forward—including historical precedence, statistical practices, and lacking formalization in theory construction—to substantiate the conclusion that the replication crisis and its reform movement ought to be qualified as being centred around issues pertaining to close rather than conceptual replicability of experimental findings. This has inevitably dragged the focal point of curative efforts onto practices which may endanger the main epistemic function associated with close replications, namely, reliability and trustworthiness of singular reports. The result has been a proverbial witch hunt for so-called QRPs, or oft-occurring practices which cause the psychological literature to be saturated with untrustworthy *false positives*. Some examples of such practices were provided in the interlude, of which *p*-hacking is likely the most widely known among practicing researchers.

The reform movement's efforts to diminish the number of false positives finding their way into the literature and facilitate the publication of null and negative findings, are laudable. However, a substantial amount of these efforts has focused on presentational factors in the sense described previously, i.e., practices which affect how data is presented when it is being gathered or after the data acquisition process has been completed. Some examples include HARKing, selective reporting of statistically significant independent variables, optional stopping, *et cetera*. To its detriment, this focus has come at the apparent cost of tackling antecedental QRPs, which have received arguably less attention. Examples include pre-analysis sample size determination, design construction in its broadest sense, statistical hypothesis formulation (e.g., opting for fixed, random, or mixed effects), and the like. These decisions are usually made prior to gathering data. The reasons for why these elements have received far less attention are multitudinous, but one definite factor is how most of those choices are more or less connected to the concept of statistical power, or the capacity of a given experimental design to detect an effect of interest by virtue of a dedicated test yielding statistical significance. Neglect for statistical power constitutes an antecedental QRP because it is a variable which can be manipulated prior to experimentation, by performing an *a priori* power analysis, thus granting a researcher more control concerning the probability of finding a true effect.

It is only in recent years that the pervasive neglect for statistical power in psychological science has been creeping into the reform movement's limelight, despite of the fact that the systematic nature of the problem was already well communicated in the 1960s and 1970s (e.g., Cohen, 1962). For example, more and more publishers and funding bodies are starting to mandate the inclusion of power-analytical considerations as part of a statistical analysis plan (e.g., JPSP has recently [see Giner-Sorolla et al., 2023] started mandating authors explicitly address statistical power, as can be observed in their imposed reporting standards; see Appelbaum et al., 2018). Also, knowledgeable authors are publishing power primers at growing rates to aid their fellow researchers who are inexperienced in the matter, and awareness of the severity of the problem is generally increasing as well. This begs the question whether the amount of attention statistical power as a topic is receiving in ever larger quantities is mirrored by an equivalent increase in the number of publications that actually include an *a priori* power analysis.

In the following paragraphs, the concept of statistical power is explicated, followed by a historical overview of the pervasiveness of the statistical power deficit in psychological science. Doing so will naturally lead back to the aforementioned question: is the practice of conducting *a priori* power analyses actually increasing? A systematic literature review will serve to address this pertinent matter.

STATISTICAL POWER AND POWER ANALYSIS

In order to understand the impact of statistical power—or lack thereof—on singular reports, sets of studies, and close replications of original studies, one must first define it. The most basic definition of statistical power is provided by Neyman and Pearson (1933a), namely, "the probability of rejecting the null hypothesis tested, H_0, when the true hypothesis is H_i" (p. 498). That is, statistical power is a probabilistic concept from the frequentist tradition of statistical inference, where a statistical null hypothesis and all of its preliminaries are assumed and data are subjected to a test of statistical significance, the outcome of which is used to make a dichotomous decision concerning the statistical null hypothesis—i.e., to reject or not to reject it. This decision can be done in error, and statistical power provides a long-term probabilistic guarantee that erroneous nonrejections occur at a specified frequency that is based on said power.

More formally, Neyman and Pearson (1933a) consider w an instance with given size of the sample space W having n dimensions, in which it is permissible for any sample point $\sum$ (with coordinates x_1, x_2, … x_n) to lie. The authors dubbed w the *critical region*, stating that if $\sum$ is defined by a set of variates which falls in w, the reality of H_0 may be considered unlikely, thus permitting its rejection. The probability of doing so correctly—i.e., when H_0 is false—is denoted as $[1 - P(w|H_0)]$, and this is equal to what is nowadays known as $(1 - \alpha)$, also coined the 'correct inference probability' (see Strahan, 1982), with α the statistical significance cut-off. Thus, $P(w|H_0) = \alpha$ is the probability of committing a type I error, or "the chance of rejecting H_0 if it is true" (Neyman & Pearson, 1933a, p. 495). Statistical power, or the chance of correctly rejecting H_0 in favour of a specified alternative hypothesis H_i , can be denoted as $P(w|H_i)$, with the critical region w again an instance with given size of the sample space W having n dimensions; it reads: the probability that a given sample point $\sum$ defined by a set of variates (with coordinates x_1, x_2, … x_n) falls into w, is the statistical power of the critical region w regarding H_i. Nowadays, one tends to refer to the statistical power of a given test T, rather than with respect to a critical region pertaining to a given alternative hypothesis, for the term ‘critical region’ is colloquially used more often in reference to α and statistical null hypotheses. The statistical power of a given test T is closely related to the type II error one can commit following an inferential decision based on T, in that they are complementary aspects of a whole. That is, “type II [errors are] made by accepting H_1 [the null hypothesis] when it is false” (Wald, 1939, p. 300).[7] Put differently, a type II error is committed whenever an effect exists, but is not found using T, such that one is inclined to deny its existence rather than accept it. Let type II errors for a given statistical test T occur at a level β, then the statistical power of T is denoted as $(1 - \beta) = P(w|H_i)$.

As stated, statistical power is a concept that emerged explicitly from the frequentist philosophy on statistical inference. In modern times, practicing researchers likely know it best by its oft-employed acronym ‘NHST’, or *null hypothesis significance testing*. NHST in practice entails that a researcher must define a statistical null hypothesis, gather data, and perform an appropriate statistical test (e.g., a classic paired-sample t test) to calculate a p-value. The p-value describes the probability of observing the pattern exhibited by the gathered data if the statistical null hypothesis were true (e.g., that a population mean μ has a specific value), including all of its assumed preliminaries (e.g., the dependent variable follows a normal distribution, or approximates it sufficiently so as not to cause systematic error). If the p-value is smaller than a preset cut-off value α, it is decided that the pattern exhibited by the gathered data is sufficiently improbable under the statistical null hypothesis, such that one feels justified in rejecting it—knowing that one will do so in error approximately α % of the time—until novel data indicates otherwise. Statistical power enters this picture, because erroneous rejection of the statistical null hypothesis is but one (type I) of two possible errors following a dichotomous decision based on a p-value and α. That is, one may also *fail to reject* the statistical null hypothesis (type II). Just as the type I error probability can be fixed by adhering to a dichotomous decision rule based on α, the type II error probability can be fixed by making sure that the relevant variables which may influence β—i.e., the chance of false nonrejection of H_0—take on values that satisfy an erroneous nonrejection probability which is deemed acceptable *a priori*.

In essence, both the statistical significance level α and the statistical power $(1 - \beta)$ must be fixed for a given test T prior to analysis, as a direct reflection of what minimum probabilities of type I and II errors occurring one deems permissible (Wald, 1939). To do so, one may conduct an *a priori* power analysis, using two connected parameters which also pertain to T: sample size and effect size (ES). These four elements are related in a way that enables researchers to determine for any given test T any single value of these parameters if the other three are given (Cohen, 1988, 1990; Cooper & Findley, 1982). In practice, an *a priori* power analysis basically entails

[7] Neyman and Pearson (1933a) and Wald (1939) denote the null hypothesis using different subscripts. H_0 in the former equals H_1 in the latter.

calculating the number of participants needed to satisfy the given values of α, $(1 - \beta)$, and ES. That is, it enables researchers to determine the sample size required, such that the dichotomous decision to be made upon calculating a p-value conforms to the implied long-term frequentist commitment of fixing the probability of committing a type II error at β for a given statistical significance level α and a minimal ES of interest. A close neighbour of statistical power analysis is a so-called *sensitivity analysis*. Oftentimes, when practicalities limit the magnitude of a sample size *a priori*, a researcher might choose to fix sample size and calculate a power curve for a gradient of ESs at a specified significance level. Doing so will indicate from which point onward a detected ES that is deemed statistically significant, is more likely lead to an erroneous decision to reject the statistical null hypothesis.

Essentially, that is what NHST is about: it is a long-term gambit for which one assumes a set of preliminaries (i.e., a set of specific statistical assumptions with respect to a population parameter of interest and for a given statistical test), and produces dichotomous decisions (i.e., to reject or not to reject the statistical null hypothesis) in a way which guarantees that the choices made are done so in error at preset frequencies (i.e., α % erroneous rejections, and β % erroneous nonrejections of H_0). These frequencies are traditionally set at α = 0.05 and β = 0.2 (because the risk of committing a type I error is generally accepted to be more problematic than the type II variant, although the values themselves are essentially arbitrary). However, dichotomous decisions pertaining to statistical null hypotheses are always done in a provisional manner, for no single test's outcome may on its own function as a means for refutation in the Popperian sense (Hacking, 1965), and because such rejection always occurs given the set of statistical preliminaries (e.g., homogeneity of variance for standard t tests [see Boneau, 1960]). Furthermore, an attentive reader may have noticed that H_0 is a *statistical* null hypothesis, and not a *theoretical* one. This difference is paramount. To wit, the frequency of correctly rejected statistical null hypotheses may perfectly satisfy the power function (as a function of α, sample size and ES), but one can never use this state of affairs to bypass the still-existing gap between the statistical hypothesis test (which essentially entails nothing more than a set of assumptions regarding the frequency of empirical random variables) and the epistemic claims concerning some theory-driven empirical postulate, the adjudication of which is based on said statistical test (Neyman, 1950, see pp. 289 – 290). It is up to the scientist to determine if the statistical inference procedure is valid. Moreover, in conducting such an analysis (power or sensitivity), one must always keep in mind that the results are probabilistic. If a researcher finds a statistically significant ES of which a prior analysis indicates that the current experimental design is not statistically powerful enough to guarantee long-term adherence to the preset type I and II error probabilities, said researcher should take into the account the very real possibility that the ES uncovered is a false positive. The reason for this is that with low statistical power, the rate of true positives However, if an ES which is found statistically significant falls within the bounds provided by a power or sensitivity analysis, there still is no guarantee that the frequentist decision that follows is not produced in error. Thus, the proper interpretation of any single statistically significant outcome must acknowledge the statistical preliminaries of the test, and may only serve as a tentative corroboration of a theoretical hypothesis; corroboration feeds into justified belief, but belief remains provisional on the possibility that new data may point in another direction (see Lakens, 2021; Levi, 1967).

The concrete effects of low power are best understood in reference to a body of publications and not any one statistical test. Because studies are generally only published when they yield a significant *p*-value, two problems arise when statistical power is chronically low: the false discovery rate (FDR) increases, and ES estimates tend to be inflated. Bartoš and Maier (2022) provide an overview of how statistical power affects FDR. The FDR reflects the proportion of false positives in a pool of significant findings (i.e., a pool comprising both false and true positives, which have been deemed ‘positive’ by virtue of a statistical significance test). That is, the proportion is defined as $FDR = \frac{number\ of\ false\ positives}{(number\ of\ false\ positives) + (number\ of\ true\ positives)}$. The number of true positives depends entirely on the statistical power of a test and the number of true alternative hypotheses (Bartoš and Maier, 2022), so if statistical power decreases, the proportion of false positives to the whole of findings deemed positive increases. That is to say, due to low statistical power, the number of false positives in the literature will increase *relative to* the number of true positives, when only statistically significant results are published.

Secondly, ES tends to inflate in published literature. To understand why, consider the following simulation (scenario and procedure by Frost, 2019). Imagine a situation where a researcher is interested in the difference in IQ scores between two independent groups (without assuming the direction of the potential difference). This researcher decides to determine the IQ scores of 10 randomly sampled individuals from each group, totalling twenty subjects. The difference between the sampled means is 15 IQ points, so the researcher decides to test the difference between these groups using a simple two-sided *t* test for independent means. They find a statistically significant *p*-value and decide to publish their results. Others see this publication and decide to replicate the research with a similar setup. Mostly statistically significant findings are published (given the file drawer phenomenon, Rosenthal, 1979). What, then, are the ramifications on the ES in the long term? Let us assume an all-knowing third person perspective. Imagine the IQ scores of both groups are distributed following a normal distribution with equal standard deviations (σ = 15), but different means (μ_1 = 100, μ_2 = 110). One can calculate the statistical power a two-sided *t* test of independent means would have in the prescribed scenario. It turns out this is $(1-\beta)=0.29$ (assuming $\alpha = 0.05$).[8] This means that we expect to detect the true effect of interest 29 % of the time. By repeatedly sampling 10 values from each population distribution and performing a two-sided *t* test, one can simulate how many tests will return statistically significant in the long run, and what the effect size estimates that are published will be.

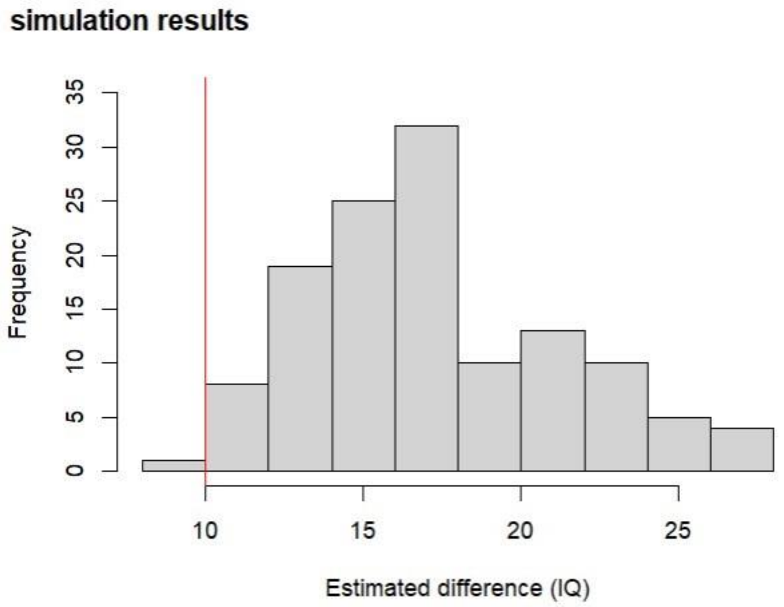
For the simulation, the process of sampling and testing as described above was repeated 500 times. 25.4 % of simulations yielded statistical significance at the nominal 0.05 cut-off, which is relatively close to the expected 29 % (if the simulation experiment were to include infinite iterations, the positive detection rate would equal 29 %). Interestingly, of those simulations which yielded statistical significance, the rounded difference between the two groups (i.e., the ES) is, on average, 17.23, and the total range is [9.95, 27.60] (see **Figure 1**). 126 out of 127 statistically significant tests yielded an overestimate of the ES (i.e., > 10 IQ points). So, low statistical power not only increases the FDR, but also inflates concurrent ES estimates (the actual difference in IQ scores is 10). If we assume that the file drawer phenomenon is not absolute, so that some statistically nonsignificant findings are published as well (say, *p*-values < 0.1), the mean ES in the simulated literature drops slightly (15.70), and the range slightly widens [7.84, 27.69]. Still, even though a slightly larger portion of the estimates resides around the true difference of 10 IQ points, it is evidently clear that the general magnitude of the estimates remains problematic. It is obvious, then, why high-powered close replication efforts are largely unsuccessful: the increased FDR implies that the probability of successful close replication in a highly powered design is lower, *a priori*, because the chance of picking a false positive to replicate from among all published 'positives' is higher; and secondly, the ES inflation makes it seem as if the effect of interest is very large when it is not (here, $d = 1.14$ and $d = 1.04$, respectively, while the actual ES is $d = 0.67$).[9] For the reader's information: to be able to detect the abovementioned difference in means with sufficient power (80 %), one would require a sample size of at least $n = 37$ for each group, so 74 in total. Also, consider the situation when there is, in fact, no true effect; i.e., no true difference between the groups described earlier. In this case, statistical power will drop to its lower bound, i.e., α. The reason is that any and all effects deemed statistically significant will necessarily be a false positive in such a case, i.e., a type I error, and all produced ES will necessarily be inflated, since the true ES is zero (for more information on FDR, alpha and power, see Bartoš and Maier, 2022; see also Wagenmakers, Verhagen, et al., 2015, for a Bayesian perspective).

[8] Calculated using the `pwr.t.test()` function from the `pwr` package (Champely, 2020) using R (R Core Team, 2022).

[9] ES is calculated by dividing the mean by the standard deviation (Frost, 2019). That is to say: $\frac{|\mu_2 - \mu_1|}{\sigma} = \frac{|110 - 100|}{15} = 0.666\ldots \approx 0.67$

Finally, for completion's sake, it must be noted that statistical power as a concept is not strictly contained within frequentist philosophy of statistical inference. It has been translated to fit Bayesian contexts (e.g., Kruschke & Liddell, 2018), but its meaning is not identical to the one described above. One of the main differences is that the ES of interest is punctate for frequentist statistics, while a Bayesian approach allows to specify distributional uncertainty. Discussion in the current thesis is restricted to a frequentist framework.

Figure 1



Note. Histogram containing binned frequencies of estimated differences. X-axis is estimated difference between every iterated pair of sampled groups which yielded statistically significant results from a two-sided *t* test at $\alpha = 0.05$. Y-axis is frequency of binned estimates. Vertical red line indicates true difference between the two populations. 126 out of 127 statistically significant results produced an overestimate of the true ES.

A HISTORY OF STATISTICAL POWER IN PSYCHOLOGY

The first systematic scrutinization of statistical power in published research articles in the fields of the psychological sciences was conducted by Cohen (1962). As mentioned previously, the concept of statistical power and type II errors had been developed long before the 1960s (e.g., Neyman & Pearson, 1933a, 1933b), but, as will become clear, this development was largely neglected by practicing researchers of psychological science. Cohen's (1962) publication was truly seminal, in the sense that it sits at the foundation of a now rich literature unveiling the nature and prevalence of what the author dubbed a "surprising (and discouraging) finding" (p. 151).

Cohen's (1962) investigation entailed a survey of a total of 70 articles from the *Journal of Abnormal and Social Psychology* (volume 61, 1960), the goal of which was to assess whether the statistical methods employed by the research reports published in said outlet actually had any chance at all of finding a small, medium, or large effect with a sufficiently low type II error probability; that is to say, to assess whether the employed designs were sufficiently sensitive. Note that only nondirectional tests at a nominal significance criterion of $\alpha = 0.05$ were considered, totalling 4.829 tests having been surveyed. As was explained before, a power or sensitivity analysis requires three out of four involved variables' values be fixed in order to determine the value of the fourth such that it satisfies the other settings. Published reports include sample sizes, which leaves ES of interest to be set. To ensure cross-study commensurability, Cohen (1962) calculated for each type of statistical test which was employed by any of the selected reports (e.g., *t* test, *F* test, χ^2 test …), a set of values which were deemed to accurately reflect a small, medium, and large effect in terms of the units of a metric-free population parameter—i.e., ESs were standardized.[10] For example, the metric-free unit of ES used for a simple nondirectional *t* test was $\frac{|M_1 - M_2|}{\sigma}$ (i.e., the statistical null hypothesis being tested states that two means are equal) and the corresponding values for small, medium and large ES were, respectively, 0.25, 0.50, and 1.00 (see Cohen, 1962, pp. 146 – 149). This medium ES would reflect, in practice, e.g., 8 IQ points difference between the mean IQs of two populations. Subsequently, for every selected article, the statistical power an exact replication would have had was calculated, assuming the previously employed design was kept intact, and this procedure was repeated for small, medium, and large effects. Note that it is implicitly assumed that the statistical hypothesis itself is valid, i.e., all preliminary assumptions are met (see Sedlmeier & Gigerenzer, 1989). The findings were quite distressing. For small effects, on average, the surveyed articles had—*a priori*—only about one chance in five of detecting an effect. Across articles, this varied between one chance in four to one in eight, with none rising above chance level. That is to say, all surveyed articles had statistical power < 50 %. For medium effects, again on average, the surveyed articles had a prior chance just below 50 % of detecting an effect. About a third of the articles had about one chance in three of detecting a medium effect using their experimental parameters, and two thirds did not reach chance level. For large effects, about five in six designs would have been capable of detecting an effect. Thus, Cohen's (1962) survey concludes that the reviewed articles "had, on the average, a relatively (or even absolutely) poor chance of rejecting their major null hypotheses, *unless the effect they sought was large*" (p. 151; emphasis added).

Throughout the years, many similar power-analytic surveys have been conducted, a non-exhaustive selection of which is discussed below. By 1976, eight such power-analytic surveys on twenty scholarly outlets had been conducted across several broad disciplines of scientific inquiry

[10] Although the chosen values for the relevant metric-free scales are systematically conceived (see Cohen, 1962), they remain highly subjective. In spite of this, it is impossible to rid oneself of the ES when dealing with statistical power, for it is the ES which effectively delineates the alternative from the null hypothesis. This creates a very interesting problem for the researcher wanting to perform a power analysis non-arbitrarily. This issue will be further discussed in the discussion section of this thesis.

(including, e.g., psychology, communication, sociology, *et cetera*; Chase & Tucker, 1976), all leading to the same general conclusion: the statistical power of surveyed samples of research articles in the social sciences is, on average, exceedingly inadequate, with most only marginally or not even reaching the currently conventional 80 % minimal power level, not even for large ESs (see Chase & Tucker, 1976, table 1).[11] Besides Cohen's (1962) survey, only two had investigated a psychological discipline (applied psychology [Chase & Chase, 1976] and education [Brewer, 1972]), and these followed the main trend: extremely poor statistical power for detecting small ES (< 30 %), inadequate statistical power for medium ES (< 70 %), and reasonable statistical power for large ES (around 80 %). Of course, at this time, there is sufficient reason to doubt the generalizability of these estimates, still. Not only because they are based on articles pooled from a small number of journals for each discipline, but also because there may be systematic differences between journals in terms of editorial requirements, methodological and reporting standards, *et cetera*. Moreover, research practice itself is likely coloured by field-related conventions, making it hard to produce valid statements on the pervasiveness of the statistical power deficit outside these specific subsamples.

The caveats put forward notwithstanding, Cohen's (1962) and others' examinations concerning the adequacy of statistical tests in the behavioural sciences *do* lay bare at least the possibility of a pervasive lack of statistical power throughout the field. A new turning point in the conversation on statistical power in psychological research was reached following Sedlmeier and Gigerenzer (1989), with the telling title: "Do studies of statistical power have an effect on the power of studies?" The authors compared arithmetic means of power-analytic studies conducted between 1962 and 1981, excluding those that were incommensurate to Cohen's (1962) findings due to methodological differences (a fact which itself called for procedural standardization across power-analytic surveys). They found that, with few exceptions, all scrutinized journals showed similarly tending statistical power levels for small, medium, and large ESs. Only for large ES did most, but not all samples reach a mean statistical power level above the currently conventional 80 %. Statistical power levels associated with medium ES tended mostly around 50 to 60 %, and these estimates tended to sink to unacceptably low values for small ES, most of which did not reach chance level. Sedlmeier and Gigerenzer (1989) concluded with respect to these findings that low statistical power is inexplicably pervasive: "Researchers paradoxically seem to prefer probable waste of time, money, and energy to the explicit calculation of power" (p. 311). Indeed, it is strange that researchers seemed to resist applying explicit *a priori* power analyses, even though the issue of statistical power had been receiving increasing amounts of attention from several scholars coming from different subfields during the preceding twenty-odd years

[11] Note that by this time, the metric-free values for small, medium, and large ESs had been somewhat revised (see Cohen, 1969).

(e.g., Cascio et al., 1978; Chase & Tucker, 1976; Cohen, 1969, 1973; Cooper & Findley, 1982; Fagley, 1985). One possible explanation for this state of affairs is that publications on the subject matter remained largely under the radar of most practicing academics, maybe due to a lack of dissemination of findings and recommendations across research domains (since rarely, if ever, have power-analytic surveys thitherto conducted crossed the boundaries of a surveyed domain, at least in terms of the journals chosen). Furthermore, Sedlmeier and Gigerenzer (1989) questioned whether such power-analytic surveys had any effect at all on the application of power analyses in the published literature, even within the confines of the originally surveyed journal. To seek answers, the authors revisited Cohen's (1962) original findings and compared them to the state of affairs of the *same journal* some twenty-four years later (i.e., 1984). Alas, for naught. A sample of 64 experiments published in the 1984 volume of the *Journal of Abnormal Psychology* [12] showed that statistical power levels associated with small, medium, and large ESs had essentially remained stable—if anything, the values had dropped further down. Specifically, the observed drop pertained to statistical power levels associated with medium ES. Sedlmeier and Gigerenzer (1989) put forward the rather ironic explanation that the lack of improvement is likely due to increased awareness of another deleterious issue, namely that of multiple comparisons on type I error rates—that is, corrections for dealing with inflated type I error probabilities due to multiple comparisons (e.g., Bonferroni correction) tend to negatively affect (already low) statistical power. If this explanation holds, it shows once again how completely preoccupied researcher of the time were in regards to type I error rates and assuring statistical significance, yet how completely ignorant in regards to statistical power.

From 1990 onward, surveys investigating mean and median statistical power levels for small, medium and large ESs have continuously been conducted. None report any kind of substantial improvement in comparison to the early 1960s. The stability of statistical power levels associated with different ES strata is especially worrisome in light of the fact that power-analytic surveys have tended to investigate increasingly sizeable samples of articles and statistical tests, and have tended to shift from single journal estimates to topically organized groups of journals. For example, Rossi (1990) investigated a sample pooled from several journals (*Journal of Abnormal Psychology, Journal of Consulting and Clinical Psychology,* and *Journal of Personality and Social Psychology*, volumes from 1982; total *n* = 221 articles, 6.155 statistical tests), and concluded that median power levels for small, medium and large ESs are 0.12, 0.53, and 0.89, respectively (using Cohen's revised [1969, 1988] nominal values for each ES stratum). This data reasserts that only if the ES of interest is truly large, it will most likely be detectable using the employed statistical procedures. A similar state of affairs was later asserted for educational psychology (estimated

[12] The journal had been renamed in 1965, and it continues to publish under said name to this day.

using means, not medians): statistical power levels associated with small ES remain unacceptably low (0.27), but those associated with medium ES slightly increased relative to previous estimates (0.71; but remain insufficient), and those associated with large ES retained their stable position at just above 0.80 (Osborne, 2008). Button et al. (2013) surveyed neuroimaging and animal model studies, but calculated power based on the reported summary ES from the studies themselves (either as Cohen's *d* or as an odds ratio [OR]). They concluded that neuroimaging studies have a dismally low median statistical power level of 0.08 (across 461 individual studies), while animal model studies varied between 0.18 for water maze designs and 0.31 for radial maze designs. Button et al. (2013) further calculated that the average sample size of such maze studies provides adequate statistical power only for instances where an ES of interest is at least $d = 1.20$, which corresponds to a real difference between two means of 1.2 standard deviations (note that $d > 1$ is generally understood to be quite rare for psychological research). Neuropsychology as a domain was investigated by Bezeau and Graves (2001), who surveyed three journals (*Journal of Clinical and Experimental Psychology, Journal of the International Neuropsychology Society*, and *Neuropsychology*). The authors deviate from the values usually ascribed to the verbal descriptors 'small', 'medium' and 'large', for they claim that the traditional conceptualization of ES terminology is too stringent for clinical applications (i.e., 'conventionally medium' does not equal 'clinically medium'). As such, they rescaled the values (in terms of Cohen's *d*) for small (0.20), medium (0.50) and large (0.80) ES to 0.50, 0.80 and 1.35, respectively. One could argue about the technicalities of such a translation, but the results were nonetheless discouraging: on average, statistical power levels associated with medium ES did not reach the conventional 80 %. One would hope that for clinical purposes, researchers pay more attention to the statistical properties of their experimental designs, yet this does not seem to be the case based on these data (even when ES values are upscaled to remain contextually meaningful).

Indicative of the fact that the concept of statistical power is gaining attention is the steady increase of power-analytical surveys to this day, investigation ever more specialized subdisciplines. For example, Helwegen et al. (2023) present the intriguing case of statistical power in network neuroscience. Data complexity in network neuroscience often requires dealing with issues of multiplicity (e.g., maintaining a stringent family-wise error rate). An informal survey of 1.300 case-control brain connectivity studies unveiled that only one in five sampled studies mentioned statistical power at all, a number of which was purely to caution readers against potential power issues in the study and not to actually calculate statistical power prior to investigation. Median power to detect an effect of $d = 0.5$ was 47 %, and to detect an effect of $d = 0.2$ it was 12 %. Only one in eight of the case-control connectivity studies surveyed by Helwegen and colleagues (2023) actually had a sample the size of which was sufficient to detect a small ES. These estimates dropped further assuming that the surveyed studies needed multiplicity corrections (statistical power of 24 % and 3 % at corrected $\alpha = 0.01$ for medium and small ES, respectively). As a final example of a discipline-wide investigation, consider Brydges (2018), who used z-curve analysis (see Brunner & Schimmack, 2020; Schimmack & Bruner, 2017) to estimate the average statistical power of 9.225 tests from gerontological psychology. The produced estimate of average power was 71 %.

As alluded to before, across the years power-analytic surveys have altered their methods, evolving from focussing on individual journals, to assessing several journals sharing topical similarities, to niche subdisciplines. Nowadays, attention is shifting toward statistical power levels of particular theoretical subjects, rather than pooling together data from across topics and/or journals. For example, Simmons and Simonsohn (2017) investigated the subject of 'power posing' from the social psychology literature. Power posing research is concerned with how physical, expansive postural stance dynamics function as regards physiological processes and general embodiment, and in relation to social power dynamics (Carney et al., 2010). Simmons and Simonsohn (2017) concluded that the average statistical power of the inspected studies ($n = 33$) was less than 14 %. This implies that, if power posing effects do exist, there was no practical chance of validly detecting it using the sampled research designs (though see Cuddy et al., 2018, for a rebuttal). Similarly, Sotola and Credé (2021) have investigated the statistical power of investigations on the topic of system justification theory (see Jost, 2018) on a sample of 180 reported *p*-values, and concluded—based on z-curve analysis—that overall average statistical power was 16 %. Interestingly, Mahowald et al. (2016) looked at the topic of semantic priming from the psycholinguistics literature, revealing that, although the surveyed publications ($n = 73$) had acceptable power overall (82 %), a subset focussing on the investigation of moderator variables of semantic priming effects turned out to be severely underpowered (53 % on average). Compared to the previous examples from social psychology, the latter example shows how research on different topics may be burdened by low statistical power in idiosyncratic ways. It follows that the power-analytic assessment of a research field can be misleading when trying to apply the results of such an assessment to the individual constituents that make up the larger field. Nonetheless, a pooled estimate retains some instructiveness; domain-wide power-analytical surveys indicate a general state of affairs which can be used to evaluate change over time, but surveys of individual research topics may unveil the cruxes underlying the asserted problems (e.g., de Vries et al., 2022; Feng et al., 2021; Mahowald et al., 2016; Nuijten, van Assen, et al., 2020).

In spite of the inherent difficulty in trying to interpret estimates that are aggregated across ever-increasing and sometimes fundamentally differing subfields (e.g., psychotherapy vs psychophysics), some authors have attempted to perform sensible cross-domain meta-analytical assessments of statistical power in the field of psychology at large.[13] They are, however, reasonably rare. To the current author's knowledge, only two attempts have been made at calculating such an aggregate estimate. A first is provided by Singleton Thorn et al. (2019), who conducted a meta-analysis across psychological fields and published their findings on the Open Science Framework (see https://osf.io/h8u9w/).[14] Spanning reviews from educational, occupational, management, clinical, psychiatric and neuroscientific literature (total *n* = 46 journals, amounting to > 8.000 individual research articles), Singleton Thorn and colleagues (2019) show that according to Cohen's revised ES benchmarks (1969, 1988), the estimated statistical power levels (and their respective 95 % confidence intervals) are 0.23 [.18 – .29] for small ES, 0.62 [.54 – .69] for medium ES, and 0.84 [.81 – .87] for large ES. In an attempt to try and assess to what extent the multiple potentially systematic sources of variance may lead to a skewed image of the pooled power levels, Singleton Thorn and colleagues (2019) performed several sensitivity and robustness analyses. This revealed minimal variability in the pooled power levels; changes in analytic design yielded at most a decrease of 0.058 on the statistical power estimate for large ES, and even smaller decreases for medium and small ES. Since the exactness of the point estimates is not immediately of concern, these decreases do not change the ultimate conclusions drawn. That is to say, according to Singleton Thorn et al.'s (2019) meta-assessment, statistical power levels in psychology tend to be (severely) inadequate for detecting medium and small effect sizes, and this has not changed markedly across time. In a similar vein, Stanley et al. (2018) have analysed 200 meta-analyses, comprising a total of nearly 8.000 individual papers, and have reported that the median of median statistical power levels in their sample is around 36 %. Only about 8 % of the surveyed studies reached the conventional 80 % statistical power level. Interestingly, experimental rather than observational designs were severely underpowered (median estimate for the former being well below 0.25, as opposed to the latter with 0.60). Stanley and colleagues (2018) also provide a deconstruction of the total estimate of the median level of statistical power into separate estimates for all investigated subdisciplines (see Figure 6, p. 1339). According to this decomposition, only behavioural genetics presents a median statistical power level > 0.80. However, behavioural genetics was also identified as having the highest average statistical

[13] An array of factors render cross-domain comparisons difficult to validly carry out. In fact, one should be careful in assessing statistical power estimates from differing research programmes within domains as well. In a very general sense, it is a fact that different theoretical concepts more or less require different apparatus, procedure and statistical methodology. Such differences, which may seem rather superficial and crude, could very well cause statistical power to vary systematically between research programmes, let alone between different domains or even the scientific fields that encompass them. As an example, neuroimaging studies usually employ smaller sample sizes due to logistical constraints, a fact which likely causes it have a systematically different amount of statistical power as compared to, say, psycholinguistic priming studies that can be done online.

[14] At the time of writing (01-08-2023), the respective manuscript is still unpublished, and to the current author's knowledge this is a deliberate choice on behalf of the manuscript authors. It is unclear whether the document has received some form of peer review.

heterogeneity, i.e., the highest ratio of observed to true effects variance (see Borenstein, 2022; Higgins & Thompson, 2002). This means that there is systematic variability across studies to such an extent that measures of central tendency should be interpreted with explicit care (the reason being that meta-analyses which have identified large heterogeneity ought to incorporate a random effects methodology to ascertain the value of an aggregate estimate, which necessarily widens the confidence interval around it [see Sedgwick, 2015]— to be clear, Stanley et al. (2018) used such a random-effects approach).

SUMMARY—AND NOW WHAT?

This brings the introduction to its necessary conclusion: the designs used in the psychological sciences are underpowered, and this deficit is apparently chronic. This is problematic, because psychology as a science is enveloped in its abundant use of NHST to adjudicate between statistical—and, by default, theoretical—claims. Underpowered studies lead to increased false discovery rates and ES overestimation. Consequently, these factors affect close replication efforts, and with it, the cumulation of scientific knowledge. Problematically, Brand and colleagues (2008) put forward that sample sizes which have been calculated to be appropriate according to an *a priori* power analysis are likely too small, still, because the ES used to conduct said power analysis is often obtained from published literature. As such, a self-sustaining cycle is generated, where underpowered studies yield overestimates, which are used to substantiate opting for small samples, thus yielding overestimates anew. The file drawer phenomenon (Rosenthal, 1979) strengthens this cycle, although it was shown in the previously presented simulation study that even when statistically nonsignificant data are published, ESs remain overestimated on average.

The historical overview which was presented seems to indicate no betterment whatsoever, but this image is not entirely correct. As was mentioned before, statistical power has been gaining attention in the last couple of years, as is reflected, for example, by review boards, funding bodies, publishers and similar entities requiring formal power analysis to be conducted if researchers want to obtain ethical approval, funding, consideration for publication, *et cetera* (Abraham & Russell, 2008). Recent years have also known a steady increase in the publication of power primers, the goal of which is to aid researchers in conducting *a priori* power analyses in a proper and informed manner, oftentimes wrapped inside a software package; for example, G*Power (Erdfelder et al., 1996), with ‘G*Power 3.1’ being the current standard (see Faul et al., 2009), but also several R packages, such as `metapower` (Griffin, 2021), `SIMR` (Green & MacLeod, 2016), `pwr` (Champely, 2020; first iteration from 2006), `mpower` (Nguyen et al., 2022), `WebPower` (Zhang & Mai, 2021), among others. Several authors have also produced simulation-based power primers for specific, often complex statistical models (e.g., for mixed-effects models

[Brysbaert & Stevens, 2018], random-intercept cross-lagged panel models [Mulder, 2023], multilevel models [Arend & Schäfer, 2019], latent profile analysis [Tein et al., 2013], complex mediation analysis [Thoemmes et al., 2010]), or for large-data designs (e.g., fMRI data [Hayasaka et al., 2007; Zarahn & Slifstein, 2001]).

Given this significant rise in attention for statistical power analysis, both in the format of power-analytical surveys, power primers, and social entities requiring it be done, one could wonder if after all this time, power analysis is perhaps starting to gain prominence in the published literature. In fact, first indications to this effect are starting to appear. A prominent example is Fraley et al. (2022), who compared statistical power levels in a sample of studies published between 2011 and 2019, from nine major social and personality psychology research outlets. For each of the nine years, one fifth of the published articles was randomly selected for analysis (totalling 1.812 articles, and 4.540 individual studies). The sampled studies were assessed in terms of whether they were adequately powered to detect an average effect size (ρ = 0.2).[15] Fraley et al. (2022) found that in 2011, six out of nine examined journals tended to publish inadequately powered studies (statistical power estimated at roughly 50 %), but by 2019, this had decreased to only one out of nine (statistical power estimated at 74 %). Eight of the surveyed journals had actually increased in terms of the statistical power to detect an average published effect size, with estimates ranging between 82 to 99 %. To quote Fraley and colleagues (2022): "At the risk of seeming hyperbolic, these data suggest that the research culture in social/personality psychology has undergone a monumental shift in research practices over the past few years" (p. 12). Indeed, the difference between these data and the estimates from previous power-analytical surveys is remarkable and promising (see also Bakker et al., 2020).

Fraley et al.'s (2022) encouraging findings beg the question whether similar improvements have taken place in other subdisciplines of psychology. That is to say, the severity of the problem as outlined in this section should definitely urge researchers to take into account the necessary role of statistical power in their designs, sooner rather than later. The reform movement's preoccupation with statistical significance has long detracted from doing so, by which means it may have inadvertently undercut its own intentions. Decreasing the amount of false positives is absolutely paramount, and tackling QRPs is part of the solution, but not at the expense of statistical power, for the number of true positives goes down with it; if anything, the decades-old neglect for statistical power may best be classified as a kind of questionable and insidious research practice itself. Neglecting statistical power is not just deleterious because of

[15] This is equal to Cohen's *d* = 0.41 (see Richard et al., 2003). The estimate is 'average' in that it is the typical ES to be expected in these particular subfields of psychology (i.e., not to be conflated with 'medium' as a descriptor of Cohen's *d* = 0.5). For the reader's information: ρ is pronounced 'rho' and represents a correlation, whereas *d* is a difference between two means.

how it affects FDRs and leads to ES overestimation, but because it undermines and weakens the very fabric of the statistical philosophy (that is, mainly NHST) by which psychological scientists hope to be able to make valid epistemic claims. To put it hyperbolically, though not entirely so: if statistical power is not addressed, the end result will be a field harbouring nothing more than a series of irreplicable *effects*, consisting mostly of false discoveries and—literally—unbelievable large ESs; all the while, empirical observations remain but loosely connected to some vague conceptualization of a 'theory', void of formality and rigour, and whose status as a valid science, as the grandfathers of psychology so ardently wished it to be, will remain questionable, for what is a science if it cannot produce cumulative knowledge?

A Systematic Review of Psychological Reports Published in 2016 and 2021

The replication crisis is hurting psychological science, and the reform movement continuously strives to mend the myriad problems which have been identified as harmful. Recent developments show how the reform movement is starting to shift its attention toward statistical power as a crucial element at the core of the replication crisis. After numerous years during which methodologists, statisticians and critics have been strenuously publishing on the nature of the statistical power deficit, its consequences, and how to tackle said deficit in a systematic fashion (i.e., by systematically conducting *a priori* power analyses), and given how institutional bodies are starting to demand statistical power be addressed if researcher want to publish reports or get funding for research, one starts to wonder if all these efforts have any desired effect? First indications to this effect are crawling into the limelight (see Fraley et al., 2022). This is promising, but more research of this kind is needed to draw a broader picture of changing practices.

With a solid and extensive background in place, the current thesis aims to further the abovementioned efforts, by conducting a systematic review of psychological reports published in a number of outlets in 2016 and 2021, to assess absolute numbers and potential trends across this five year span. The nature of this systematic review is entirely exploratory and descriptive, meaning that no use will be made of statistical inference procedures by aid of NHST or other formal methods. The main hypothesis to be challenged is as follows: based on the fact that statistical power has been neglected for so long, and given the unwieldy nature of scientific institutions, it is hypothesized that *there is no practically significant change* in the amount of *a priori* power analyses being reported in psychological research literature. The goal of this systematic review is to try and find indications to the contrary.

Additionally, some ancillary issues will be addressed. First of all, given the importance of ES to statistical power analysis, in cases where it is reported, it is of interest to see if sufficient detail is provided, both in terms of the nature (standardized or not) and exact value of the ES, and whether it is based on previous literature, a rule of thumb (e.g., Cohen's benchmarks) or something else. Secondly, it is of interest to investigate if there exist any notable differentiable trends between different subdisciplines of psychological science. For example, it could well be that the 'softer' subdisciplines, such as social psychology, are less likely to include *a priori* power analysis, for they often entertain less formalized theories, making informed statistical power analysis harder to conduct. Alternatively, 'softer' subdisciplines might also be *more* inclined to include statistical power analyses in their reports, because the statistical methods employed may not require complex simulations as compared to, e.g., neuroimaging research. As such, it would be interesting to map such trends as well, if any do exist.

In summary, the current research contends that no sizeable difference exists between the two time periods (2016 vs 2021), yet strives to find evidence to the contrary. Besides the absence of a positive trend, it is contended that the absolute numbers themselves remain unacceptably low, i.e., only a small proportion of published reports includes an *a priori* power analysis to justify the employed sample size. The current research additionally aims to explore how the papers that do include a statistical power analysis, go about it in terms of determining an ES of interest (i.e., based on theory, previous literature, other methods?). Finally, there might be substantial differences between subdisciplines, so, this avenue will be explored as well.

Methods

To accomplish the abovementioned goals, articles were sampled from prominent APA journals with publications in both 2016 and 2021. First, three psychological subdisciplines of interest were chosen from the list of subdisciplines recognized as such by the APA. It was decided to include 'Social Psychology', since the replication crisis has a rich history in this specific subdiscipline (as was extensively discussed throughout the introduction), making it an ideal candidate to assess whether improvement is happening. The other two selected subdisciplines are 'Educational Psychology', and 'Neuroscience and Cognition'.

To be able to actually start sampling articles published in these subdisciplines, the following steps were undertaken. First, all journals recognized by the APA as belonging to the selected subdisciplines were ordered according to their impact factor (IF; the exact values were obtained from Web of Science [www.webofscience.com]). When articles were catalogued under several subdisciplines, it was coded as belonging to the subdiscipline in which it had the highest IF. The initial goal was to draw at random one journal from the highest (Q1) and one from the lowest (Q4) IF quartile. However, prior to doing so, journals were screened as to whether they comprised exclusively articles meeting the exclusion criteria (see later). Because of this reason, and because some subdisciplines did not comprise any or few lower quartile journals, the lowest possible quartile was always preferred for the second journal to be selected. Only journals whose contents could be downloaded using Web of Science were considered.

Eventually, the following journals were chosen. Note that the IF of each journal may have changed since the period of data acquisition. For Social Psychology: *American Journal of Orthopsychiatry* (AJO; IF = 3.407; Q1), *Cultural Diversity & Ethnic Minority Psychology* (CDEMP; IF = 4.035; Q2), and *Psychology of Men & Masculinities* (PMM; IF = 2.893; Q3). The third journal did not contain enough admissible articles (only 5 in 2016, and none in 2021), but this had only become clear during article coding, so a replacement journal was selected.[16] For Educational Psychology: *Journal of Counseling Psychology* (JCP; IF = 5.088; Q1), and *School Psychology* (SP; IF = 2.945; Q2; whose original name *School Psychology Quarterly* had changed between 2016 and 2021). For Neuroscience and Cognition: *Journal of Applied Research in Memory and Cognition* (JARMC; IF = 4.6; Q1), and *Behavioral Neuroscience* (BN; IF = 2.154; Q4).

Next, articles were sampled. Due to practical constraints, it was impossible to inspect all published articles from both years and of every journal. Instead, for the 2016 and 2021 volumes, ten articles were sampled from each selected journal, totalling 120 articles (sixty from each year). The sampling procedure went as follows: first, articles were filtered in Web of Science to be of a type 'article' (to exclude, e.g., editorials) and language 'English'. Next, based on the total number of filtered articles *k* published in a given year, in a given journal, every $n^{\text{-th}}$ article was selected from the filtered list, such that $1 + 9n \leq k$, and n was maximized. So, for example, for a journal comprising 86 publications in a given year, every 9th article was selected, starting from index $= 1$, and jumping to the 10th, 19th, 28th, *et cetera*, until the 82nd $(= 1 + 9{*}9)$. When such a selected article met one or multiple of the exclusion criteria (see below), it was catalogued in an 'excluded' file, and the next article in the row was selected (so, in the example above, if the 1st article met an exclusion criterion, the 2nd was considered). If all articles between two consecutive $n^{\text{-th}}$ articles had to be excluded, the second $n^{\text{-th}}$ article was skipped, and the articles comprising the next batch between the second and third $n^{\text{-th}}$ article were iteratively considered. So, to stay with the previous example, if all articles with index ∈ {1, 2, …, 9} were excluded, the 10th was skipped (as it was preselected), and the process started anew for all articles with index ∈ {11, 12, …, 18}.

The exclusion criteria were: qualitative research, reviews, meta-analyses, methods papers. If an article did not belong to any of these classes, and yet did not contain an empirical investigation, it was also excluded. Analysis of individual articles was done manually, and information was coded in an Excel spreadsheet. The following identification tokens were saved:

- *APA discipline*

[16] As may be noted, AJO has a lower IF than CDEMP, yet is classified as belonging in a higher quartile. The current author acknowledges this mistake, but wishes to clarify that journal screening, IF screening, quartile binning and eventual journal selection was performed by third parties, before the current author joined the larger project.

- *Journal*
- *Publication Year and Month, date of export*
- *Title, authors, and keywords*
- *Times cited across databases*
- *Number of pages*

For each article, all reported empirical investigations were analysed. Of each empirical investigation, the following bits of information were coded:

- *Number of participants in the analytic sample*
- *Presence of power analysis (PA)*
- *If PA:*
    - *a priori or post hoc*
    - *mentioned in 'methods', 'discussion', 'results', or 'separate'*
    - *desired level of statistical power (e.g., 95; as a percentage)*
    - *desired alpha level (e.g., 5; as a percentage)*
    - *desired effect size (non-standardized and standardized)*
    - *desired variance (non-standardized)*
    - *desired values obtained from 'pilot', 'literature', 'theory', 'not mentioned', or 'other'*
    - *elaboration: 'short', 'medium', or 'long'*
    - *primary analysis (e.g., t test, SEM, hierarchical regression, …)*

Note that more variables were screened than those mentioned above, but they are not reported here. The current research was part of a larger effort, and so presents only information pertaining to the current goals. Also, oftentimes, articles mentioned more than one analysis, as most reports include whole sequences of analyses. To the best of the current author's abilities, the analysis which most fundamentally targeted the main hypothesis of an investigation was coded as 'primary analysis'. The reason the types of analyses were coded, was to explore whether there might exist a pattern in the absence or presence of *a priori* power analyses depending on the complexity of the primary analysis. Finally, all data was analysed using R (R Core Team, 2022). Visualizations were made using the `ggplot2` package (Wickham, 2016).

RESULTS

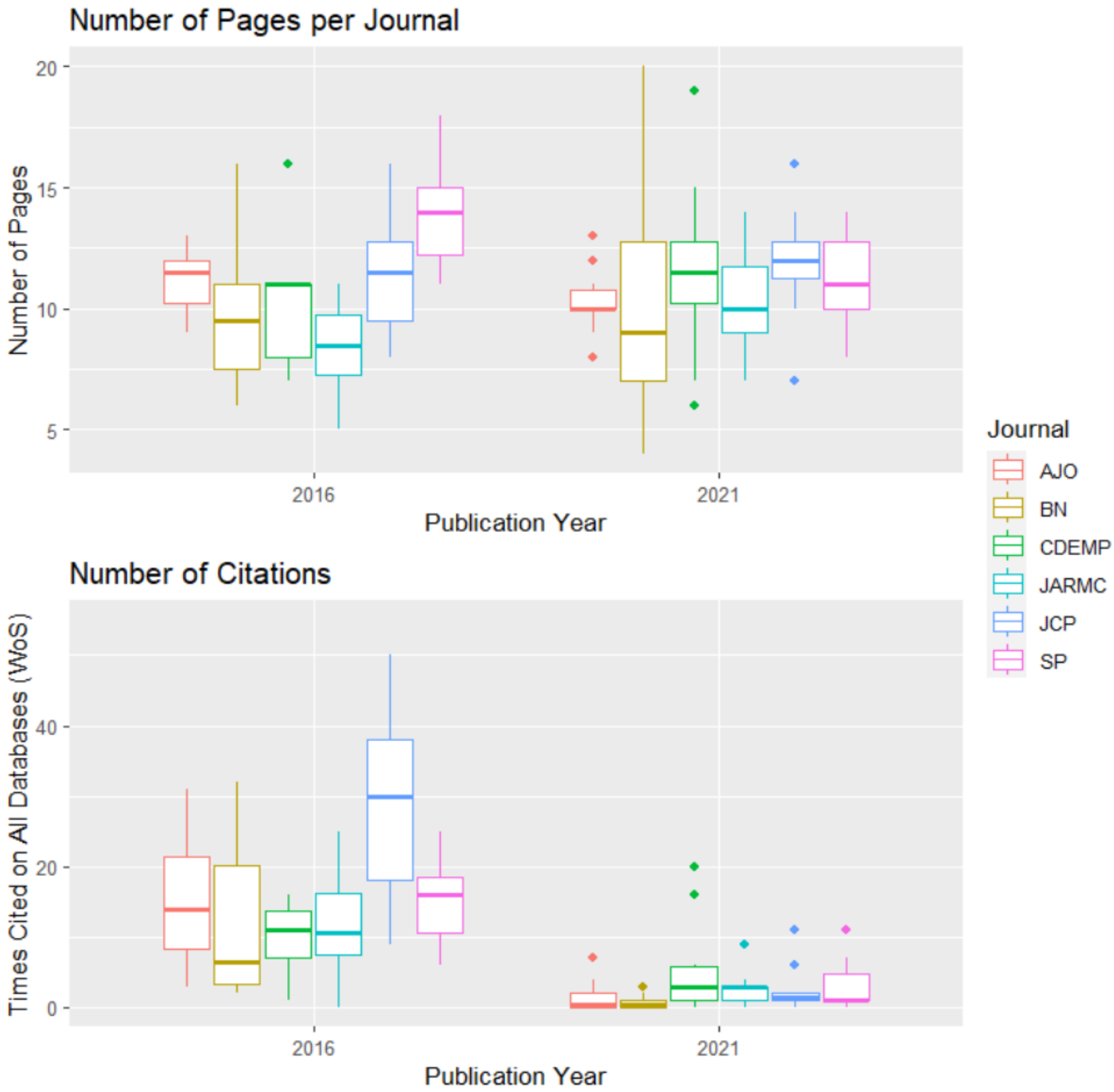
In total, 120 articles were screened, sampled from two publication years (2016 and 2021), three disciplines (social psychology, educational psychology, and neuroscience and cognition), and six journals (see Methods). First, some preliminary analyses are presented, after which the core results pertaining to the central research questions are discussed.

Data was gathered across several sittings, and by one person only (i.e., the current author). As already touched upon in the Methods section: the lower IF quartile journal for the social psychology subdiscipline did not contain enough articles that met the inclusion criteria, and for this reason a new journal was selected by the supervisor as per the outlined process. Of all selected journals, the lower IF varies between Q2 to Q4, which handicaps our ability to make descriptive, categorical comparisons between high and low IF journals. For this reason, such a comparison is omitted. A list of analysed and excluded articles can be found in appendix. In Figure 2 is presented an overview of the number of pages and citations, aggregated per journal, per publication year. The average length of the sampled articles did not change between publication years (μ2016 = 10.9, μ2021 = 11; aggregated over journals), but the number of citations is notably smaller for sampled articles published in 2021. This makes sense, however, since publications from 2016 have been circulating in the scientific literature for a longer period of time as compared to publications from 2021.

In terms of the presence of a power analysis in the sampled publications, there were eight articles from the 2016 sample (10 %) and ten from the 2021 sample (16.7 %) which included such an analysis. A decomposition of these numbers into journals is presented in Figure 3. Overall, only Educational Psychology journals show a decrease in the aggregate number of articles that include a power analysis (2016: *3*, 2021: *2*). However, within journals, JCP is stable (1 out of 10 for both publication years), and the decrease is situated in SP (2 out of 10 in 2016, but only 1 out of 10 in 2021). Social Psychology shows an overall increase (2016: *2*, 2021: *4*), produced entirely by the fact that the sampled articles from CDEMP did not include any power analysis for the 2016 batch. Apart from this, both batches from AJO and the 2021 batch from CDEMP included two articles wherein a power analysis was reported. The largest increase in number of articles that include a power analysis is found in the subdiscipline of Neuroscience and Cognition (2016: *1*, 2021: *4*). However, this increase is encapsulated entirely within JARMC, since the sampled articles from either publication years did not yield any power analyses in BN.

In Figure 3 is also included a further decomposition of the sampled articles that had reported a power analysis. The decomposition shows how many of these were *a priori* power analyses and how many were *post hoc*. The majority of reported power analyses were performed *a priori* (12/16). AJO contained one *post hoc* power analysis in the 2016 batch, CDEMP contained one as well in the 2021 batch, and JCP only contained *post hoc* power analyses in either batch. Of

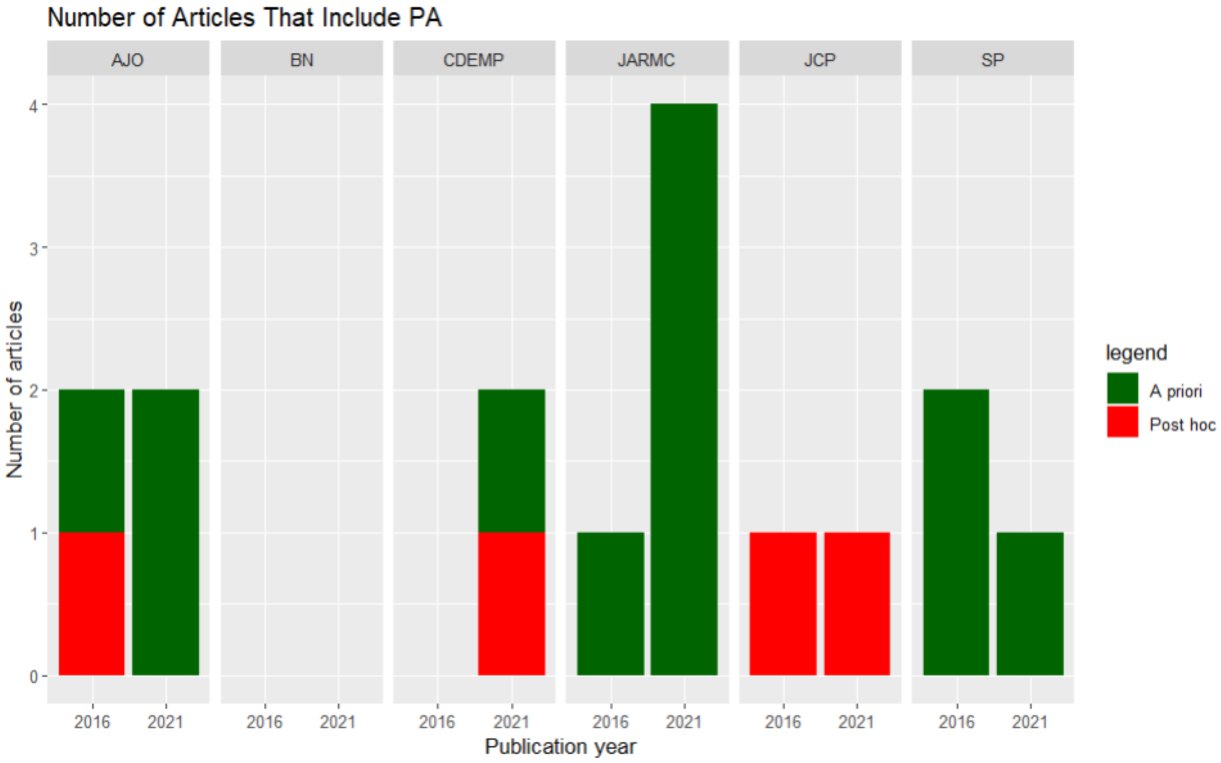
Figure 2



Note. TOP: Boxplots representing the number of pages of the sampled articles (n = 10) for each journal, divided by publication year. BOTTOM: Boxplots representing the number of citations of the sampled articles (n = 10) for each journal, divided by publication year. LEGEND: AJO (*American Journal of Orthopsychiatry*), BN (*Behavioral Neuroscience*), CDEMP (*Cultural Diversity and Ethnic Minority Psychology*), JARMC (*Journal of Applied Research in Memory and Cognition*), JCP (*Journal of Counseling Psychology*), and SP (*School Psychology; School Psychology Quarterly*).

all 16 power analyses, only eleven actually mention a desired power level; the mentioned desiderata ranged from 78 % to 95 %, six of which were exactly 80 %, the nominal statistical power level. Furthermore, only nine out of 16 power analysis reports actually specified the statistical significance level used to conduct said analysis. In all cases, it was 0.05. Interestingly, only three cases involved an unstandardized ES (two from SP_{2016}, and one from AJO_{2021}), but only one of these also defined a standard deviation to go with it (one in SP_{2016}). Ten others used some kind of standardized ES. The mentioned ES metrics were RMSEA (1; JSP_{2016}), rho-squared (1;

Figure 3

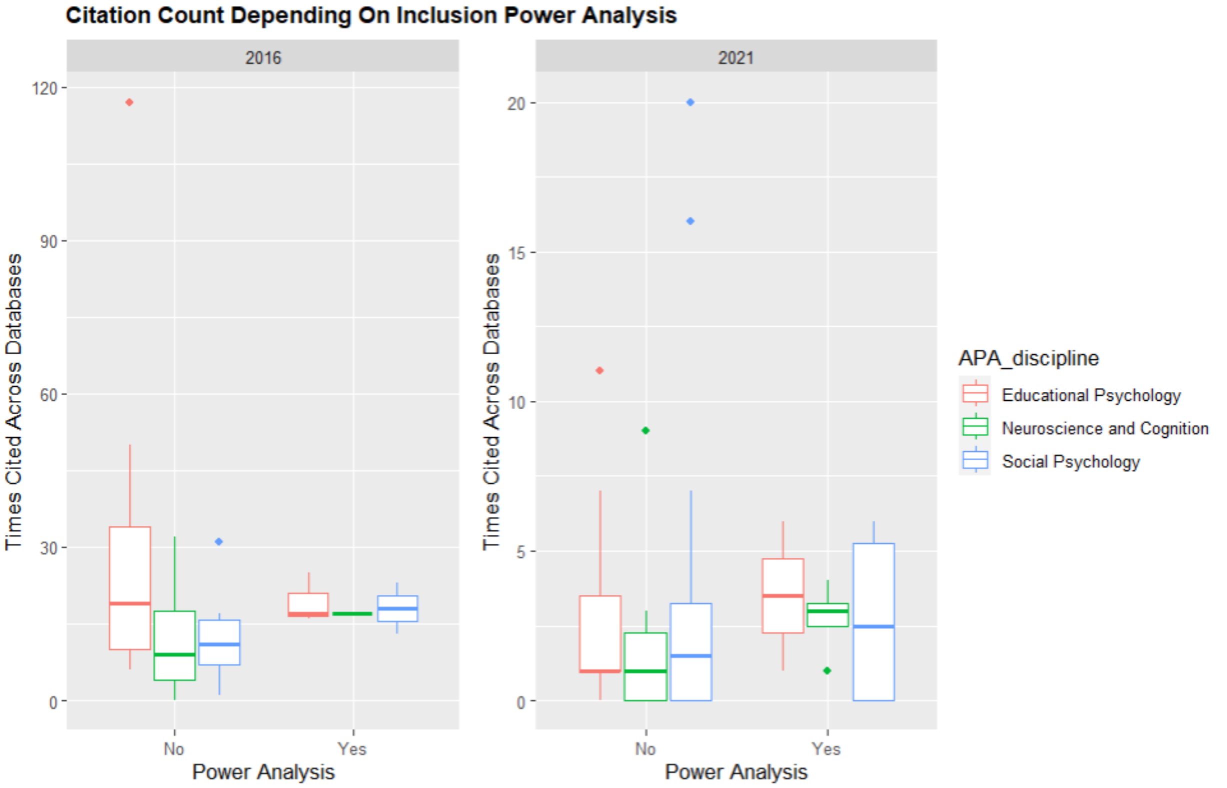


Note. Number of sampled articles that reported having conducted a statistical power analysis. Counts are divided according to journal and according to whether the reported statistical power analysis was conducted *a priori* or *post hoc*. LEGEND: AJO (*American Journal of Orthopsychiatry*), BN (*Behavioral Neuroscience*), CDEMP (*Cultural Diversity and Ethnic Minority Psychology*), JARMC (*Journal of Applied Research in Memory and Cognition*), JCP (*Journal of Counseling Psychology*), and SP (*School Psychology; School Psychology Quarterly*).

CDEMP$_{2021}$), Pearson's correlation (2; JCP$_{2021}$ and CDEMP$_{2021}$), Cohen's *d* (3; two in JARMC$_{2021}$, and AJO$_{2021}$), Cohen's *f* (1; JARMC$_{2021}$), and eta-squared (1; SP$_{2021}$). One paper referred to its ES of interest as "medium" (JARMC$_{2021}$), but provided no metric. Two others provided a value (as in, e.g., "the ES of interest was 0.17"; both AJO$_{2016}$), but no metric, verbal descriptor or further context. Only four reports indicated the origin of the ES of interest's value, and in all cases it was based on previous literature. Most reports did not elaborate much or at all on how exactly the analysis itself was conducted. In terms of primary analyses, no patterns worthy of mention arose from the data, except for the fact that only two articles that had reported a power analysis, employed an arguably 'easy' statistical method—that is, a simple Pearson's correlation significance test and an independent samples *t* test.. All other articles that had reported a power analysis tended to employ more complex statistical models (e.g., SEM, hierarchical regression, repeated-measures ANOVA, *et cetera*), although it must be stated that 'complexity' is somewhat of a subjective measure. The exact analyses are not discussed here, but can be consulted in appendix.

In Figure 4 are presented the citation counts per publication year, aggregated within subdisciplines, divided according to whether a power analysis was conducted. The boxplots reveal a

Figure 4

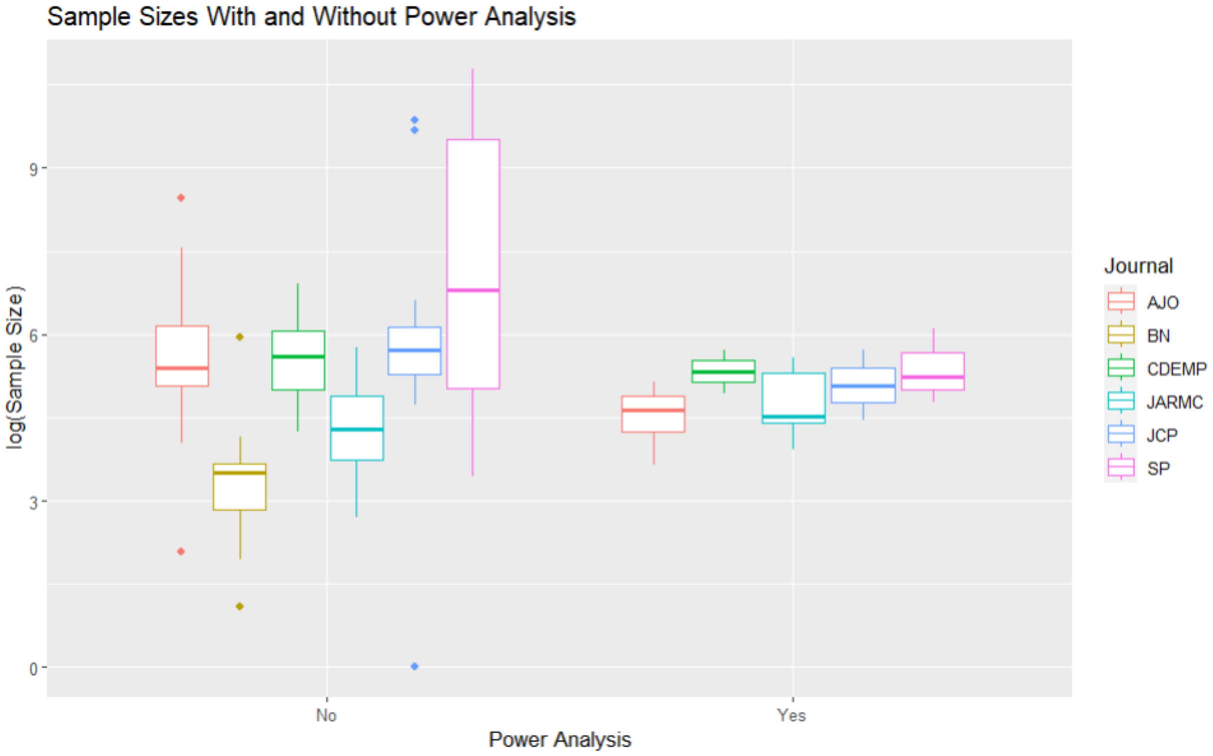


Note. Boxplots representing citation count across databases for articles that did and did not report having conducted a statistical power analysis. Counts are aggregated within subdisciplines, and divided by publication year to account for the temporal advantage of the 2016 batch.

soft trend where articles that included a power analysis tend to be cited slightly more than articles that did not. This trend is most outspoken for 2021, but is overall difficult to assess given the presence of some outliers in both publication years. Notably, these outliers occur only for articles that did not include a power analysis.

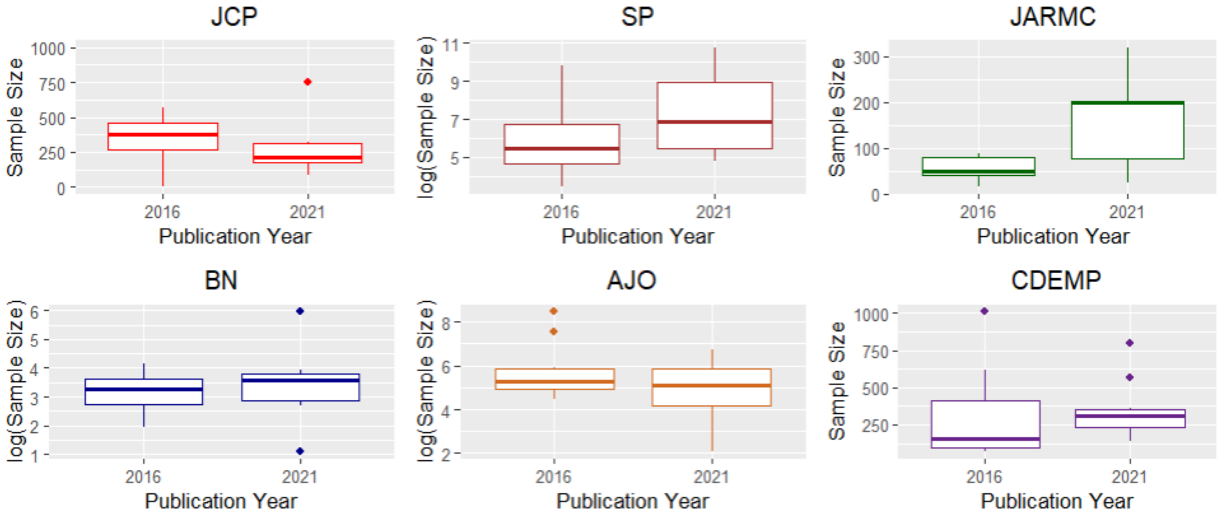
Finally, the articles were analysed in terms of sample size. In Figure 5 are presented the sample size boxplots, divided according to whether a power analysis was or was not conducted and reported. Inspection of this graph reveals that sample sizes show more spread overall when no power analysis has been conducted, while articles which included a power analysis seem to end up with analytic samples that lie much closer together. Note, however, that the y-axis is in a natural log scale, to compensate for the fact that some papers employed very large samples (e.g., *n* > 4000). It is therefore advisable not to take the graph at face value. Given the presence of extreme outliers, the differences are best numerically assessed using measures of central tendency that are relatively robust in the face of extremities, such as medians and interquartile ranges. Median sample size for articles without reported power analysis was m = 188.5, and with reported power analysis was m = 131.5 (also when *post hoc* analyses are excluded). Interquartile range for articles without reported power analysis was IQR = 329.5, while for articles with reported power analysis, it was IQR = 132.5 (m = 103.25 without *post hoc* analyses). In Figure 6 are presented boxplots of sample sizes within each sampled journal, divided by publication year (irrespective of whether a power analysis was included). Apart from JCP and—less clearly so—AJO, the sampled articles from each journal indicate that sample sizes have increased in 2021 relative to 2016.

Figure 5



Note. Boxplots of sample sizes by journal and divided according to whether a statistical power analysis was conducted. Y-axis is natural log-transformed to account for extreme outliers. LEGEND: AJO (*American Journal of Orthopsychiatry*), BN (*Behavioral Neuroscience*), CDEMP (*Cultural Diversity and Ethnic Minority Psychology*), JARMC (*Journal of Applied Research in Memory and Cognition*), JCP (*Journal of Counseling Psychology*), and SP (*School Psychology; School Psychology Quarterly*).

Figure 6



Note. Boxplots of sample sizes by journal, divided by publication year. Absolute scales have been natural log-transformed for SP, BN, and AJO, to compensate for extreme outliers. LEGEND: AJO (*American Journal of Orthopsychiatry*), BN (*Behavioral Neuroscience*), CDEMP (*Cultural Diversity and Ethnic Minority Psychology*), JARMC (*Journal of Applied Research in Memory and Cognition*), JCP (*Journal of Counseling Psychology*), and SP (*School Psychology; School Psychology Quarterly*).

Discussion

It is said that the psychological sciences are amid a replication crisis. From 2011 onward, it has become exceedingly clear that published effects that have long been deemed canonical, are in fact not replicable. Such lacking replicability casts doubts on the credibility of the field, and more specifically on its ability to generate cumulative knowledge. Coordinated attempts at producing close replications have produced overall disappointing results, and the main conclusion to draw is that a substantial proportion of the published literature is comprised of false positive findings that have failed to receive adequate scrutiny until recently. The concurrent reform movement has subsequently identified a series of QRPs which are said to inflate the false positive rate in the published literature, and which are believed to be relatively widespread in terms of engagement. As such, the reform movement has largely focussed on curative efforts the aim of which is to reduce QRP engagement. Unfortunately, the reform movement has largely ignored a major contributing factor: the persistent and historical neglect for statistical power in psychological research. When statistical power is systematically and chronically low, it results in an increase in the FDR of a body of work by reducing the probability of finding a true positive, and leads to overestimation of ESs. Decades of methodologists and statisticians surveying the literature, pointing out harsh facts, and proposing curative efforts have for the most part not had any considerable effect.

However, statistical power has started to gain attention in recent years. Publishers and funding bodies are starting to require *a priori* power analyses be conducted if authors wish to make use of their services, and knowledgeable and able persons are helping their fellow researchers by building and freely providing easy-to-use software packages for power analyses, and by publishing power primers on how to conduct an informed power analysis. A recent power-analytical survey by Fraley et al. (2022) has provided some preliminary empirical indications that statistical power in published literature is, in fact, positively evolving. To help build a bigger picture of changing practices, a systematic review of three psychological subdisciplines was conducted, investigating whether publications actually report having conducted a power analysis, and comparing whether such practices have increased in 2021 relative to 2016. Two journals from different IF quartiles were selected from Social Psychology, Educational Psychology, and Neuroscience and Cognition, and articles were randomly sampled from each of the two publication years of interest. The sampled articles were screened for whether they included a power analysis, and some ancillary data were gathered as well, e.g., the used sample size, the ES of interest, *et cetera*.

Based on the fact that statistical power analysis has been neglected in the literature for so long, the falsifiable hypothesis was posed that there would still be no general increase in the inclusion of statistical power analyses in the surveyed literature between 2016 and 2021. Exploration of the sampled articles reveals to the contrary: whereas only 10 % of articles contained a power analysis in 2016, this had risen to 16.7 % in 2021. These data speak against the main hypothesis, and so the conclusion is warranted that there seems to be a general increase in the inclusion of statistical power analyses in the surveyed samples between 2016 and 2021; however, these data cannot formally refute the stated hypothesis, they merely corroborate the alterative. However, the absolute number of sampled articles which had included a power analysis remained relatively small (i.e., both < 20 %). Provided that the samples are representative of their subdisciplines, this is a discouraging finding. It shows that much work is to be done, still.

There were observable differences between the surveyed subdisciplines and between individual journals, but in light of the limited total number of sampled articles which included a statistical power analysis, it is safer to assume that the current data are not representative. Educational Psychology showed a drop rather than an increase in sampled articles that included a statistical power analysis from 2016 to 2021, contrary to the other. However, the absolute numbers are so close together and the differences between years so minimal, that it is hard to imagine they represent a larger trend. Worthy of mention is the somewhat starker difference between 2016 and 2021 within the subdiscipline Neuroscience and Cognition (an increase from 1 to 4), which was carried entirely by one of the two sampled journals (i.e., JARMC). It is somewhat surprising that the BN sample did not include any article with a statistical power analysis, since Neuroscience and Cognition as a subdiscipline is often thought to be more of a ‘hard’ variant of psychology than its ‘softer’ siblings. To the extent that this finding is actually representative, it indicates that prejudices in terms of hardness and softness of psychological subdisciplines do not necessarily translate well to actual statistical practices of individual journals from such a subdiscipline.

A closer look at the sampled articles which reported a power analysis reveals that 75 % were conducted *a priori*, and the others *post hoc*. On the hand, this is promising, as *a priori* analysis is the proper way of conducting one, but on the other, it is worrying that, given the small absolute number of statistical analyses in the sample, a substantial proportion is conducted *post hoc*. This is problematic, because *post hoc* power analyses are often misinterpreted as providing additional information on the study that was just completed, while any statistical power analysis is inherently prospective (more on this later). Apart from one, all sampled articles that included a statistical power analysis had specified a desired power level of 80 % or more. This aligns with canonical recommendations to strive for a maximal type II error probability of 20 %. Surprisingly, only nine out of 16 reported statistical power analyses included an explicit desired type I error probability in their report. On the one hand, one may think that this is not an issue, as most investigations adhere to the standard 0.05 alpha level. On the other hand, however, if a researcher knows beforehand that an analysis will, e.g., require a multiple testing correction, said statistical power analysis would have to take this into account by fixing the desired alpha level at the corrected value (e.g., for a Bonferroni correction for a sevenfold multiplicity, $\alpha \approx 0.007$). Thus, it would be good practice to always explicate the desired alpha level which was used in the power analysis, for if a multiplicity correction was wrongly omitted, a statistical power analysis may underestimate the total required sample size (i.e., alpha corrections generally deflate statistical power). To give a concrete example: if a researcher conducts seven two-sided t tests for independent means, an *a priori* power analysis for $d = 0.6$ at the nominal alpha level with desired power 0.8, results in a required sample size of $n = 45$. However, using the corrected alpha of 0.007, the required sample size increases to $n = 72$, or a 60 % increase. *Ceteris paribus*, if one would use only a sample size of $n = 45$, the family-wise type II error probability (i.e., of all tests combined) would be 46 % upon repeated sampling (i.e., if this particular experiment were repeated an infinite number of times).[17]

In terms of ES, three articles reported an unstandardized ES of interest, but only one actually provided the necessary information to calculate the corresponding standard deviation (without which it is impossible to conduct a statistical power level). Ten others reported using a variety of standardized metrics. Problematically, the final three provided rather vague descriptions of the ES of interest, with one referring to a “medium” ES without providing a metric, and two others providing a numerical value without metric, descriptor or context. “Medium” likely reflects the interpretation of Cohen’s d, although this is not certain. Taken together, the main

[17] Calculations performed using the `pwr` package in R (Champely, 2020; R Core Team, 2022).

takeaway from this is that there is need for standardization on how to report the input factors of a statistical power analysis, for the benefit of reproducibility. However, it also indicates that ambiguous statements on the constituents of a performed statistical power analysis get past peer review and editorial process, which is itself problematic, irrespective of whether these particular sampled cases are not representative of the larger groups they were drawn from. Additionally, most statistical reports were relatively non-elaborate.

In terms of whether there is a connection between primary analysis complexity and the inclusion of statistical power analyses, strong claims cannot be validly produced to a lack of data and systematicity in the acquisition of these datapoints. However, the plausibility of there being some connection remains intriguing, so future research should definitely look into this. Interestingly, the presence of a power analysis seems to be correlated with the total number of subsequent citations of an article. Especially for articles from the subdisciplines of Neuroscience and Cognition, and Social Psychology this seems to be the case in both the long and short term. Educational Psychology shows less clear of a pattern, but the sampled articles included a number of practically significant outliers in terms of citation count. One ad hoc explanation for this trend is that researchers put more trust in reports that include a power analysis, but it is equally plausible that the observed trend is entirely spurious due to the small number of datapoints and the unequal distribution of articles in the groups which did and did not include a statistical power analysis. To be able to make substantive causal claims, larger samples (preferably of equal size) selected on the basis of the presence or absence of power analyses are required, ideally in a (quasi-)experimental context.

Finally, the sampled articles were also analysed in terms of sample sizes. Interestingly, across journals and disciplines, articles which include a statistical power analysis tended to have smaller variance as compared to articles which did not contain such an analysis. Again, the necessary caveat of small number of datapoints applies. Aggregated across journals, articles with statistical power analyses had smaller median samples, the difference being > 50 subjects. The interquartile range of sample sizes between articles that did not include a statistical power analysis was almost 2.5 times wider than its counterpart. The interquartile range difference further decreases if only *a priori* analyses are considered (i.e., almost 3.2 times wider). These are promising, because they imply that including a statistical power analysis may be able to lower total logistical costs due to putting targets on sample sizes prior to data acquisition. This may inhibit over- and undershooting required sample sizes. However, the changes in median and interquartile range may also partly be explained by considering that when a statistical power analysis was conducted, it was almost always done using an ES of a similar standardized value (i.e., roughly $d = 0.5$). This could have drawn the sample size estimates toward each other across all journals. If more articles investigated smaller and larger ESs, one could expect the interquartile ranges to widen again. Also, given the fact that most true ES in psychology are rather small *a priori*, due to the multitude of covariates working in on psychological constructs, one could expect that medians would go up if they were included. Interestingly, the sampled articles reveal a trend of sample sizes increasing in general. Thus, lacking reports of statistical power analysis may not necessarily be bad, for larger samples generally increase statistical power.

In summary, the current systematic review suggests that reports of statistical power analysis are generally increasing, albeit to a limited extent, and the absolute numbers remain unacceptably small. The suggested trend itself is only tentative, however, given the limited number of datapoints. Somewhat noteworthy is the fact that one of the journals from Neuroscience and Cognition did not contain any mention of statistical power analysis among the sampled articles. Even though it is likely that a broader sample from BN would have resulted in the presence of statistical power analyses among the sample, this speaks against the oft-held prejudice that relatively 'hard' branches of psychological science are more methodologically rigorous, for a methodologically rigorous psychological discipline which relies on NHST would undoubtedly include statistical power analysis or sensitivity analysis in all its publications; that is, NHST as it is used today requires rigorous statistical power or at least sensitivity analysis be conducted. If anything, the overall small proportion of sampled articles that included a statistical power analysis speaks against any of the surveyed psychological disciplines possessing sufficient methodological rigour—inasmuch as such an extrapolation is warranted based on the limited number of sampled articles and journals. Moreover, among the negligible number of articles that included a statistical power analysis, a nonnegligible proportion had used a *post hoc* analysis. This is problematic, because *post hoc* power is often misinterpreted; that is, if researchers calculate statistical power using the sample size and *obtained effect size*, this 'observed power' is isomorphic to the observed *p*-value (Hoenig & Heisey, 2001; Pek et al., 2022), and, as such, adds no new information. Only when an ES of interest is used to calculate the statistical power level, some interpretative value may be gained, yet it would still only apply to a prospective study, for statistical power is essentially a measure of prospective probability of a certain kind of dichotomous decision error. The systematic review also indicates that there is need for standardization with respect to how statistical power analyses are reported. There was notable heterogeneity in the nature the presented information; some included an ES metric, some did not; some specified the alpha level of interest, some did not. Additionally, some preliminary indications seemed to suggest that the inclusion of power analysis could put bounds on sample sizes, such that there is no random and unnecessary over- or undershoot of effort and logistics. One slightly positive aspect of the data is that sample sizes seem to increase overall, which itself is beneficial for statistical power.

General Discussion

The replication crisis is a challenging subject for a multitude of reasons. The main problem is that it is a multifaceted topic grounded in a multibranched history and sociology of science. The meta-scientific literature is equally multifaceted and often diffuse, rife with simplifications, and scattered among differing philosophies of science. A tip of the iceberg has been addressed throughout the current thesis. First, a relatively thorough treatment of the notion of 'replication' was provided, on the different variants, several typologies, and the broader epistemic functions it may serve. It was argued that replicability as repeatability is desirable, but insufficient, in the same was that reliability of measurement is desirable, but not a sufficient criterion for its validity. It was argued that the classic division between *direct* and *conceptual* replications (Schmidt, 2009) was essentially reductionistic, but provides a solid ground for fruitful discussions on their merits and shortcomings. Whereas *more direct* or *closer* replications provide an opportunity to assess the psychometric invariance of the outcome of a specific experimental procedure (see Fabrigar & Wegener, 2016), *more conceptual* replications provide a means for assessing generalization and extension. However, not all nonreplicable findings are false, and not all replicable findings are true. A triangulated approach for assessing empirical postulates by targeting them through differing theoretical lenses, auxiliary hypotheses, procedures, *et cetera*—i.e., through causally independent replications—is therefore required (see Irvine, 2021; Radder, 1992). Following this deconstruction, it was argued that the current replication crisis is best qualified as one centred around close replications. This claim was justified based on the fact that all major replicatory efforts (e.g., Boyce et al., 2023; Ebersole et al., 2016; Klein et al., 2018; OSC, 2015) have explicitly aimed to stay as close as possible to original designs and methodologies, often by involving the original authors; but said claim was also based on the fact that 1) the reform movement has focussed on filling the void created by the file drawer phenomenon, and 2) psychological science simply does not currently possess the level of formality in theory construction that is needed to actually be able to conduct properly conceptual replications. To quote again van Rooij and Baggio (2021): "[M]ethodological reform so far seems to follow the tradition of focussing on establishing statistical effects, and, arguably, the reform has even been entrenching this bias" (p. 683). The current reform movement has focussed on mending those practices which inhibit close replicability and endanger its epistemic function, broadly captured under the acronym of QRPs, or questionable research practices. A distinction was made between presentational and antecedental QRPs. The former encompasses those practices which enable a researcher to present findings in a distorted or idealized way, after or during data acquisition. The latter involves the tweaking of experimental design prior to data acquisition, including, e.g., the amount of data to be gathered (e.g., sample size), the statistical method to be employed (e.g., within or between subjects), *et cetera*. Subsequently, it was argued that the reform movement has mainly been focussing on presentational QRPs, such as *p*-hacking and HARKing, and has long ignored the importance of antecedental QRPs, such as uninformed sample size determination and neglect for statistical power analysis. This observation concluded the first section of the current thesis.

In the second major part, a relatively thorough treatment of the concept of statistical power was provided. Its origin and place in Neyman-Pearson frequentist statistical philosophy was illustrated, as was the concept of type II errors. It was shown how insufficient statistical power increases the FDR of a body of experimental literature, and a simulation experiment visualized how chronic low statistical power inflates published ESs in the absence of null and negative outcome reporting. With this knowledge in mind, an extensive historical overview of the neglect of statistical power in psychological science was provided, from the 1960s to 2023. Although care is required in interpreting the individual reports on psychological science's lacking statistical power, the overall picture is crystal clear: statistical power in psychological science is subpar, FDR is likely elevated, and published ES are generally overestimated. In light of the fact that the field seems mostly obsessed with establishing *effects*, it is ironic that even this admittedly restrained scientific ambition is undercut to the extent that it apparently is.

However, recent developments speak against an overly pessimistic interpretation of the field's current state of affairs. Publishing companies, funding bodies and peer reviewers are starting to mandate the inclusion of *a priori* power analyses, and Fraley and colleagues (2022) have presented first indications that actual improvement regarding statistical power is happening. This recent change of affairs prompted the current thesis' main research question: has statistical power actually gained footing in psychological science, specifically in the format of publications including dedicated reports of having conducted statistical power analyses? To investigate this possibility, a systematic review of reports published in six scientific outlets across three major psychological subdisciplines was conducted, and a comparison was made between reports published in 2016 and 2021 to investigate the possibility of a trend. Contrary to the pessimistic hypothesis of no change, a slight increase of reports including a statistical power analysis was observed between the two surveyed publication years. This is a substantial difference compared to early and contemporary surveys on the prevalence of statistical power analysis in published reports, which tend to reveal that no more than 5 % of their sampled articles actually perform and report statistical power analyses (see, e.g., Bezeau & Graves, 2001; Fritz, A., Scherndl, & Kühberger, 2013; Olsen et al., 2018; Osborne, 2008; Tressoldi & Giofré, 2015; Vankov et al., 2014; though see Tressoldi et al., 2013). However, the absolute numbers remain quite small (< 20 %). In fact, the number of reports that included a statistical power analysis was so small that comparisons within and between subdisciplines were arguably futile, let alone within and between individual journals. Several aspects of how statistical power was conducted and reported were quite salient, however. For example, a quarter of statistical power analyses was reported as having been conducted *post hoc*. Moreover, there was a distinct lack of regularity and elaboration on how the ES of interest was determined and reported (i.e., standardized vs non-standardized ES, derived from theory or previous literature, short or long elaboration on methods, *et cetera*). Finally, two interesting patterns emerged from the sampled data in terms of sample sizes. The first concerns the fact that articles that had included a power analysis generally tended to showcase less spread across the sample sizes of individual articles, and this was the case for all sampled journals. Of course, given the small number of articles within each journal, it would be foolish to make strong conclusions at the level of individual outlets, but aggregated across subdisciplines the difference remains. The second salient pattern involved the overall increase in sample sizes across journals from 2016 to 2021 (although some did show small decreases).

In summary, one could view the general increase in reported statistical power analyses across journals and subdisciplines with cautious optimism. The fact that the current, relatively small-scale systematic review can corroborate Fraley et al.'s (2022) report affirms the belief that things are indeed getting better, as is indicated by changing publication and funding requirements. This is good news. However, as outlined, several issues remain problematic and these will be the topics of discussion for the remainder of the current thesis. Specifically, it is interesting that reports of having conducted a statistical power analysis remain so few in number, while, at the same time, sample sizes seem to be increasing across the years (at least, insofar as the sampled data is representative of the larger literature). Recently, calls have been made to simply increase sample sizes overall, for this would guarantee a similar increase of statistical power levels in published literature (e.g., Asendorpf et al., 2013). It is true that doing so would be beneficial for statistical power in a rough-and-ready kind of way. However, in the next segment it will be argued that this buckshot approach to solving issues of statistical power is not ideal on ethical and practical grounds. Furthermore, it will be argued that merely increasing sample sizes would still require a researcher to conduct a sensitivity analysis instead; statistical power is not a goal in itself, it is part of a an essential tetradic tool in the frequentist statistician's toolbox (i.e., statistical power, sample size, effect size, and significance criterion). By illustration of an *ad hoc* simulation, it will be shown how incessance on increasing samples by virtue of their having statistical power is not a sufficient solution, for ES overestimates remain and without statistical power *analysis* the extent of this overestimation cannot be captured. If anything, blind faith toward larger samples' capacities may further entrench the problems associated with publication bias as we know them today.

This will bring this thesis to its closing arguments. A very simple question will be asked, namely, *what keeps researchers from simply performing a power analysis?* The answers shall be found in history. It will be argued that several historical and persistent causes interact to create a so-called ‘perfect storm’. Firstly, questionable uses of statistics have historical precedent in inadequate statistical education, which persists to this day. Secondly, psychology researchers have a nearly unimaginable preponderance toward creating and employing ill-devised rules of thumb at every corner. These two issues will be discussed in tandem, for they are closely related. And thirdly, it will be argued that the idea that increasing sample sizes would be a satisfactory (part of the) solution fits perfectly in the history of psychological science and its inability to tackle issues at their core; as it stands, it fits perfectly in the current reform movement’s bias toward close replication and its apparently superficial conception of the crisis as merely one of *replication*, instead of symptomatic of a far graver issue. Said issue was preluded in the beginning of the current thesis: a nigh complete absence of formal and mechanistic theories of psychological constructs and phenomena. It will be shown that only with such theory can an informed statistical power analysis be properly conducted and its results inform research, by illustrating the difficulties of doing so in the absence of it. A call for major theoretical reform will close the final argument. A small conclusionary segment will then finalize the current thesis.

SAMPLE SIZE: THE ONLY WAY IS UP?

As was extensively argued, statistical power has not been a main subject of discussion in the larger discourse surrounding the replication crisis. Of course, this does not mean that statistical power has not been mentioned at all—quite the opposite—, it just means that it was not a priority like presentational QRPs were/are a priority for the larger reform movement. Often, when statistical power is mentioned, it is done almost in a transient manner, and almost always in reference to sample sizes; that is, extensive coordinated replicatory efforts (e.g., OSC, 2015) are adamant on making sure their samples are large enough to replicate the intended effect, should it actually exist, and this is virtually always achieved by simply increasing the sample size relative to the original experiment. In and of itself, this practice is not a peculiar fact, for sample size is a core element of statistical power. However, the focal point *at least seems* at times to be especially condensed on sample sizes in the respective literature, notwithstanding the fact that some excellent efforts have been made to expound on the subject in a more nuanced fashion. But is increasing sample sizes really all there is to it?

The notion that small samples are not a positive sign is not new. Several reviews have pointed out the fact that a nonnegligible number of articles published in psychological science journals presents findings from experiments that are based on arguably small samples (Cochrane & Duffy, 1974; Holmes, 1979, 1983; Holmes et al., 1981). Depending on the study, half or more of the sampled articles comprise a sample size of no more than $n < 50$. Dukes (1965) finds that quite a sizeable proportion of the literature leading up to the 1950s is based on $n = 1$ studies. There are, of course, distinct differences between subdisciplines—for example, experimental psychology in the lab is bound to employ smaller samples than, say, observational field studies (see, e.g., Muchinsky, 1979)—, but overall ballpark figures assume a magnitude of around a few dozen. Some exceptions exist, however. For example, Muchinsky (1979) and Shen et al. (2011) have studied the sample sizes used in *Journal of Applied Psychology* and found that between 1957 and the early 2000s, most published articles employed sample sizes of $n \approx 100$ to $n \approx 175$ (see also Reardon et al., 2019). In recent years, similar estimates have been published. For example, Marszalek et al. (2011) investigated four APA journals, comparing sample sizes between 1977, 1995 and 2006, and found that across abnormal, applied, developmental and experimental psychology journals, the reported median sample size was generally $n < 35$. For applied psychology, specifically, it had even decreased from $n = 32$ in 1977 to $n = 21$ in 2006. Another example comes from Nuijten, van Assen, et al. (2020), who conducted a cross-domain meta-meta-analysis on intelligence research, and with the exception of behavioural genetics, found that most research employs sample sizes between $n = 49$ and $n = 65$. An investigation by Simonsohn (2015) on sample sizes in *Psychological Science*, a prominent APA outlet, found that, based on reported degrees of freedom for t tests, the median sample size employed for said tests was about 20 per cell, and most did not cross the threshold of $n = 150$ per cell. Sample sizes along this order of magnitude (into the triple digits) may seem quite reasonable, but a brief skimming of Cohen (1992) will quickly show that such figures are only adequate if the ES of interest resides in the upper echelons (say, Cohen's $d \geq 0.5$), especially when few degrees of freedom apply. One could argue that average and median sample sizes as those reported by the abovementioned authors are adequate because true ESs in psychology are, in fact, large, such that statistical power is guaranteed. But this is not the case. Sedlmeier and Gigerenzer (1989) found that studied effect sizes were generally what is colloquially understood to be 'small-to-medium' (i.e., Cohen's $d = 0.30$). This result is strongly and consistently corroborated by other reviews. For example, Haase et al. (1982) investigated the distribution of 11.044 ESs (denoted as η^2) from 701 manuscripts published in *Journal of Counseling Psychology* between 1970 and 1979, and report a median $\eta^2 = 0.083$, which corresponds to a colloquially 'medium' effect. However, mindful of critiques on measures of central tendency (e.g., McShane et al., 2020), it is perhaps wise to study the graphed distribution as well: it shows a marked skewedness of the η^2 distribution, with nearly 6000 instances of $\eta^2 \in [0.0, 0.09]$, most of them likely residing quite a bit further down from $\eta^2 = 0.09$ based on the extremity of the skew observed in said graph (see Haase et al., 1982, Figure 1). Alternatively, Cooper and Findley (1982) report that psychology textbooks are filled with oversized ES; the mean ES of their sample was $d = 1.19$ ($SD = .62$). However, it is reasonable to assume that textbooks will include studies that report very clear, distinct and large effects for didactic purposes, and which are therefore hardly representative of the larger field. More recently, Richard et al. (2003) compiled over 25.000 studies, finding that social psychology can be expected to yield ESs of $r = 0.21$ ($SD = 0.15$), which corresponds to a colloquially ‘medium’ ES. An estimate from the individual differences literature yields an equally small value ($r = 0.19$; Gignac & Szodorai, 2016). Finally, a domain-wide investigation of ESs reported in psychological literature by Stanley et al. (2018) shows that, based on 200 meta-analyses (pooling over 12.000 individual ESs), the mean absolute standardized mean difference is $d = 0.389$ (median $SD = 0.21$), subdivided as follows: 108 meta-analyses focussing on correlational research ($d = 0.458$), and 92 meta-analyses focussing on mean differences ($d = 0.291$). One must keep in mind that all review-based ES estimates are derived from published findings, and given the fact that statistical power is chronically low and null findings remain largely unpublished, it is appropriate to treat the reported values as probable upper bounds on the median/mean of the distribution of true effects. That is to say, the actual median ES is likely what is colloquially known as ‘small’ instead of ‘small-to-medium’ or ‘medium’.

So, in summary, most research published in psychological literature seems to be comprised of relatively small samples, and the true ES of interest—though, of course, differences between subdisciplines may be substantive—is likely relatively small overall. The consequences of these two facts coinciding is hardly trivial, and the case for simply increasing sample sizes is easily made using the elements so far provided in the current thesis. Small samples pull down statistical power. Low statistical power results in a higher FDR and ES overestimation. Presentational QRPs and publishing culture practically deny the existence of negative or null findings, so the published literature becomes rife with false positives that are overestimated. Then replicability enters the frame and, to put it somewhat sardonically, all hell breaks loose. As such, it makes perfect sense to want to find a kind of one-size-fits-all solution, for example, by simply increasing sample sizes in all cases. In fact, Cohen (1962) himself presented the idea that “[f]ormally, at least, the answer is simple: increase sample sizes” (p. 151).

The idea that sample sizes must go up was already proclaimed in the early days of the current replication crisis. For example, Asendorpf et al. (2013) urged editors, reviewers and readers of psychological science to “insist on [bigger sample sizes]” (p. 110). The authors acknowledged the different purposes attached to using a smaller and a larger sample (namely, exploration and determination), but still argued that systematic use of larger samples is needed to confirm effects found in smaller pilots. Another example is Perugini et al. (2014), who argued that “if the replication study has a considerably larger sample size, everything else being equal, it is arguably a better piece of evidence” (p. 330). This idea that ‘smallness’ is somewhat synonymous to ‘unconvincing’ is, to this day, quite entrenched in the psychological literature. A telling example comes from Bakker et al. (2020), who argue for clear sample size planning, if possible as part of pre-registration or registered reports, yet also state that “a more general shift to larger sample sizes in psychological research is still needed” (p. 12). That is to say, a nuanced approach is appreciated, but a buckshot approach will do the trick. To this end, they provide two weblinks to initiatives that may help increase sample sizes, for example, by making different research groups work together. The authors "hope that these initiatives will finally result in better-powered research in psychology which results we can trust" (p. 12). Some authors present arguments in favour of simple rules of thumb regarding sample size. For example, Simonsohn (2015) argued authors of replication studies could try to obtain a sample size that is 2.5 times the original study's sample size, except when original studies already have quite large sample sizes. Finally, some journals have started automatically denying to publish reports that are based on small samples, because small samples lead to inflation of measured standard errors, which taints scientific inferences (Smith & Little, 2018). The question is whether the simplicity of this buckshot approach really outweighs the benefits of conducting a statistical power of sensitivity analysis. The answer is: it depends.

Increasing sample sizes is a crude measure, but if done consistently, it is a perfectly sound solution to the problem of low statistical power. It is a mathematical fact within the confounds of theoretical statistics: all else being equal, if sample size increases, statistical power increases with it. However, it is precisely this crude, rough-and-ready approach that is perhaps least readily applicable in realistic settings, because increasing a sample size is rarely practical, and nearly always more expensive (Rossi, 1990). Baguley (2012) summarizes the following:

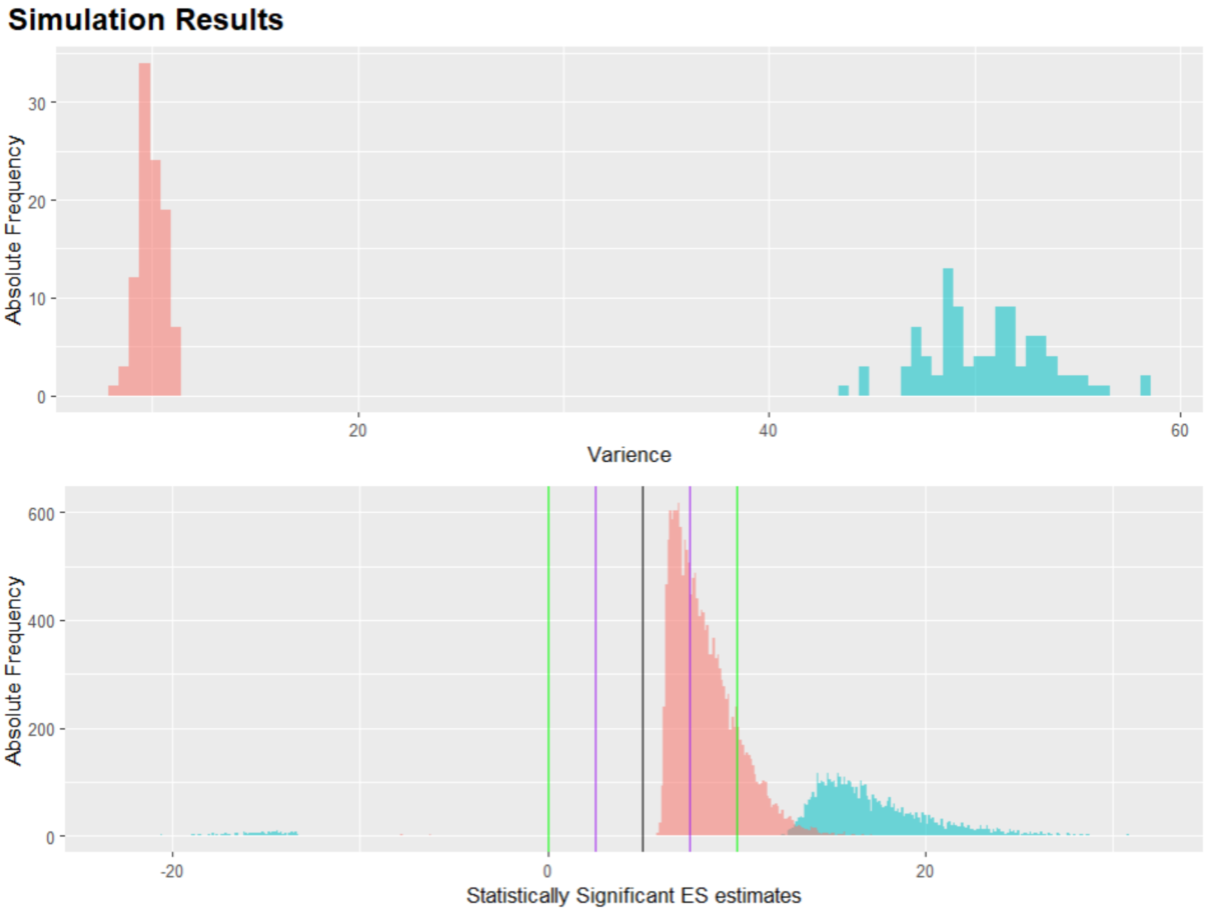
> "Research that is underpowered is wasteful of resources that might be used more productively and may expose people and animals to unnecessary risk or harm (given that the study has little prospect of finding anything useful. Conversely, overpowered studies may have a very high probability of detecting important or interesting effects, but at the cost of exposing more people or animals than necessary to adverse consequences."
> (p. 280).

It is a simple truth that limited resources restrict the scope of research that could possibly be conducted. For example, research may be costly due to financially demanding manipulations or treatments, and/or due to participants requiring compensation (Abraham & Russell, 2008). Other logistical constraints involve the availability of specific subpopulations (e.g., clinical samples; Clark, 2009), how much time can be invested in a particular study (Timmons & Preacher, 2015), and, depending on the used materials (e.g., brain imaging techniques), the financial demand of an investigation may increase exponentially. On top of that, psychological research receives less federal funding compared to life sciences in general. Stamm and colleagues (2017) report that in 2016, 3.2 % of United States' federal research funding went to psychological research (a little over 2 billion US dollars), compared to an astounding near 50 % for life sciences (accumulating to approximately 32 billion US dollars). These unfortunate truths force psychology researches to manage their funding with due diligence. The dictum of increasing sample sizes stands diametrically opposed to the required strategy in dealing with limited resources. As such, this approach is suitable only in those cases when there are resources aplenty. Note that non-WEIRD (Henrich et al., 2010) research groups—for example, in countries that have less resources to spend overall—are likely unable to commit to any approach that involves increasing sample sizes to some undefined magnitude. If, for example, journals and reviewers subsequently reject manuscripts solely based on the magnitude of such a sample, those who do not have the resources—of which a substantial part resides in non-western countries—will automatically fall behind and face even more unequal treatment than they currently already do (see, e.g., Bajwa & König, 2019; Begeny et al., 2018; Hida et al., 2019; Mason et al., 2021; Rosenstreich & Wooliscroft, 2006).

Secondly, apart from exploiting limited resources in a responsible manner, researchers also carry the responsibility of not exposing human or animal participants to unnecessary risk (Doke & Dhawale, 2013). Because psychological research pertains to human (and animal) subjects in particular, research ethics are especially important to the field. The American Psychological Association published the *Ethical Principles in the Conduct of Research with Human Participants* (American Psychological Association, Committee for the Protection of Human Participants in Research, 1982) as a guide for its members, but nowadays, researchers are often (if not always) required to receive some form of fiat from an ethical committee before they can actually conduct their research. General principles include, for example, the practice of informed consent, anonymity, protection from unnecessary mental and physical harm or danger, *et cetera* (Ethical Principles in the Conduct of Research with Human Participants, 1973). It is obvious that simply increasing sample sizes is, again, opposed to these fundamental principles. If true ESs are relatively small, larger sample sizes make sense, but just increasing sample sizes will inevitably violate the principle of not exposing more participants (e.g., mice, primates, humans) to unnecessary harm or otherwise adverse consequences than is strictly necessary. A potential counterargument may be that underpowered research due to small samples also violates said principles, as the outcomes of such studies are untrustworthy and inhibit proper inference—that is to say, the caused harm occurs without payoff. While that is certainly true, it does not therefore authorize researchers to abide by a buckshot dictum and find themselves at the other end of the spectrum. Both unnecessarily overshooting and carelessly undershooting sample sizes is ethically questionable.

Apart from practical and ethical concerns, the dictum to merely increase sample sizes would not nullify a responsible researcher's obligation to quantify the type II error probability. If sample size is maximized *a priori*, a sensitivity analysis is still required in order to determine the lower bound on the ES space from which point onward the result of a statistical significance test at a nominal alpha level of 0.05 can be guaranteed to adhere to the frequentist's notion of a prespecified statistical power level (e.g., 80 %)—that is to say, only then is the long-term gambit of frequentist statistics actually worth the pursuit. The specific problem lies in the fact that, even though larger sample sizes constrain the variance of the estimates of the same outcome—i.e., the spread of the estimate distribution becomes tighter—, and even though it does indeed result in higher levels of statistical power, what exactly those 'higher levels' are remains unknown as long as some indication of an ES of interest is absent. If a sample size is quintupled, from 100 to 500, on the regular, for a two-sided unpaired two-sample *t* test, statistical power levels would intuitively seem to skyrocket, but if—unbeknownst to the research team—the true ES of the phenomenon under study is Cohen's *d* = 0.1 (i.e., a colloquially 'small' ES), power increases only by a factor of 3.248, from a measly 10.8 % to a little over 35 %. In fact, the required sample size for the abovementioned situation is $n = 1571$ *in each group* in order to obtain a desired statistical power level of 80 %. Let us now consider a simulation of the above circumstances. Specifically, let us run one hundred simulations of a 500-fold replication of a two-sided unpaired two-sample *t* test to investigate the difference between two groups, having means $\mu_1 = 10$ and $\mu_2 = 15$, both with $\sigma = 50$, such that $d = 0.1$ (so, in essence, a structurally identical simulation as before is run, but now repeated 100 times to illustrate the behaviour of distributions of ES estimates). Results are presented in Figure 7. As was mentioned, one can immediately observe that merely increasing sample size (here: 100 to 500 participants) systematically reduces the variance of groups of estimates ($var(\sigma^2_{100}) = 8.16$ and $var(\sigma^2_{100}) = 0.39$); that is to say, if one systematically employs larger sample sizes, the obtained ES estimates will tend to cluster together. However, if only statistically significant outcomes are published, as often remains the case in contemporary literature, this quintuplicating of the sample size *continues to produce* overestimates of the true ES. In the *ad hoc* simulation study, 0 % (!) of 'publishable' estimates from the $n = 100$ group were smaller than 10 and larger than 0 (while the actual difference was 5). In the $n = 500$ group, this becomes 84 % for that quite wide interval, and diminishes quite rapidly to only 40 % if said interval is halved (i.e., [2.5, 7.5])—and, to wit, all of which are still overestimates. Now, consider this scenario, but one does not know the ES of interest, and one fails to calculate statistical power. Unless all research is suddenly published instead of only (nearly) statistically significant results, samples comprising 500 subjects will not suffice to actually get a hold of the true magnitude of the ES of interest. Now, let's imagine what would happen if a researcher *did* calculate statistical power, even if only by means of a sensitivity analysis. Then they would know that for $n = 100$, the lower bound on measurable ES given the preliminaries on the required level of statistical power, the significance criterion and sample size, is Cohen's $d = 0.4$, and for $n = 500$ it is $d = 0.18$. Then the researcher would at least be aware of the potential inferential problems when, upon replicating the design multiple times, publishable ES estimates return smaller than the safe lower bound. In fact, in the case of $n = 500$ from the simulation, said researcher would be operating within acceptable type II error bounds in only about 30 % of publishable simulated cases, but which would, based on the simulation, on average, deviate from the true ES by about 100 % (i.e., the true ES = 5, and, on average, the subgroup of estimates which behave according to the bounds set by a sensitivity analysis, resides, on average, about 5 units above the true ES).

Figure 7



Note. Results of the *ad hoc* simulation study. TOP: Variances of all simulated ES distributions according to sample size (red: 500, blue: 100). It is clear that larger sample sizes will result in ES estimate distributions that cluster closer together, and that are significantly smaller compared to smaller sample sizes. BOTTOM: Statistically significant ES estimates across all simulations, divided by sample size (red: 500, blue: 100). Vertical lines indicate the true ES (black), a large interval around the real ES [0, 10] (green) and a smaller interval [2.5, 7.5]. Larger samples result in more accurate ES estimates than smaller estimates, but the majority of ESs is still significantly out of acceptable bounds around the true ES. Notice the small blue blip near x = -20: this is due to the *t* tests being two-sided, meaning that samples which are too small can accidentally stumble upon an opposite effect that is statistically significant.

To extrapolate from this simulation to the real world: it is clear that failing to conduct a sensitivity analysis when statistical power is supposed to be a ‘fixed issue’ by simply increasing sample sizes can have serious consequences. The most important issue is that by failing to address statistical power in a direct manner, one implicitly assumes that whatever is the value of a true ES of interest, the sample size which has been ‘increased’ will likely cover it. But, of course, there is no prior guarantee that this is the case. What *is* known from meta-analytical research and surveys is that published research is mostly comprised of relatively small ESs, and the concrete estimates likely set but upper bounds on true ESs in light of extant publication bias and the generally underpowered nature of conducted research. Note that the problem's severity may increase substantially when interactions rather than main effects are studied, but this depends in large part on the pattern of means, standard deviations and correlations which substantiate the interaction of interest (see Lakens, 2020, for an elaborate primer on the subject). But, again, the extent of this potential problem cannot be assessed if one but blindly pursues a 'larger sample size' instead of conducting a power or sensitivity analysis. One can also look at it from another perspective: as was shown using the *ad hoc* simulation, if ESs are small—which they likely are—, increases in statistical power via increases in sample size are not even necessarily effective, in that the statistical power level will not necessarily be at or above the nominal 80 % threshold. But without conducting sensitivity analyses, there is virtually no way of ascertaining whether statistical power has actually increased to an acceptable level in the first place. That is, the only way one would be able to notice a lack of effect would be by observing that close replication efforts remain largely ineffectual. This is hardly a productive way of tackling as pertinent an issue as the current one. Lastly, there exists also the risk that when researchers erroneously believe their samples are sizeable enough to virtually guarantee statistical power, no matter the circumstances, they are likely to overvalue the practical significance of a statistically significant ES—or, from a pessimistic viewpoint, overvalue them *even more*. That is to say, misguided confidence in a statistical procedure due to belief in large samples may lead one to unduly overrate the verisimilitude of a finding to the extent that that is not already the case.

RITUAL AND RULES OF THUMB

All things considered, simply enlarging samples will not suffice. It is wasteful, just as is research with unnecessarily small samples. It is deontologically ill-advised, and blind faith in its ability to increase statistical power is worth nothing if there is no understanding of inherent limitations. One question is worth asking at this point: *What is keeping the field from simply conducting statistical power analyses en masse?* As has been thoroughly outlined, statistical power is necessary, and statistical power analysis is a necessary tool, but why do researchers fail to use it? At first glance, it is not at all obvious why this might be the case. Cohen (1962) describes that the observed lack of statistical power is—at least in part—attributable to the fact that sample size determination is often done non-rationally, for example, on the basis of tradition; *because everyone else does it in such and such a manner*. The fallacy behind such a lack of heuristics is likely fuelled by publication bias. If a study with a relatively small sample finds an effect to be statistically significant, and by that fact alone it finds its way into the literature, a novice might be easily tempted to simply copy aspects of procedure and method, because one may expect senior peers to conduct and publish research in good faith—that is, by having employed sound design. This way, convenience becomes rule of thumb, becomes standard practice. Additionally, it has been argued by many that statistical power is simply not all that well understood as a concept. In fact, Cohen's (1988) handbook was made for this very reason: to fill an apparent void in the curriculum. It is important to emphasize this point: statistical power as a concept had been defined a long time ago by Neyman and Pearson (1933a, 1933b), but it was simply not elaborated in statistical textbooks (Cohen, 1962). For example, Sedlmeier and Gigerenzer (1989) write that Guilford's (1956) influential textbook on fundamental statistics did not even address the issue of power, because it was deemed too complicated to discuss (also Gigerenzer, 2018). On this topic, John (1992) writes: "[i]t has been claimed [that] the statistics curriculum in US graduate schools fails to equip students to tackle many types of research problems of current interest (Aiken et al., 1990)" (p. 145), also

> "The teaching of statistics largely ignores basic philosophical issues pertaining to knowledge production (Dar, 1987) and, in distinction to other areas of psychology, avoids study of the central theoretical controversies and disputes in favour of instruction in the cookbook application of various statistical techniques, so that students come unthinkingly to apply tests of statistical inference routinely as a kind of knowledge increase ritual." (p. 146)

Osborne (2008) argues that this lack of exposure to basic concepts pertaining to statistical power has largely remained throughout the 20th and into the 21st century. Most researchers' understanding of statistical power is likely limited to its definition (Rossi, 1990), and there is solid evidence that researchers' intuitions on the matter are greatly flawed. For example, Bakker et al. (2016) conducted surveys with research psychologists and found that a substantial number of the surveyed researchers relies on rules of thumb for determining adequate sample sizes (53 % indicated they do not generally conduct statistical power analyses at all, and 23 % explicitly referred to applying rules of thumb, for example, 20 subjects per cell). Notably, nearly 10 % expressed feeling confident in using a sample size maximization strategy like the one currently being proposed as a solution to the chronic statistical power deficit. When Bakker and colleagues (2016) conducted a statistical power analysis based on the suggested 'typically sufficient sample sizes per cell', they found that for an independent samples *t* test, trimmed mean statistical power across respondents' suggestions was about 40 %. When respondents were asked to intuit on the statistical power of a specific design, it was found that they were not able to; most (89 %) respondents could not properly assess the statistical power of a small ES scenario. This is clearly problematic, as most published ES are small-to-medium, and true ES values are likely even smaller. Obrecht et al. (2007) studied laypeople's intuitions on statistics (undergraduates,

$n = 203$). They found that their participants had only a 'limited' intuitive sense of how statistical information of a sample's data is properly used in assessing pairwise comparisons (i.e., information on sample sizes, mean differences, and standard deviations). They concluded that "[a]lthough [participants] were more confident in a pairwise difference when mean difference and sample size were large […] and when variance was low […], subjects gave these factors far from the equal weight that would be expected if their intuitions were in line with statistical power" (p. 1152). Another aspect of flawed intuitions pertaining to statistical power is related to belief in the *Law of Small Numbers* (Tversky & Kahneman, 1971). It is repeatedly found that professional researchers are susceptible to the fallacious inclination to attach great representative value to findings from small samples. Specifically, in a whole array of scenarios, Tversky and Kahneman (1971) famously found, among other things, that researchers systematically overestimate statistical power and tend to trust patterns stemming from initial small samples as being stable. In light of such strong beliefs, a statistically significant p-value following some testing procedure is likely to be interpreted with undue enthusiasm, especially if it is based on a small sample (Dingledine, 2018; see also Guy, 1988).

Inadequate understanding of statistics and statistical power is likely also a factor which contributes to research psychologists' apparently incessant desire for rules of thumb. The statistical power threshold of 80 % was devised by Cohen (1962), but the field seems to have taken it as a mandate (see Bakker et al., 2016). The significance criterion of 5 % or 1 % itself is based on Fisher's intuition that it was convenient (Fisher, 1973), but this particular value is essentially arbitrary and it has been suggested to instead make a deliberate and justifiable choice (Alifieris et al., 2020; Lakens et al., 2018; Manderscheid, 1965; see also Machery, 2021). Similarly, the idea that increasing sample sizes will solve issues related to low statistical power is essentially a rule of thumb approach to a pertinent issue. In fact, it is highly reminiscent of another particularly interesting such 'rule of thumb': the case of '$n \geq 30$'. It is a persistent and popular myth that $n \geq 30$ is a good benchmark for sample sizes, perhaps as a guiding principle that once a sample size crosses this magical boundary, statistical significance testing is 'safe'. Such a sample is supposedly good enough to start trusting confidence intervals, omit type I and II errors at reasonable rates, and guarantee a level of precision that suits the scientific endeavour. Of course, no responsible statistician or methodologist would ever suggest such things, and, yet, this fiction is surprisingly widespread. The origin of this rule of thumb is likely a tragically pragmatic one: Fischer (2011) argued that textbook publishers found that at $n \geq 30$ the t distribution's p-values approximate normal tail probabilities using z-values, to the extent that the limited printing space on a textbook page could be spared by referring to z-tables for $n \geq 30$ (see Zhang et al., 2022). The continuing belief by some in this rule of thumb is problematic, because it goes hand in hand with other misunderstandings of theory of statistics. For example, Zhang and colleagues

(2022) surveyed social science researchers and students and found that the $n \geq 30$ 'rule' is likely grounded in a fundamental misunderstanding of Central Limit Theorem (CTL). Briefly, classic CLT has to do with the process of repeated sampling of any given distribution with finite mean and variance. According to CLT, the sample means obtained by such a process will, as the amount of samples being drawn tends to infinity, themselves tend to follow a standard normal distribution. More specifically, the random variable $\frac{\sqrt{n}\,(\bar{X}-\mu)}{\sigma}$ will follow a standard normal distribution (Heyde, 2006; Zhang et al., 2022). The rule of thumb is misguided, because, as Zhang et al. (2022) explain, it assumes 1) that all distributions follow this version of CLT (there are others, see Heyde, 2006), and 2) that they all do so at the same rate. Both assumptions are unfounded. That is, any distribution that does not adhere to the restrictions put on the moments of the distribution of the random variable at hand (i.e., non-finite mean or variance), will not converge to a normal distribution in the manner explained above (e.g., Cauchy distribution); and heavily skewed distributions require more repetitions of the sampling process for the resulting random variable's distribution to start converging to something akin to a normal distribution (e.g., χ^2 distribution).

The above example shows how small yet misguided rules of thumb can have serious consequences in the long run. The apparent refusal by many to just incorporate *a priori* statistical power analyses in favour of following buckshot-approach rules of thumb fits nicely among other, similar "cookbook" ways of using statistics in science. For example, Gigerenzer (2018) argues that the lack of statistical power is largely, if not entirely due to the fact that statistical analysis in psychology has become a ritualistic application of a messy hybridization of Fisherian and Neyman-Pearson statistical theory, the goal of which is to eliminate the subjective nature of researcher judgment by adhering to a procedural recipe of statistical inference. The ritual is threefold: first, one defines a null hypothesis of no difference, then one applies the conventional 5 % significance criterion, and, finally, one accepts the posited research hypothesis to a varying degree of significance when the null can be rejected (often signed with one, two, or three asterisks; see also Sedlmeier & Gigerenzer, 1989). No care is given as regards statistical power, consistent with Cohen's (1988) assessment that power is not discussed, even in reports where it is clearly relevant. The procedure is deterministic, i.e., resistant to subjective influences from researchers' judgments of statistics. This is, of course, incredibly ironic, given the fact that the whole procedure itself is based on a hardly objective understanding of the frequentist philosophy of statistics. Neyman and Pearson (1933a, 1933b) expressly stated that the null hypothesis ought not to be set at nil, and that the procedure requires defining a *specific* alternative hypothesis, while it is currently almost always denoted in the form of some nonformal and relatively vague verbal description that simply indicates a non-nil state of reality. Importantly, Neyman-Pearson decision theory does not involve a dichotomous *reject* or *non-reject* decision pertaining to the null, but explicitly includes a broad area of *doubt*, where data cannot reasonably be believed to corroborate either hypothesis. The inclusion of a significance criterion in the ritual is a useful remnant of Fisher's methods, who, by the way, vehemently rejected the entire notion of statistical power. Fisher (1955) deemed type II errors about as useful as "the type of mental confusion in which it was coined" (p. 73), stating that the concept has only come about due to the logic of significance testing being confused with that of acceptance procedures.[18] Furthermore, Neyman-Pearson statistical inference procedure adheres to a strict concept of frequency, in that a categorical statement pertaining to the truth value of any single hypothesis based on any single test cannot be reasonably made. Instead, they thought of hypothesis testing as a tool to guide on how to behave in the absence of more evidence than what had been gathered, and explicitly not as a justifiable measure of belief (see Fidler, 2005). Yet, nowadays, a statistically significant finding is often taken as a direct corroboration of the non-nil alternative hypothesis—i.e., as justifying belief in a vaguely defined non-nil alternative—, and any deviation from nil is considered supportive to this effect. The new rule of thumb to simply increase sample sizes to some undefined magnitude is additionally problematic, in that it likely fuels such misguided belief justification of merely non-nil alternative hypotheses: the idea that power must at all costs be maximized is counter to the philosophy of frequentist statistics, because as samples grow larger, *any effect* can, in time, be shown to be statistically significant, and not just *effects of interest*. As such, if samples are just *large enough*, any non-nil value may become statistically significant and, thusly, alternative hypotheses are always corroborated, for they are almost always defined as simply *deviating from nil* and current NHST practices leave little room for the concept of doubt.

An intuitive solution to both lacking statistical education and the abundance of ill-advised rules of thumb in psychological research could be to simply ameliorate educational programmes. If researchers in training are taught an improved curriculum, this could supposedly reduce the incessance in employing NHST and instead inspire the use of alternative methods, such as Bayesian, nonparametric and simulation designs. It could also minimize the use of mere rules of thumb. However, once again, this would likely be less of a grounded solution and more of a symptomatic treatment brought forth by erroneous intuitions. Stating that statistical education is the solution is akin to stating that 'bad statistics' is *the problem*, and while it is certainly so that NHST practices as they are so thoroughly and mindlessly practiced today are absolutely problematic, it will not solve the crisis. Similarly, stating that blind faith in rules of thumb, either by fault of lacking statistical education or by fault of mere adherence to tradition, is akin to stating

[18] Fisher's convenient system of rejecting a null hypothesis if empirical data resides in the tail of, e.g., a *t* distribution is fundamentally flawed in the absence of an alternative hypothesis, because, as Oakes (1986) explains: "without reference to an alternative class of hypotheses, there is no apparent reason to choose the tail area as a region of rejection. Any part of the sampling distribution with area of 0.05 would serve a similar purpose" (p. 122).

that rules of thumb are *the problem*; and, again, while blindly applying ill-devised rules of thumb is by no means ‘good’, merely ridding the field of these practices will not save it. The reader might wonder what, then, the solution ought to be? Well, as was already shown in the introduction, the abovementioned ‘solutions’ are not good solutions to *the problem*, because *the problem* is not statistical power, or, in fact, replicability. That is to say, these elements are highly problematic, as was thoroughly argued throughout the current thesis, but they are but symptoms of an underlying cause, and this cause is what must be tackled; not the symptoms.

The current thesis contends that what binds all aforementioned symptoms together is a pervasive lack of formal and mechanistic theory of psychological constructs and phenomena. It was already argued in the introduction that the ‘replication crisis’ is somewhat of a misnomer, for there seems to be a disconnect between what the reform movement focuses on to solve *the problem* and what meta-scientists have argued is the actual fundamental problem. In the introduction, it was argued that the current crisis is treated as one of replicability as repeatability, as is characterized by all major coordinated replication efforts and the referents of proposed solutions. For example, in all major coordinated replication efforts there has been adamance on including or getting feedback from original authors in order to guarantee to a higher degree that original and replicated effects are commensurate. The early reform movement’s efforts and critiques were directed at reducing the file drawer problem, because there was no certainty about whether published effects were real. Curative efforts are almost exclusively directed at ridding the field of QRP engagement, since those practices inflate the file drawer problem. Most importantly, however, it was argued that the current replication crisis *must be* one of close replication, because conceptual replications are technically impossible to carry out in light of the informality and descriptive incompleteness with which theories are expounded. There is simply no way of knowing whether a replication attempt is more or less conceptual than another, because the theories about which the results of those replications ought to tell us something are so thoroughly nondescript, that it is simply impossible to quantify the causal independence of two independent replication attempts. Thus, the reform movement, in wanting to improve replication, cannot but focus on *close* replication. The current thesis has mostly discussed the notion of statistical power—an issue so neglected by the reform movement at large—within the context of QRPs, showing how lacking statistical power inflates FDRs and ES estimates; these aspects reside in the world of close replication, still.

To argue this point, the following elements will be addressed. First, the observation of there being a lack of formal and mechanistic theory comes naturally from wanting to perform a proper *a priori* power analysis, in that one needs to choose an ES of interest. Furthermore, it will be argued that both procedural ritual and adherence to rules of thumb are necessary in light of absent such formal and mechanistic theory. For illustration’s sake, it will be argued that some major QRPs directly follow as well. Also, it will be shown that there is vast historical precedent on the issue of lacking theory. When this is done, the nature of formal and mechanistic theory will be clarified, and the natural progression from their existence to a reduction in ritualism, rules of thumb and QRPs, and an increase in informative statistical power analysis, close replicability, conceptual replicability and cumulative knowledge building.

ON THE NEED FOR FORMAL AND MECHANISTIC THEORY

In the current thesis, an account has been provided of what is a statistical power analysis and which benefits it entails, and which decrements may come from avoiding it. However, an account of the practicalities of conducting such an analysis has largely been omitted. This omission will now be rectified. The practical aspects of conducting an informative statistical power analysis will automatically reveal the natural problems of there being no formal and mechanistic theory in psychological science.

In practice, a statistical power analysis may be employed for either of three purposes: 1) to determine the statistical power achieved for a specific statistical test given a specified sample size, ES, and significance criterion, 2) to determine the lower bound of ES such that the employed statistical test may be able to detect it at a given statistical power level, sample size and significance criterion (previously denoted as a sensitivity analysis), or 3) to determine the lower bound of the sample size of a statistical test required given a statistical power level, ES, and significance criterion. It is also possible to calculate a significance criterion for a given sample size, ES, and statistical power level, but since it is traditionally set at the convenient 5 % level, such calculations are almost never performed (Sedlmeier & Gigerenzer, 1989). Ideally, one would conduct an *a priori* statistical power analysis of the third kind, i.e., to determine the sample size needed for the other specified values to hold. Since statistical power and the significance criterion are virtually always set beforehand, the ES remains. This procedure fits in the philosophy of scientific realism, for which effects one wishes to study exist independently from the research taking place, and by doing research, it is the scientist's aim to unveil this external reality. An *a priori* statistical power analysis serves to inform about which statistical design may be more or less capable, given a set of assumptions and preliminaries, to detect said effect. In practice, researchers need to assume an ES of interest, set the significance and statistical power criteria to one's wishes, and calculate the required sample size for the future research to accord with these parameter settings. Of course, as was already explained, sometimes one cannot flexibly adhere to the minimum sample size determined by such an analysis, in which case one may instead conduct a sensitivity analysis, the outcome of which is an assessment of the minimal ES obtainable and interpretable within the confounds of the set preliminaries.

It is here, with the first step, that things go awry: one needs to assume an ES of interest. Cohen (1962) dubbed this aspect “the most difficult problem […] in performing a power analysis of an experimental plan” (p. 146). Lipsey (1990) elaborates that ES is an essentially problematic parameter in this whole process because the true population ES is almost by definition unknown, and, more importantly, difficult to guess. This is the case not just because statistical methodologies virtually all rely on some set of assumptions concerning the samples and the population from which they are drawn (e.g., in terms of distributional properties of a variable of interest), such that the ES that is established via an experiment is almost guaranteed to be technically false, but also because the ES itself is influenced by the posed theories, none of which can guarantee *a priori* to yield results of the nature or in the direction of the actual ES. Senn (2002) provides the following illustrative metaphor: “The difference you are seeking is not the same as the difference you expect to find […]. An astronomer does not know the magnitude of new stars until he has found them, but the magnitude of star he is looking for determines how much he has to spend on a telescope” (p. 1304). In fact, it is perhaps not much of a stretch to state that, in many cases, the definite and *real* ES is unknowable, although it is also perfectly possible to obtain it if, for example, completely reliable census data can somehow be amassed (Kelley & Preacher, 2012). A recent survey has indicated that psychology researchers are likely very much aware of the problem of assuming an ES of interest: the commonest rationale mentioned for completely abstaining from conducting such an analysis was that researchers feel there is often simply no basis for determining an ES of interest (Washburn et al., 2018). In another recent survey, Collins and Watt (2021) asked researchers about their attitudes toward and usage of statistical power analysis. Nearly 60 % of the surveyed sample responded that statistical power is “very important”, and over 90 % deemed it at least “somewhat important”. However, the authors also included a questionnaire on why researchers do not choose to perform a statistical power analysis, even if it is highly informative for determining adequate sample sizes. Those that did not (or are at least discouraged to) perform such analyses often did not do so because they do not understand, feel unsure about, or lack information critical to perform a proper statistical power analysis, and did not know how to acquire it. Collins and Watt (2021) cite one particular participant who told them they felt especially discouraged to even try performing proper prospective statistical power analyses because of sheer co-author ignorance on the subject. The broader survey confirms this sentiment to a certain extent, by finding that of those researchers having no experience regarding *a priori* statistical power analyses, over 60 % could not sufficiently define the concept or did so incorrectly. Surprisingly, of those who *did* have experience, about one quarter defined statistical power incorrectly or did not know how to properly define it at all. Interestingly, only one surveyed participant actually mentions having had difficulty estimating proper *a priori* ES estimates, thus somewhat implying that the others might not. However, given the fact that, even though ES reporting is—very gradually—becoming the norm, a definition of ES is seldomly provided (12 %) and almost never justified (4 %; see Sun et al., 2010), it is highly likely that most researchers are not familiar with the intricacies of ES in the framework of statistical power analysis, or at perhaps falsely belief that their practices are adequate. While some have argued that although ES *per se* is a punctate concept, using a power function relative to changing parameter values is perhaps more appropriate (e.g., using a gradient of ES value in accordance with a more or less rough conceptualization of an alternative hypothesis; Maxwell et al., 2008). That is to say, a range of ES values requires less stringent a prior assessment of the part of the researcher; but it requires a *valid* assessment, still.

The reason why an ES of interest must be chosen beforehand is inherent to the function of a statistical power analysis. For any given investigation, the aim is to subject a hypothesized idea about external reality to a severe test, the outcome of which ideally provides the researcher with enough information such that they can reasonably choose to reject the hypothesized idea or find that sufficient evidence is provided to declare that the data corroborates an alternative hypothesis. In less ideal circumstances, the obtained information is ambiguous and the researchers must reasonably remain in doubt. The goal of a statistical power analysis is to provide the researcher with a means to control, prior to examination, the probability that if a choice to reject a hypothesis is taken, it is not done in error for a set frequency of iterated cases. In the case of significance testing following the Neyman-Pearson tradition, this is done by formulating an exact null hypothesis, an exact alternative, and acceptable probabilities of type I and II errors. The ES of interest is the difference between the null hypothesis and the alternative, either denoted as some form of difference or relation. Ideally, the ES is conceptualized as a minimal meaningful effect. Hence, it is clear that if one has no concept of ES, one, in fact, has no concept of either or both hypotheses. It is no understatement to say that if a researcher has no concept of ES, the whole statistical procedure of significance testing is effectively worthless. The outcome of such a procedure is uninterpretable, and any attempt to interpret it anyway and insert that interpretation into the literature would inevitably result in a corruption of said knowledge base.

But where does one get such an ES of interest? Two strategies are often employed by practicing researchers: either one uses an ES that has been previously reported in the literature, or one uses a standardized effect size with colloquial denotations like ‘small’, ‘medium’, and ‘large’ effects. Oftentimes, they are used in tandem. All of these strategies are highly problematic. First, it was implied before that, given the state of the literature, it is imprudent to simply take reported ES estimates as valid ES of interest, for they are likely overestimated. When statistical power analyses *are* conducted, oftentimes one makes use of so-called *standardized ES* values. In the systematic review included in the current thesis, almost all reports which did include a statistical power analysis mentioned employing some standardized metric, such as Cohen’s *d*, Cohen's *f*, correlation coefficients, *et cetera*. In general, effects can be reduced to either a difference or an association between two variables (Baguley, 2012). For continuous outcome variables, the corresponding ES metrics are contained within either the *d*-family or *r*-family, respectively (Rosenthal, 1994). For discrete outcomes, different classifications exist as well, such as phi coefficients, odds ratios, *et cetera* (see Fleiss, 1994). There are multiple reasons why standardized ES metrics seem useful in practice. For instance, due to their being standardized, commonly used ES metrics from different families can be transformed into each using easily applicable equations (e.g., see Rosenthal, 1994). Furthermore, they often have standard verbal descriptors. For example, Cohen (1962) proposed to define 'medium' as being visible to the naked eye of a careful observer, and purported that a standardized mean difference of 0.5 sufficiently approximates this for different subdisciplines of psychological research. The definitions of 'small' and 'large' then follow from the scale of the standardized mean difference; i.e., 'small' being around 0.25 and 'large' being around 0.80. One may easily observe the arbitrary yet practical nature of this standardized approach to ES. Not only does it enable a somewhat standardized form of scientific language in talking about magnitudes of effects, it also supposedly allows to compare different outcome scales that purport to measure the same underlying psychological construct.

However, others have opposed this approach, arguing that enforcing a general rule of thumb for 'small', 'medium' and 'large' goes against the inherent heterogeneity of psychological subfields. For example, as mentioned before, Bezeau and Graves (2001) argue that for clinical psychology, Cohen's 'medium' (1962, 1988) does not translate well to 'clinically medium' and ought to be rescaled. Nonetheless, they adhere to the notion of standardized language. However, there is arguably a more fundamental problem, which disqualifies the use of such standardized language altogether. Broers and Otgaar (2021) explain that the utility of standardized ES measures is reflected by their ability to generate transferrable information on ES values from different studies that investigate the same theoretical entity with different outcome scales. However, they go on to emphasize that standardized ES must be interpreted in relation to the original scale and cannot be detached from it. For example, "the .50 […] could be reflective of a medium sized absolute effect established with great precision, but it could just as well be a large absolute effect established with a not so very reliable instrument" (Broers & Otgaar, 2021, p. 14). The authors also enumerate in great detail several methodological choices that may influence standardized ES values upwardly by toying with the precision of the unstandardized measure, such as the number of levels of an independent factor, or the reliability of the measure. This makes it difficult to assess how standardized ES values may be graded to reflect 'meaningfulness' with respect to the effect implied by formulating a null and alternative hypothesis. That is, since there exists a certain malleability between actually measured ES, influenced by an array of elements, and the one being proposed to be measured, such that the interpretation of the former is easily conflated with the desired latter. In these circumstances, the notion of standardized language necessarily becomes nullified, because how can two numerically identical standardized ES values from two different investigations be commensurately 'sized' if they denote qualitatively different things? 'Small' in study A is not 'small' in study B, both in terms of how 'small' is interpreted in relation to the abstract construct and to the lower empirical circumstances. Broers and Otgaar (2021) conclude that, only if the absolute outcome scale is such that one is able to prespecify what constitutes a 'meaningful difference' (or association), does the use of standardized ES values for sample size determination via statistical power analysis become appropriate. That is to say, there is no way of omitting the specification of a rationale behind the ES that applies to one's specific research circumstances, because, as was stated, the ES that denotes the difference between a null and alternative hypothesis must be *meaningful* in terms of how it may be interpreted and used to qualify an adjudication of some theoretical postulate and its alternative(s). Hence, across several more or less close replications of an investigations, the same must apply; incommensurate outcomes are not interpretable as a whole, because their standardizations are incommensurate, still. The reader is reminded of a quote by the eminent Tukey (1969): "Being so disinterested in our variables that we do not care about their units can hardly be desirable" (p. 89). Indeed, for ES values to be usable in the context of statistical power analysis, they must be specified and interpretable *given the circumstances of the empirical investigation*.

A critic might argue that these reservations are valid in a theoretical sense, but do not apply to psychological science because the employed metrics *are* meaningful, and because close replications can mimic original circumstances to a close enough extent so as to circumvent the issue raised by Broers and Otgaar (2021). And maybe that is true for some 'harder' subfields of psychology, such as psychophysics. Fields that frequently employ psychophysiological measurement techniques involving heart rate, skin conductance, *et cetera* will likely be more able to defend the meaningfulness of their metrics. However, any reader would likely agree that the majority of psychological research does not employ these methods in isolation, but almost always with some *questionnaire* or *survey* containing a series of questions asking the same thing in opposite ways, or by briefly exposing a participant to some mundane stimulus and then make them commit to a dichotomous choice, the outcome of which supposedly tells us something deep about the human psyche (behavioural economics, social and personality psychology, softer branches of cognitive psychology and the likes are especially 'good' at devising such nonsense scenarios; cf. social priming). These scales are arbitrary, and, as such, inherently meaningless. And so are their standardized metrics. To wit, the aforementioned reservations with respect to standardized ES metrics apply specifically to taking previously reported ES metrics as a foundation for statistical power analysis. And not just because empirical circumstances differ and conceptualizations of scales contain variations which remain tacit and badly understood. Most often, different ES values between a previous and a current investigation are incommensurable, irrespective of how close the replication is, *because even the unstandardized measures are ill-conceived*. Consider the following: oftentimes, outcomes are transformed into a standardized metric and reported as such, because standardized metrics have an air of being easily interpretable via the aforementioned standardized scientific language. Apart from the problem of overestimation, picking an ES of interest from previous reports is unwise, because, unless a replication is *exact*—which it can never be—, it is a matter of high probability that what one is interested in, the psychological construct, will not be the thing replicated, for the replicated effect will likely be comprised merely of the ratio of the absolute effect to the precision of its measure. The argument is simple, and connects back to Tukey's (1969) remark: units of psychological variables are often *intrinsically meaningless*, such that a difference or association between two values is equally meaningless, and its division by the standard deviation in a sample remains meaningless. Tukey (1969) provides a specific elaboration on nonsensical correlations:

> "Why are correlation coefficients so attractive? Only bad reasons seem to come to mind. Worst of all, probably, is the absence of any need to think about units for either variable. Given two perfectly meaningless variables, one is reminded of their meaninglessness when a regression coefficient is given, since one wonders how to interpret its value. A correlation coefficient is likely to bring up the unpleasant truth—we *think* we know what r = -.7 means. *Do we?* How often? Sweeping things under the rug is the enemy of good data analysis. Often, using the correlation coefficient is "sweeping under the rug" with a vengeance." (p. 89).

Indeed, the regression coefficient of a meaningless scale is uninterpretable. That is to say, any interpretation that goes beyond a half-baked observation that two variables seem to somewhat go together in groups of participants, or that a difference exists on some amalgamated measure of construct X—itself constructed by clogging together standardizations of three different, rather arbitrary scales—stems from intellectual dishonesty. Psychological theories, by vice of their ill-conceived measures, can, as such, go no further than statements of the sort "if you pull on it, it gets longer" (see Tukey, 1969; Cohen, 1994). Meaningfulness, or a meaningful scale, may best be understood as one "where ordered steps in the scale can be clearly understood by the consumer in terms of progression on the construct the scale intends to represent" (McGrath, 2004; p. 128). To understand why a psychological measure is problematic, one need but to look at how it is designed: it is almost always an aggregate of Likert scale responses, yes/no dichotomizations and the like. McGrath (2004) provides an illuminating example:

> "An air temperature of 78 °F always means the same thing in terms of heat, although it may require personal experience with that temperature to develop a full sense of its personal and social meanings. […] Contrast this with scores on a depression scale generated

> by summing varying indicators of severity on a series of items that represent at best random sampling of domains relevant to depression. On a simple 30-item, true-false measure of depression, there are over 100 million different combinations of responses that could result in a score of 15. Is there any way that one can understand a score of 15, or an increase from 15 to 16, as having a consistent meaning under these conditions?"
> (p. 128).

Any single person who is somewhat familiar with psychological science will immediately recognize the arbitrariness of such scales and the meaningless inherent to them; and, worst of all, the pervasiveness of their existence. And they are pervasive! Fiske (2004) provides an enumeration that will surely be recognizable to many a research psychologist:

> "We measure most of our variables on verbally anchored rather than absolute metrics. Our dependent measures include Likert scales far more often than dollars or blood pressure readings. Our independent variables include the presence or absence of some context (e.g., salience) or attribute (e.g., race) that would be hard to quantify on a scale common to other independent variables (e.g., positive and negative feedback). Our variables are often categorical (yes or no) or ordinal (less or more); scales are rarely ratio, and even more rarely on a shared metric. […] the fact that we do not attempt to predict such data as absolute levels of bacteria, temperature, or dollars probably contributes to our verbal orientation." (p. 133)

No single serious person would dare deny that psychology is filled with such 'scales'. Consequently, if one ask 100 participants to fill in a Likert scale and one reports a standardized mean difference with some control group, after which another researcher uses this ES value to calculate a sufficiently large sample size, it is trivial that the second researcher is not setting up their future experimental design for successful replication of the purported meaning behind the initially reported ES, but is, in fact, setting up to replicate the ratio of the absolute effect to the extent it is 'reliably', let alone 'validly' measured with some arbitrary scale. A pessimist may look at this situation and ask themselves the pertinent question whether the label of *effect size* is really at all meaningfully applicable to this situation; a situation that is not so uncommon in psychological science. In fact, the softer a psychological subdiscipline, the more problematic their use of standardized metrics likely becomes, since it is mostly in those fields that creating a meaningful measure can be very challenging. Note that this argument is only somewhat of a slight toward the softer psychological subdisciplines, and it is mostly a mere stating of fact. Meaningful and reliable measurement construction in those subdisciplines is extremely difficult, because one needs to find out how more or less arbitrary scales relate to an abstract empirical relative, such as anxiety, which can be a cumbersome task (Schäfer, 2023). For example, it is easier to devise a psychophysiological measure of attention (e.g., pupil dilation as an indicator of attentional processes; Goldwater, 1972) than it is to devise a valid measure of something as ubiquitous as intelligence (for a history on the latter, see Gottfredson & Saklofske, 2009), while, arguably, both concepts—attention and intelligence—are very difficult to define as constructs. Even more difficult, then, is trying to devise a measure of, e.g., implicit race bias, a concept from social psychology that is societally very important, but, it seems, nigh impossible to measure validly (see Schimmack, 2019; Tinkler, 2012). But, as mentioned, the previous statement is a slight, still; if a scale is known to be inconsistent or dubious, or cannot be deemed reliable and valid beyond reasonable doubt, it should not be used—at least not by standardizing supposed *effects*. The difficulty of constructing valid and meaningful measures notwithstanding, it remains so that when a measure is inherently meaningless, so is its standardization, and so is the statistical power analysis the latter is used in. In spite of published doubts and criticisms (e.g., Baguley, 2009; Schäfer, 2023; Schäfer & Schwarz, 2019), standardized effect sizes are by far the most used and reported. And given that no two studies are the same, even when one tries to make it so by carefully adhering to descriptions of procedure (which are often incomplete and informal to some degree), it remains a difficult if not impossible task to try and compare outcomes of different studies purported to address the same subject. Yet, many researchers who conduct a statistical power analysis will use a previously reported standardized ES on the assumption that all aforementioned pitfalls are somehow avoided, without detailing *how* they are avoided.

ES metrics of the sorts described above are categorically uninformative for statistical power analysis, and they disqualify the utility of such analyses by infringing on the ability to adjudicate between refutation and corroboration of null and alternative hypotheses, let alone of candidate theories. It is exactly here that psychological science often starts to stumble over itself, and it is exactly here where the true problem of psychological science is to be located: in most of psychological science as it exists today, *there is no way of devising an informative and meaningful ES of interest*, and the reason for this is that in the majority of psychological science, the constructs being studied, the measures developed to study them with, and the connections between these two in the adjudicatory process which substantiates their formation, are partially or completely underspecified, ill-conceived, merely verbal, nonformal, and/or all of the above.

Let us take a step back and return to the essentials: the whole endeavour of science is to accumulate valid and reliable knowledge about the world, and the endeavour of psychological science is to accumulate valid and reliable knowledge about psychological constructs and phenomena. These are our primary explananda: capacities of human psychology (see Cummins, 2000; van Rooij & Baggio, 2021). Unfortunately, these capacities are often so complex (e.g., logical reasoning, attention processes, memory retrieval, *et cetera*) that no automated method can fruitfully devise computational algorithms that emulate these capacities and capture their essence. Rich et al. (2021) provide convincing proof of this by showing that inferring the exact nature of cognitive processes based on observing a cognitive system's behavioural output is an intractable problem, even in idealized situations; there is simply no way of devising any so-called "efficient abductive inference procedure" (p. 3037). Instead, research psychologists must resort to the formulation of "plausible explanations" of capacities. The point of the scientific process is, then, to assess the plausibility of proposed explanations and adjudicate between them. An essential ingredient in this process is that plausible explanations are sufficiently detailed and formalized, such that said progress can be rigorous and cumulative. The goal of an empirical, simulation or otherwise investigation is, then, to provide information to the researcher such that assessments and adjudications of these sorts can be performed. Corroboratory information strengthens an explanation, and refutative information must be used to iteratively adapt explanations, expose their weaknesses and hidden assumptions, and devise better explanations whose predictions are consistent with observation. Again, it seems almost trivial to explicate it here, but only if plausible explanations are sufficiently formal can a sufficiently risky prediction actually be made; risky, in the sense that a test of a prediction ought to be severe, and the riskier a test some explanation survives, the more corroboratory the outcome of said test is to said explanation. Ideally, a capacity of interest is approached from causally independent frameworks (in the sense of Radder, 1992). As Buzbas et al. (2023) argue: "[A]ccumulation of scientific evidence in support of a finding requires epistemic iterations and confirmation by independent approaches and methods to achieve specific scientific objectives" (p. 20). Scientific progress may then be understood as devising ever more plausible explanations of capacities, the fundamental goal of which is not to establish truth, but verisimilitude (Niiniluoto, 2014).

Contemporary psychology does little to achieve this ideal goal, and it actively abstains from even a rough translation of the abovementioned process. As was mentioned in the introduction of the current thesis, it is instead occupied with establishing *effects*, which are treated as self-contained entities and replace, in practice, capacities as explananda of interest. For instance, how often does one not read in psychological literature a statement of the following sort: "phenomenon such and such emerged from our data, and may be primarily explained by [insert convenient effect]". These are clearly established effects in the literature, but they are not explananda. Instead of theoretical entities which form part of a more or less formally conceptualized theoretical explanatory construction harbouring ideas and relations, *effects* are often employed as explanations in and of themselves. They are treated both as the explanandum and as the explanans. For example, De Houwer (2011) finds that "cognitive learning researchers who study the mental process of association […] often use the presence of classical conditioning effects as a proxy for the formation of associations in memory. Whenever classical conditioning is observed […] they conclude that association formation has taken place" (p. 203). De Houwer (2011) rightly points out that doing so violates the necessary gap between explanandum and explanans. By *explaining* association by referring to its proxy, which also functions as the explanation itself, the explanandum becomes the explanans. The result is a circular rhetoric, where the explanandum is explained by an explanans, which itself is, in fact, an explanandum, for it is treated as a proxy to the unobservable primary explanandum. Gawronski and Bodenhausen (2015) provide another example: the 'unconscious thought effect' refers to the phenomenon where a moment of distraction may lead one to make better decisions than when that same moment is used for active deliberation. The explanans in this situation is unconscious thought, for it explains the observed effect of distraction on decision-making. However, like associations in De Houwer's (2011) example, unconscious thought is itself non-observable, so a proxy is needed for empirical measurement. Problematically, the proxy is the effect itself: "[U]nconscious thought (the explanans) is empirically defined as the beneficial effect of distraction on decision quality (the explanandum)" (Gawronski & Bodenhausen, 2015, p. 69), and thus the thing-which-explains is equated to the thing-to-be-explained. Effectively, this style of reasoning is akin to stating that one can remember things because humans have the capacity for memory. This conflation does not aid in wanting to devise a meaningful ES of interest. An ES of interest is meaningful when its postulation is testable and the outcome thereof consequential with regard to a theoretical construction, but in conflating theoretical postulates of psychological constructs with their empirical relatives, an ES derived from said empirical element cannot be used to say anything meaningful about the theory under investigation. By deriving an ES of interest from such a corrupted measure, it becomes but tied up in the same circular rhetoric.

The oft-present confusion of explanans with explanandum, and vice versa, is synonymous with psychological science's obsession with *effects*. To state that it is *an obsession* is no overstatement. In fact, it is likely no overstatement to say that most of psychological science, by vice of said confusion, deals in *denotation* rather than *explanation*. As Cummins (2000) has stated, the field is "overwhelmed with things to explain, and somewhat underwhelmed by things to explain them with" (p. 120). To illustrate more explicitly what is meant, consider the following quote by Hempel (1952):

> "A scientific theory might [...] be likened to a complex spatial network: Its terms are represented by the knots, while the threads connecting the latter correspond, in part, to the definitions and, in part, to the fundamental and derivative hypotheses included in the theory. The whole system floats, as it were, above the plane of observation and is anchored to it by rules of interpretation. These might be viewed as strings which are not part of the network but link certain points of the latter with specific places in the plane of observation. By virtue of those interpretive connections, the network can function as a scientific theory: From certain observational data, we may ascend, via an interpretive string, to some point in the theoretical network, thence proceed, via definitions and

> hypotheses, to other points, from which another interpretive string permits a descent to the plane of observation" (p. 36).

One may easily recognize the notion of the *nomological net* (see Cronbach & Meehl, 1955). The fact of the matter is that most psychological scientists will, if pressed on the issue, admit adherence to a philosophy of science that is quite like the one detailed by Hempel (1952), while at the same time, their theories and ways of science as seen in their literature is nothing like it. Upon closer inspection, several important factors are often missing from any one theory of a psychological construct or phenomenon. For example, as was already argued, the observational elements are often poorly conceived or easily altered due to the malleability of the connection between desired ES meaning and actual ES meaning acquired from a specific set of empirical circumstances. It follows that the connection between the ground-level plane of observation and the floating plane of theory is, almost by definition, brittle and volatile. The meaninglessness of psychological metrics is often the result of lacking interpretive strings between empirical observation and theoretical postulate. However, the dearth of robust interpretive strings is not uniquely attributable to difficulties in devising uniform measures. An arguably much larger problem may be found in the conceptual foundation of psychology's objects of investigation—i.e., our *capacities*. Maraun (1998), whose ideas are based on Wittgenstein's philosophy, dubbed these *common-or-garden concepts*:

> "In marked contrast to technical concepts, common-or-garden concepts are not developed, laid down or modified at the outset of empirical investigation. This is because these concepts already have meanings, as manifest in their everyday use, use being governed by grammar. Hence, there exist grammatical restrictions on what one may legitimately do with them […] it is not the case that common-or-garden concepts *must* provide the conceptual foundation for empirical work in psychology, but merely that if the phenomena they denote are to be the focus of investigation, coherent empirical work necessitates that they be employed correctly. For when the meaning of a concept is subverted, the link between the phenomena and the concept that was supposed to denote them is severed: The denotational link is not established." (p. 454)

Franz (2022) clarifies that the problem is that objects of scientific investigation in psychology are inherently less malleable, relatively unchanging, because there meanings predate the adoption by research psychologists. Therefore, they 'arrive' in the scientific knowledge base of psychology without the level of conceptual clarity that would normally be required of a scientific object. Consequently, Hempel's (1952) theoretical plane is filled with underspecified notions of folk psychological concepts. It is impossible to make an interpretive string between such a concept and its empirical relative rigorous, because both the concept and the empirical relative constitute imprecise notions of ideas, much like a cloud of particles where nothing really touches each other.

Ideally, a concept is defined before it is translated to an empirical context, but in psychological science, notions are borrowed from outside and their colloquial, nonformal, imprecise meanings imposed. One cannot build a system of theoretical constructions if its building blocks consist of preconceived ideas. To make the problem tangible, consider a translation to psychology of the natural concept of ‘water’ by Weisberg (2006) in *Water is* Not *H_2O*: terms from folk psychology are much like water, in that at some point in time, a thing was “baptized” with a name, without knowing the intricacies of the thing just named, or, in fact, whether the name really applies to all other things for which the name is subsequently used. For example, At some point, the meaning of the name will have to be made explicit, likely in a scientific context, and where for water that is chemists, for folk psychology concepts that ought to be research psychologists. What should happen is a quick dismantlement of said term, a realization that the name of the thing is practically worthless because the name’s referent is diffuse at best. Such things like ‘anxiety’, ‘working memory’, ‘emotion’ *et cetera* have all been adopted from natural language, but are generally ill-specified for purposes of scientific knowledge cumulation. The reader is challenged to find a uniform and accepted definition of any of the core psychological constructs—they do not exist. It is quite logical, then, that psychological theory is hardly formal; theoretical constructs are, to wit, hardly specified at all. To give a practical example, consider the concept of *emotions.* Moors (2009) clarifies that one of the main theoretical problems in emotion research is the pervasively lacking consensus of its definition—which is, arguably, *the* foundational aspect of any theory. Theories of emotion have in common that they assume a series of different components needs to come together for a psychological event to be considered ‘emotional’, but there is no consensus on which components ought to be included, excluded, sufficient, extraneous, *et cetera* (see Moors, 2009, 2014). The difficulty of studying emotion may be found in the fact that it is a concept imposed on the field of psychology, and not one that evolved from the aforementioned ideal process of scientific progress. However, what seems to have happened is that, whereas chemistry recognizes ‘water’ as a product of natural language and not one of scientific language (see Weisberg, 2006), psychological science seems to be unable to let go of such imposed concepts. If such vague and abstract concepts are to be the object of scientific inquiry, it is almost necessary to resort to proxies in measurement—which is not a bad thing *per se*—, but in the absence of formality, these proxies are vulnerable to becoming the explanandum or for theoretical postulates to be clouded in quietly held assumptions and necessary vagueness, as a kind of brittle replacement for the vague concept that is now aimlessly wandering inside Hempel’s (1952) rather empty floating plane of theory. Of course, meaningful ES conception then becomes impossible; in fact, meaningful scientific practice becomes impossible. But, to be fair, psychology has been dealt a bad hand with its research subject. Smaldino (2020) summarizes the state of affairs best:

> “The social, behavioral, and cognitive sciences have, historically, relied on the power of the word. Words are powerful. Rich analogies can resonate in the minds of readers, appearing to illuminate the mysteries of nature. I’m talking about verbal theories—descriptive explanations of complex phenomena. Most theories are probably more workmanlike than poetic, but they generally rely on a property of most languages, whereby phrases can carry several possible implicatures—consider, for example, that words like ‘perception,’ ‘category,’ ‘identity’, ‘learning,’ and even ‘response’ are sufficiently ambiguous to allow for multiplicity of interpretations. That is, language is inherently (and adaptively) vague and ambiguous […] This is ultimately a problem for scientists, because we need to be exceptionally clear regarding *what we are talking about* in order to advance useful theories of the universe.” (p. 207)

Psychological theories cannot be clear, because their building blocks are rarely, if ever, clear. Meehl (1990) states that there are hardly any formal deductions in psychological science, let alone soft psychology, because how does one do so if theories are but verbal? Note that he makes the fair and necessary observation that psychological theories are not completely informal or merely verbal; consider, for example, subdisciplines which employ a mathematical model to explicate *part* of a theory under investigation. That partiality remains problematic, however, because “[the theorist or experimenter] often relies upon one or more ‘obvious’ inferential steps which, if spelled out, would require some additional unstated premises” (Meehl, 1990, p. 199). The problem of verbal theories is that they are very good for deceiving those who attempt to interpret or critique them. A theory that is richly verbalized, with deep concepts and complex rationales, is rather difficult to attack, because the outcome of a critical investigation likely depends more on one’s verbal cleverness, wit and perhaps even pedantry, in that minutiae can be strung together to explain away every potential critique in ways that would make Wittgenstein turn is his grave, than it depends on the internal consistency of the actual argument. Verbal theories do not need to be consistent, they just need to be convincing. Consequently, ES estimates for statistical power analysis do not need to be meaningful, one merely requires them to be easily transposed into the muddled language of the theory.

Unstated premises of theoretical constructs are part of the auxiliary hypotheses substantiating the main theory of interest, but which are hardly ever made explicit in psychological science. Auxiliary hypotheses span potential elements that stand between an experiment and the theory about which it ought to tell us something; it includes validity and reliability of instruments, of experimental manipulations, of statistical assumptions—in fact, of any non-explicated assumption in the experimenter’s belief system. Lakatos (1978) stipulates that empirical investigations usually challenge only those auxiliary hypotheses, “whereas the ‘hard core’ of the theory is protected from the *modus tollens*” (Dar, 1987, p. 148). *Prima facie*, it is unclear how one would go about doing so if auxiliaries are vague at best, and unstated at worst. The being problematic of unstated auxiliaries is not qualitatively unique to inexact sciences (Meehl, 1978); all sciences have to validate their instruments and quantify the nature and efficacy of stimulus inputs. However, in hard sciences, auxiliary hypotheses sometimes verge on derivability from core theories, exactly because Hempel's (1952) floating plane of theoretical constructions can be incredibly formally tight in the hard sciences. In soft sciences like psychology, with its lacking formality, auxiliary hypotheses can be easily—too easily—altered in the face of refutative data.

The following should ideally happen: a formal theory is devised, of which most, if not all, auxiliaries are defined, their restrictions known and explicated, and the wiring of the system logically consistent within itself and in its connections to other explanations. From this theory, an interpretive string links to an empirical postulate, a *prediction*, which may be a more or less precisely defined observable 'thing'. If one wishes to adhere to a frequentist statistical procedure, one has to formulate a null hypothesis which stands in opposition to the posed theory, and whose reality would be logically refutative of said theory. Subsequently, one may calculate an ES of interest, which encapsulates the empirical difference between the refutative case and the prediction of the theory of interest. If the system is formal and sufficiently specified, it should be possible to conduct a statistical power analysis that tells you if the nature of your devised experiment is such that a decision based on the outcome of a significance test would, in the long-term gambit that is frequentist statistics, adhere to preset type I and II errors. However, the outcome of an investigation likely leaves room for doubt, and, as has been stated before, singular investigations are likely insufficient for making bold claims. Nevertheless, the outcome of the study will be informative to some degree, even if it is not entirely convincing. An outcome may not be of the magnitude or in the direction that was expected, and this may be due to an auxiliary not having been tested, having remained tacitly assumed, or not having been identified. However, the formality of the system allows one to explore all auxiliaries, and devise tests for each of them; tests that are themselves theory-driven, and not just made up *ad hoc*.

Meehl (1967) explains how this idealized concrete example is explicitly *not* what often happens in psychology, and his explication reverberates with contemporary practices. Most notably, a decent proportion of psychological scientist commits to a more ignoble practice of *ad hoc* auxiliary testing whenever an outcome is not as per one's wishes. When a vaguely defined theory ought to be refuted—or, less stringently speaking, *fails to be corroborated*—psychology researchers often conduct a series of little experiment to mediate the original outcome away from the *modus tollens*. Meehl (1967) states:

> "[A] zealous and clever investigator can slowly wend his way through a tenuous nomological network [of auxiliaries], performing a long series of related experiments which appear to the uncritical reader as a fine example of an 'integrated research program,'

> *without ever once refuting or corroborating so much as a single strand of the network*. Some of the more horrible examples of this process would require the combined analytic and reconstructive efforts of Carnap, Hempel, and Popper to unscramble the logical relationships of theories and hypotheses to evidence. Meanwhile our eager-beaver researcher, undismayed by logic-of-science considerations and relying blissfully on the 'exactitude' of modern statistical hypothesis-testing, has produced a long publication list and been promoted to a full professorship. In terms of his contribution to the enduring body of psychological knowledge, he has done hardly anything." (p. 114)

The reader will perhaps not agree with Meehl's implied conclusion that most of psychology is filled with nonsense, but the reader will have to agree that to this day the described recipe of 'doing science' is very recognizable. Lacking formality in auxiliary hypotheses combined with loosely defined verbal theories allow one to use any empirical postulate which would naturally lead to its refutation, as the building block of some *ad hoc* adjustment to a once but merely vague conceptual idea that is now progressively and quickly becoming a patchwork of loose threads between observations and whatever beastly hybrid of theory is suffering its way through Hempel's (1952) floating system of platonic reality, until it is abandoned out of what seems to be mere *boredom* (see Meehl, 1978). One could ask the question whether in these circumstances it is really possible to abandon a theory out of any other consideration *but* boredom, since logical refutation is impossible to carry out, as is logical corroboration. Dar (1987) explains: "It is not just that theories persist in the face of anomalies or that researchers get involved in extensive revisions of their auxiliary hypotheses [...] the essence flow progression is the lack of accumulative knowledge" (p. 149). Lakatos (1978) decried this state of affairs in the social sciences quite vehemently, that is, the continuous stream of "patched-up, unimaginative series of pedestrian 'empirical' adjustments", arguing that "this theorizing has no unifying idea, no heuristic power, no continuity. They do not add up to a genuine research programme and are, on the whole, worthless" (p. 88). Note that it is not the adding of new auxiliaries itself that is problematic, but the ease with which it may be done in the social sciences relative to, say, physics. In hard sciences, the links between main and auxiliary hypotheses are far tighter, such that any successful challenging of an auxiliary necessitates a revision of the main hypothesis. In psychology, however, "*ad hoc* challenges to the auxiliary hypotheses are often little more than afterthoughts that do not have any real consequences for the substantive theory" (Dar, 1987, p. 149). A reader may object that these criticisms do not apply today, but the reader is challenged to provide an example of a research programme in psychological science today that actively stays away from any of the aforementioned shortcomings. And maybe they find one or two such programmes, but no serious reader of psychological literature can state unabashedly that the field has improved in terms of theory hardness compared to the twentieth century (save, perhaps, for behaviourism and some branches of computational psychology).

The vagueness of core theories and their auxiliaries lends itself quite easily to the development of rules of thumb and statistical ritual as discussed previously. In fact, such a development may be necessary to somewhat successfully cover up the inherent flaws of these theories. The most prominent example is NHST. Instead of devising meaningful null and alternative hypotheses, researchers often merely assume a null hypothesis of no effect, and when an effect—*any* effect—is found in the general direction of what could be expected if the verbal theory were right, said theory is at once accepted. If an effect cannot be found, a researcher is likely inclined to make an offhand reference to statistical power *maybe* being low, of the questionnaire *maybe* being invalid, of there *maybe* having been unforeseen 'demand characteristics' (Dar, 1987). The specifics of the ritualism will not be repeated, but their consequences deserve emphasis. The undesirable consequence is that *any effect* becomes an ES of interest. The nil hypothesis is always false—if only a sample is big enough to make a statistical test spit out $p < 0.05$—and, as such, its refutation hardly 'strong'. Meehl (1978) states: "the usual use of null hypothesis testing in soft psychology as a means of 'corroborating' substantive theories does not subject to theory to grave risk of refutation *modus tollens*, but only to a rather feeble danger" (p. 821; emphasis in original). He continues: "it follows that the probability of refuting [the null hypothesis] depends wholly on the sensitivity of the experiment", and "[p]utting it crudely, if you have enough cases and your measures are not totally unreliable, the null hypothesis will always be falsified, *regardless of the truth of the substantive theory*" (p. 822; emphasis in original). Bolles (1965) was among the first to point out that the NHST practice in psychological science is the result of a catastrophic conflation of what it means to refute a *statistical hypothesis* and a *scientific hypothesis* (or *substantive hypothesis*). That is to say, whereas a statistician can refute a null hypothesis at the 0.05 significance level and, consequently, decide to behave in accordance with the alternative merely *non-nil* hypothesis until more evidence is acquired, a scientist has no such privilege, for they make the additional assumption that the scientific hypothesis is adequately represented by its statistical counterpart. Bolles (1962) writes:

> "In assessing the probability of his hypothesis, [a scientist] is obliged to consider the probability that the *statistical model* he assumed for purposes of the test is really applicable. The statistician can say "*if* the distribution is normal," or "*if* we assume the parent population is distributed exponentially." These *ifs* cost the statistician nothing, but they can prove to be quite a burden on the poor [experimenter] whose numbers represent controlled observations not just symbols on written paper." (p. 639).

It is clear that if sample sizes were to merely increased, statistical power would increase, but the goal of scientific psychology would not be more easily reached. If anything, larger samples would just mean that even smaller effects become statistically significant and publishable, but it would not solve their lacking interpretability. Lakatos (1978) has stated that the misuse of significance testing in the social sciences makes one wonder "whether the function of statistical techniques in the social sciences is not primarily to provide a machinery for producing corroboration and thereby semblance of 'scientific progress' where, in fact, there is nothing but an increase in pseudo-intellectual garbage" (p. 88). Dar (1987) adds that education programmes focus too much on statistics and too little on philosophy of science and method. It is implied that novice researchers simply do not have the tools to know whether a theory is actually worth anything, apart from whether a test can be devised that elicits $p < 0.05$ on some metric relevant to said theory.

To wit, these reservations are not new. The concept of NHST may have developed in the second half of the twentieth century, but the dearth of progress in psychological science has been criticized many times before. In fact, crisis literature is so distinct across all of psychology's history that it may as well be considered its own *genre* (Wieser, 2016). Declarations of psychological science being in crisis go back to the late nineteenth and early twentieth century (e.g., Bühler, 1926; Driesch, 1925; Kostyleff, 1911; Line, 1931; Willy, 1899), and the field does not seem to have moved away from this declaration ever since (Sturm & Mühlberger, 2012). The crisis of the early 1900s can perhaps best be described as a period of philosophical turmoil between humanists, descriptivists, associationists and others, from which emerged several opposing paradigmatic branches of psychology; e.g., behaviourism, Gestalt psychology, psychoanalysis, *et cetera* (Radzikhovskii, 1992; Sturm & Mühlberger, 2012; see also Tonneau, 2011). However, these philosophical troubles were then already viewed as but strawmen of the real problem. For example, at the time, Line (1931) argued quite despondently that "[psychologists] appear to glory in furthering the cause of individual systems, each interested in disposing of other—and, therefore, less insightful—points of view, rather than searching for the fundamental similarities and truths in apparently diverse doctrines" (p. 495). Note that this state of affairs is reminiscent of the current need for epistemic diversity (Devezer et al., 2019). Also reminiscent of contemporary problems is Kostyleff's (1911) criticism that psychology lacks "system in its methods and objects of research, presenting too much variety and planlessness in its investigations" (as cited by Buchner, 1912, p. 2). Both criticisms are as true today as they presumably were back then: systematic methodology, as well as philosophy of scientific psychology were lacking. Talks about crisis grew smaller from the 1930s onward, though occasionally they would briefly resurface (e.g., Cronbach, 1957; Eissler, 1950; Koch, 1951; see Sturm & Mühlberger, 2012), but in the 1960s and 1970s, the notion of crisis was gradually rekindled (e.g., Bakan, 1967; Deutsch, 1976; Elms, 1975; Lewin, 1977; Morrison & Henkel, 1969; Ring, 1967; see Lakens, 2023). This second major time of crisis was particularly violent within social psychology. It would be wrong to state that it was not typified by philosophical qualms at all, but it must be noted that the emphasis was often more on methodological and practical problems than ontological ones. Nederhof & Zwier (1983) provide some "highlights" of the crisis literature at the time, and it is, again, interesting to see that most of these issues find some translation in modern-day crisis literature and most of the topics discussed in the current thesis:

> "Some of the more controversial issues focus on the social psychological experiment. They are objections concerning the ethical abuse of the laboratory situation […], the contrivedness of the experimental situation […], and the usefulness and applicability of specific statistical concepts such as the Null hypothesis […]. Moreover […], the role that social psychology plays, or should play, in contemporary society has been intensely debated […] and some have argued that social psychological research is socially irrelevant […] or of little use […]. At a theoretical level, concern has also been shown over the non-integration of the various levels of analysis […], the problematic nature of theory in social psychology […], or even the lack of theory […]." (p. 256).

Again, several elements seem to return: statistical procedure is questioned, research methodology is viewed as contrived, the objects of analysis are non-integrable and theory is problematic (for specific references, see Nederhof & Zwier, 1983). The authors go on to mention other issues, such as the "philosophical naiveté" of psychological science practitioners, who seem to have a misguided and generally euphemistic view of the history of their field, its inner fragmentation and its lacking connections to other disciplines. In summary, this shows how *the crisis* throughout psychological science's history is made up of an assortment of different issues, causing turmoil by interacting in complex ways, thus creating crises of identity, paradigm and confidence at once (e.g., Elms, 1975; Mills, 1979). The feeling of crisis has not died off since the 1980s (Nederhof & Zwier, 1983; also Epstein, 1980). In the years that followed, discussions have remained quite active (e.g., Altman, 1987; Bakan, 1996; Giorgi, 1987; Kim, 1999; Radzikhovskii, 1992; Salzinger, 1996; Thompson, 2004; see also Parker, 2007). The crisis of the twenty-first century is, much like its predecessors, exceedingly multifaceted. However, most of the discussions that are being held in the literature of today, have known precedents in the crisis literature of the 1970s and before. Recently, Lakens (2023) has provided an extensive overview of some of the most glaring similarities, the most prominent of which are "replicability of findings, the strength of theories, the societal relevance of research, the generalizability of effects, and problematic methodological and statistical practices" (p. 2).

Based on these historical facts and similarities, one cannot but abduce that research psychologists are either exceptionally bad at identifying the problems that plague their field, or at adhering to solutions that have been proposed. Given that enough literature exists that clearly defines the problems at hand, and given that more than enough literature exists on how to tackle these problems, it is likely that research psychologists, as a social group, are extremely stubborn learners. The current reform movement exemplifies this stubbornness somewhat, in that its constituents, as was extensively argued in the introduction to the current thesis, have actively misidentified the nature of the crisis as one of mere replication, and have subsequently failed to address the actual core issues laid out throughout the current thesis. The only positive thing that has come from the reform movement is the rich literature on everything researchers can do wrong to inflate type I error rates—i.e., QRPs—and how to cure the discipline of those particular ailments. However, these are but peripheral issues in the grand scheme of things. They are emergent symptoms of a discipline that actively disallows its own evolving into a proper science. That is to say, either the realized 'solutions' address an issue relatively successfully (e.g., preregistration and registered reports to solve issues pertaining to QRP engagement), but fail to address a deeper, more fundamental problem, or realized 'solutions' are based on a complete failure to recognize fundamental problems and lead to nothing but more ritualism, proceduralization and bureaucratization of psychological science. The former is somewhat admirable, all things considered, but the latter is hardly differentiable from ignorance. To give an extreme example, recently, an article was published in which it is argued that psychological research outlets should get rid of discussion sections altogether, because they "allow for an inappropriate narrativization of research that disguises actual results and enables the misstatement of true limitations" (Schoenegger & Pils, 2023, p. 1). Solutions like these do not solve any *real* problem; if anything, they may further obfuscate the nature and pervasiveness of these 'inappropriate narrativizations and misstatements'. The idea that merely increasing sample sizes would somehow improve the state of psychological science is equally misguided. In fact, forcing researchers to conduct statistical power analyses is equally misguided, because it is almost guaranteed that researchers will rely in ever increasing numbers on ill-conceived rules of thumb for what constitutes a 'small', 'medium', or 'large' effect; whatever those verbal descriptors may really mean. As Smaldino (2019) argues: "Much digital ink has been spilt describing ways to improve replicability in science. Preregistration. Open data. Open code. These are all necessary, but insufficient". The idea that psychological science can be 'saved' merely by focusing on the merit of a procedure, from standardized, cookbook methods to an obliteration of the power of rational speculation, is severely irrational. On social media and in blogs, individuals put forward 'solutions' such as the automatization of abstracts, the negation of narrative storytelling, the dropping of introduction or discussion sections, increased ritualism, proceduralization and bureaucracy via mandatory preregistration, as if the presence of qualitative discourse, methodological diversity, and messiness in scientific practice are core problems. *They are not*. The only real and fundamental problem is that psychological science deals in ambiguity and has adopted and entrenched a scientific practice which disallows corroborating and refuting any of its major theories because they are made up of immaterial, intangible nebulae of vague terms and even vaguer deductive relations to the empirical relative. All other problems either grow directly from this theoretic lacuna, or serve to cover it up.

The only real and systemic solution is translating verbal theories as they currently exist into formally rigorous models. Such models, mathematical or computational, of complex processes are normal for mature sciences (Smaldino, 2020). But what is meant by this statement that "verbal theories ought to be translated into formally rigorous models"? Smaldino (2020), van Rooij and Blokpoel (2020) and van Rooij and Baggio (2021) provide extensive and understandable overviews of how this may be achieved. What is needed first and foremost is a sharp delineation of the explanandum of interest—a *capacity*. The need for this was exemplified earlier: if explananda are not clearly delineated, given current practices in psychological theory formation, it is well possible that at some point in the scientific process, a circular rhetoric comes about due to a conflation with an explanans. When a capacity-as-explanandum is clearly delineated, a verbal theory that is based on first intuitions can be used to draw a superficial sketch of how the relations verbalized in said theory may be transformed to fit the formative restrictions of a computational-level theory (Marr, 1982). What this concretely means, is that a verbal theory must be restated as a functional relationship between an input and an output, the latter of which is a specification of the explanandum. Note that the exact computations between input and output need not be specified at this stage; a sketch may be reasonably *rough*. However, what *is* essential, as van Rooij and Baggio (2021) explain, is that not just any initial sketch is accepted. A sketch of a computational translation of a verbal theory needs to be attuned to one's initial intuitions, but they require most of all a complete absence of informality, the goal of which is to inhibit ambiguity from infecting our thinking about an explanandum. A sketch must have "all the requisite properties and no undesirable properties (e.g., inconsistencies)" and if need be, either the sketch, one's intuitions, or both need revision (van Rooij & Baggio, 2021, p. 686). What is important during this process is that both the input and the output of a computational-level theory are sufficiently specific. Smaldino (2020) provides the following illustrative example: if one is asked to model a contained system where people enter and leave an elevator that travels between floors of a building, there is an innumerable number of ways one could do so given the absolute absence of any kind of direction in the modelling question; that is, there is no question in this example, only a system. Computational modelling, even if only at this preliminary stage of *sketching*, requires a clear referent problem. A system needs to be decomposed into relevant parts, and the nature of the decomposition and what is eventually focused on is determined by the nature of the question being asked (Smaldino, 2020).

It might be tempting to immediately start finetuning a reasonable sketch at the computational level of explanation by comparing it to empirical predictions. Indeed, if a computational model is sufficiently specified, the logical next step may seem to start descending from interpretive strings onto the plane of observation (Hempel, 1952). One has just spent quite some time devising the right question to ask, revising intuitions and computational parameters and relationships—even if only at a rudimentary level—so it comes naturally to want to put one's effort to the test. However, van Rooij and Baggio (2021) argue that at this stage, diving into the nitty-gritty of empiricism may be premature. Assessing the verisimilitude of a computational-level theory first requires a theoretical cycle to be gone through, where one attempts to prove the computability and tractability of the problem as conceived by the restrictions of the computational-level theory. The reason is simple: the goal of a theory is not to merely represent our admittedly reductionist understanding of reality, but to do so in a way which is true to a certain level of verisimilitude. A useful theory is good, but a verisimilar theory is ideal. In empirical sciences, theory adjudication is a matter of finding the least *false* theory. As Tichý (1976) explains: "If [discarding a false theory in favour of another false theory] is to be meaningfully qualifiable as a step forward, it must make sense to say that the new theory [...] is closer to the truth than its discarded predecessor" (p. 25). A formally rigorous theory of a psychological construct or phenomenon needs to adhere to principles of tractability and computability if psychological science's aim is to find *truths*. If a capacity of interest is postulated to be brought forth by an algorithm that cannot provably do so, then what is the utility of such a theory? If a tractability assessment shows that the devised sketch is *a priori* impossible, then one is to return to the drawing board—not necessarily to discard the theory as a whole, but to introduce further constraints to its system (van Rooij & Baggio, 2021). If such a sketch turns out incomputable, how may it then meaningfully reflect the workings of an actual capacity?

A theoretical cycle may consist of many more different aspects of interest that may be checked beforehand, prior to empirical testing (see van Rooij & Baggio, 2021). However, once one is ready to do so, the aforementioned problems associated with the absence of formal theory are likely to disappear, or at least become far less intrusive in the process of theory construction and adjudication. The goal of empirical is to subject a substantive theory to tests which, if met, can strengthen our belief concerning the verisimilitude of said theory. The more *risky* such a test—i.e., the more the negative outcome of a test would be detrimental to the tenability of a substantive theory—the stronger it is deemed to be. This is what Salmon (1984) alternatively defines as follows: the positive outcome of such a test would constitute a "damn strange coincidence" if the substantive theory were *false*. As was outlined before, the ideal scenario—if one is of a frequentist persuasion—is to define a specific null and alternative hypothesis, the difference between which constitutes the ES of interest. A formal and sufficiently mechanistic computational-level theory allows quite naturally to delineate and specify empirical relatives which ought to be observed; to use Hempel's (1952) metaphor once more: such theories enable quite automatically the development of an interpretive string (or, alternatively, a *derivation chain* [Meehl, 1978, 1990]) between the theoretical plane and the observational one. A series of probable scenarios can thus be appreciated, which would either corroborate the verisimilitude of the posed theory, or lead to its logical refutation *modus tollens*. Subsequently, a researcher is to actually construct an experiment, assemble their variables of interest, subjects, instruments, and carry out the research. As Lakatos (1978) essentially argued, the strings connecting the theoretical plane to its empirical relative are composed of auxiliary assumptions and hypotheses, which should ideally link formally to the core theory at play, and it is most often, if not only those auxiliaries that will be attacked *modus tollens*, or corroborated. The riskier the test, the higher up in the theoretical plane its consequences will carry, in that riskier tests involve more corollaries of the posed theory.

Multiple problems have been identified with current use of ES metrics and their application in statistical power analysis. It is clear that a sufficiently formal and mechanistic theory of the sort described above resolves those issues quite naturally. Remember that current NHST practices are necessary because the conceptual paucity of psychological theories simply do not allow to conceivably and reasonably postulate empirical realities that would logically follow from a posed theory, and which circumstances would be highly unlikely, if not impossible in the same circumstances. A formal and mechanistic theory allows one, through rational considerations and by omitting ambiguity altogether, to follow the logical corollaries and consequences of a central theory to its empirical postulate; one is able to simply descend, using the rules of deductive reasoning, from platonic to empirical reality and formulate *what ought to be observed* in a systematic fashion, and explicitly *not* via verbal games. Logical inconsistencies with these descending derivation chains then constitute candidates which, if they were to be observed, will lead to the inevitable conclusion that the posed theory is implausible and in need of correction, revision, or refutation—depending on the severity of the test and the magnitude of the inconsistency between empirical reality and theoretical supposition it has unveiled. Whereas current NHST practices disallow researchers to formulate any such thing to formulate adjudications about, formal and mechanistic theories circumvent the shaky foundations of ambiguous verbal theories and actually allow one to extract real consequence from the empirical cycle. The nature of the inconsistency between refutative and corroboratory empirical realities, translated in terms of the observable units of a metric of interest, then simply *dictates* what would be a reasonable smallest ES of interest. Thus, a statistical power analysis can actually be conducted in an informative manner.

Moreover, the whole ordeal of working with standardized ES metrics may become less of a hassle and more systematic. Remember that the main problem was that inconsistencies of empirical circumstances between investigations were impossible to assess due to both being clouded by verbal ambiguities; there was simply no way to systematically assess the commensurability of circumstances and outcomes, such that standardized ES metrics which emerged from them and that were imported from other studies to one's own statistical power analysis are completely uninformative. The goal of an analysis thus became necessarily reduced to a comparison of the ratio between true ES and the reliability and validity—read: precision—of an arbitrary and intrinsically meaningless scale, based entirely on the unfounded *assumption* that the ES metric somehow transcends the empirical circumstantial constraints of either investigation. When theory is formal and mechanistic, and so are its auxiliaries, these problems become less intrusive. The measure of a scale is itself an auxiliary, and if one is consistent in their formal conceptualization of all auxiliaries, one can—theoretically speaking—reason one's way through the nomological network that connects them all. That is to say, the empirical circumstances of one study can be traced back to their theoretical postulates and auxiliaries, and may thence be derived toward the empirical circumstances of the current study, such that one can at least logically formulate, if not quantify the commensurability of two empirical investigations. The magical transcendence of the standardized ES metrics becomes thus demystified, because there is far less ambiguity standing between two or multiple studies. In fact, one can reorient one's own research to better fit the formal constraints of another, such that commensurability can at least be logically defended. This discursive process is what makes ES comparisons meaningful. Of course, it will always be the case, especially in social sciences, that substantial elements of an empirical relative cannot be entirely disambiguated; it would be a Herculean task. But a formal and mechanistic *and internally consistent* unambiguous central theory will at least move the field forward in such an idealized direction as the one described above.

By a generally similar procedure, NHST practices will no longer be necessary to obfuscate the imprecisions of most psychological theory. The cascade of *ad hoc* auxiliaries will become untenable; the disconnect between their ambiguity and the clarity of a formal and mechanistic theory would surely elicit too insurmountable a degree of cognitive dissonance. Meehl's (1967) logical nightmare (see quote above) would no longer present itself, for *ad hoc* auxiliaries would not be easily fitted inside the formally robust theory. Equally, the destabilization of an auxiliary, such as questionable survey validity for a depression scale, would have consequences for the central theory, for the survey would be formally derived from and intimately connected to it. It would necessitate a return to van Rooij and Baggio's (2021) drawing board; an off-hand remark on questionnaire validity or 'demand characteristics' (Dar, 1987) would no longer be sufficient—if they ever were. If the focus is taken away from establishing *any effect*—which encourages QRPs such as *p*-hacking and HARKing—toward *an ES of interest* that is formally and/or mechanistically derivable from disambiguated verbal theories, Neyman-Pearson significance testing may be conducted the way it was meant to be. In fact, the engagement in QRPs may sometimes even become impossible, for formal approaches can lead to veritable knowledge without the need for inductive statistics, but simply by fitting formal models to observed data and noticing that significant parts of the data seem to escape the clutches of the formal constraints. Finally, conceptual replications might become more of a possibility in psychological science. If theoretical postulates are formalized and sufficiently delineated, the difference with an alternative theory that is substantiated by a whole different array of such postulates, auxiliaries, connections and rational assumptions could be devised to logically imply an equivalent empirical postulate as the one described by the first such theory. To reiterate from the introduction: there is a difference in inductive strength that lies in replicating an experiment based on an entirely different theoretical construction, such that one may move beyond the inductive restrictions posed by a singular approach. As such, a truly coherentist approach to scientific practice is actively possible.

In summary, formalization of verbal theories comes with several benefits, and many have called upon the field to finally start adopting the rich toolkit that formalization practices offer (e.g., Guest & Martin, 2021; Robinaugh et al., 2021; Smaldino, 2020; van Rooij & Blokpoel, 2020; van Rooij & Baggio, 2021). The main purpose of formalization is disambiguation of verbal theories, which allow to obfuscate inconsistencies that would normally lead to a refutation of theory, its auxiliaries, or both. Formalization aids in making ES values meaningful and commensurate across studies. It omits the necessity for hybridized and poor statistics, it potentiates replications across the spectrum of causal independence—for causal independence can actually be achieved *and* quantified. Only if formal theorization is not just normalized, but incentivized and encouraged will the slow increase of statistical power analyses as shown in the herein included systematic review bear no poisoned fruits. It is either that, or continue to trudge along the peripheries of symptoms and strawman nonproblems; it is either that, or continue to remain a field in search for something to fill its epistemic void.

Conclusion

As with most—if not all—movements, a persistent group of sceptics ceases not to voice doubts concerning the severity of the seemingly endless stream of crises in the history of psychology; in fact, they not so much view it as crises, but rather as a “set of tractable problems localized in specific subareas or [as] a result of the sensationalism incited by disaffected researchers” (Morawski, 2019, p. 220). At its core, crisis sceptics argue that the apparent non-replicability of certain findings reported in the literature should not be viewed as symptomatic of a larger, domain-wide, systemic issue, but as resulting from local, explicitly delineable methodological deficits in arguably niche subdisciplines. The contents of the current thesis have spoken to the opposite of this claim. The only resemblance it bears to the sceptics is to the notion that the problem is delineable; though it is by no means local or niche.

The current crisis in psychological science was kindled by multiple observations of non-replicability concerning several supposedly canonical findings. Coordinated replicatory efforts have only exacerbated the necessary consequence of this observation: psychological science lacks a robust knowledge base; in fact, it may be said to lack any cumulative character at all. The reform movement that followed has focused primarily on ridding the field of questionable research practices, which are argued to cause somewhat of an infestation of false positives in the published literature. Curative devices are therefore mainly oriented towards minimizing type I error rates; or, at least, towards keeping their frequency at the set alpha level of statistical significance. Unfortunately, this preoccupation has obfuscated, much like curative efforts of purported crises in the past, the real problem. The reform movement's initial critique toward false positives has spiralled out of control, into a dictum that any given investigation must be as replicable as possible. The crisis as seen from the perspective of the reform movement is one of mere replication; and, in fact, one of mere *close* replication, as is exemplified by its insistence on keeping replications as close as possible to original designs on the premise that the consequences of a failed replication that is not close are null and void.

The issue of statistical power is one mostly ignored by the current reform movement. It is only in recent years that psychology's lack of statistical power has been gaining attention, despite the issue having been identified already in the 1960s. A historical negligence toward the subject is changing, and researchers are slowly becoming aware of the severe consequences that are attached to a failure to take into account statistical power for any given investigation. Publishers are starting to require such analyses be included in reports, more emphasis is being placed on sample size justification, knowledgeable authors are providing the field of power primers which serve to aid those who are uninitiated. Recent surveys have shown that, indeed, some improvement seems underway, but clear evidence is nonetheless lacking. To provide more such evidence, the current thesis has committed to a systematic review of three psychological subdisciplines, six journals, and two publication years. It was revealed that, indeed, the inclusion of statistical power analysis reports seems generally on the rise, but absolute numbers remain unacceptably low. Moreover, several reporting practices reveal that statistical power analysis as it is mostly conducted seems to have several problems inherent to it. The most prominent of these is the fact that effect sizes of interest are rarely clearly defined or interpreted, while this aspect is arguably the most important one. Researchers almost always rely on a standardized metric and an agreed upon description of which standardized effect is 'small', 'medium' or 'large', or they import an effect size that is reported in a previous investigation. However, this practice is misguided, because effect sizes across studies are incommensurate, whether they are standardized or not. Surprisingly, instead of tackling the issue of how, then, to device meaningful and interpretable effect sizes, prominent at the forefront of the reform movement has come about the notion that sample sizes should just be increased to solve issues pertaining to statistical power. Using simulations and rational considerations, the tenability of this ‘solution’ was shown to be questionable. Furthermore, it was argued that psychological science is caught up in its own web of ritualistic procedures, at the centre of which is its attempt at statistical inference—a poor conception of Neyman-Pearson significance testing. Exemplificatory to this effect are the tendencies of research psychologists toward ill-conceived rules of thumb, their generally poor education on topics related to statistics and philosophy of science, and engrained traditions.

In the final segments of the current thesis, it was argued that at the core of all of the above lies the ambiguity that is inherent to how psychological theories are traditionally constructed. Psychologists’ verbal proclivities obscure the inherent inconsistencies of their theories, and the informality with which they are conceived allows neither to refute nor to corroborate hypotheses derived from them. Psychological science’s traditional ways obstruct any meaningful cumulation of verisimilar knowledge, and, instead, theories merely come and go, and are forgotten. A case was made to rebuild psychology from the ground up, not by ridding the field of its theories, but by formalizing their semantics, such that strong and *testable* theoretical frameworks may actually come about. It was briefly explained what this exactly entails, and it was argued that if psychological theories are formalized and mechanistically conceived, the problem of devising meaningful effect sizes becomes much more tractable. Incentives for QRP engagement are likely to dissipate with it, as well as bad statistics, *if and only if* strong and systematic adherence to principles of formality is achieved and sustained.

If the field of psychological science is truly to break free of its stubborn adherence to ill-conceived practices, it is imperative not only to normalize formalizing verbal theories, but to necessitate it. By obviating the need for compromised statistical approaches, rules of thumb, *et cetera*, formalization enables the field to become truly cumulative; replications are then enabled to span the entire spectrum of causal independence, because the latter will be attainable and quantifiable. Neglecting this imperative could render the slow and incremental improvements that we are seeing in the mere reporting of having conducted *a priori* statistical power analyses, unproductive. Either this transformative course is embraced, or the field shall further wither away in the peripheries of the problem, or worse, strawman nonproblems—losing itself in symptomatic investigations and solutions that bear no fruit. To put it theatrically: the field is at a juncture, and this juncture is represented by a pivotal choice. Either the field advances towards substantial resolutions, or it perpetuates its quest to fill an epistemic void with ambiguous verbal theories that have no importance, no continuity, and no verisimilitude. It is up to the researchers which of these two options they would like to pursue.

BIBLIOGRAPHY

Abraham, W. T., & Russell, D. W. (2008). Statistical power analysis in psychological research. *Social and Personality Psychology Compass: 2*(1), 283-301. https://doi.org/10.1111/j.1751-9004.2007.00052.x

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger, H. L., Scarr, S., Kazdin, A. E., & Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45*(6), 721-734. https://doi.org/10.1037/0003-066X.45.6.721

Aldhous, P. (2011, May 5). Journal rejects studies contradicting precognition. *NewScientist.* https://www.newscientist.com/article/dn20447-journal-rejects-studies-contradicting-precognition/

Alifieris, C. E., Souferi-Chronopoulos, R., Trafalis, D. T., & Arvelakis, A. (2020). The arbitrary magic of p<0.05: Beyond statistics. *Journal of the Balkan Union of Oncology, 25*(2), 588-593. Retrieved from: https://jbuon.com/25-2/

Altman, I. (1987). Community psychology twenty years later: Still another crisis in psychology? *American Journal of Community Psychology, 15*(5), 613-627. https://doi.org/10.1007/BF00929914

American Psychological Association, Committee for the Protection of Human Participants in Research. (1982). *Ethical principles and the conduct of research with human participants.* American Psychological Association. https://doi.org/10.1037/10084-000

Angell, M. (1989). Negative studies. *The New England Journal of Medicine, 321*(7), 464-466. https://doi.org/10.1056/NEJM198908173210708

Anvari, F., & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology, 3*(3), 266-286. https://doi.org/10.1080/23743603.2019.1684822

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist, 73*(1), 3-25. https://psycnet.apa.org/doi/10.1037/amp0000389

Ardila, R. (2007). The nature of psychology: The great dilemmas. *American Psychologist, 62*(8), 906–912. https://doi.org/10.1037/0003-066X.62.8.906

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods, 24*(1), 1-19. https://doi.org/10.1037/met0000195

Artino, A. R., Driessen, E. W., & Maggio, L. A. (2019). Ethical shades of gray: International frequency of scientific misconduct and questionable research practices in health professions education. *Academic Medicine, 94*(1), 76-84. https://doi.org/10.1097/ACM.0000000000002412

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Vanaken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*(2), 108-119. https://doi.org/10.1002/per.1919

Bajwa, N. u. H., & König, C. J. (2019). How much is research in the top journals of industrial/organizational psychology dominated by authors from the U.S.?. *Scientometrics, 120*, 1147-1161. https://doi.org/10.1007/s11192-019-03180-2

Baguley, T. S. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*(3), 603-617. https://doi.org/10.1348/000712608X377117

Baguley, T. S. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences.* Palgrave Macmillan.

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

Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation*. Jossey-Bass.

Bakan, D. (1996). The crisis in psychology. *Journal of Social Distress and the Homeless, 5*(4), 335-342. https://doi.org/10.1007/BF02092909

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science, 27*(8), 1069-1077. https://doi.org/10.1177/0956797616647519

Bakker, M., Veldkamp, C. L. S., van den Akker, O. R., van Assen, M. A. L. M., Crompvoets, E., Ong, H. H., & Wicherts, J. M. (2020). Recommendations in pre-registrations and internal review board proposals promote formal power analyses but do not increase sample size. *PLoS ONE, 15*(7), e0236079. https://doi.org/10.1371/journal.pone.0236079

Bakker, M., & Wicherts, J. M. (2014). Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS ONE, 9*(7), e103360. https://doi.org/10.1371/journal.pone.0103360

Bamberger, P. A. (2019). From the editor: On the replicability of abductive research in management and organizations: Internal replication and its alternatives. *Academy of Management Discoveries, 5*(2), 103-108. https://doi.org/10.5465/amd.2019.0121

Banks, G. C., Rogelberg, S. G., Woznyi, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology, 31*(3), 323-338. https://doi.org/10.1007/s10869-016-9456-7

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*(2), 230-244. https://doi.org/10.1037/0022-3514.71.2.230

Barr, A. S. (1932). A study of the amount of agreement found in the results of four experimenters employing the same experimental technique in a study of the effects of visual and auditory stimulation on learning. *Journal of Educational Research, 26*(1), 35-45. URL: https://www.jstor.org/stable/27525567

Bartoš, F., & Maier, M. (2022). Power or alpha? The better way of decreasing the false discovery rate. *Meta-Psychology, 6*, MP.2020.2460. https://doi.org/10.15626/mp.2020.2460

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology, 74*(5), 1252-1265. https://doi.org/10.1037/0022-3514.74.5.1252

Baumgaertner, B., Devezer, B., Buzbas, E. O., & Nardin, L. G. (2019). Openness and reproducibility: Insights from a model-centric approach. *arXiv*. https://doi.org/10.48550/arXiv.1811.04525

Baxter, B. (1940). The application of factorial design to a psychological problem. *Psychological Review, 47*(6), 494–500. https://doi.org/10.1037/h0055537

Begeny, J. C., Levy, R. A., Hida, R., Norwalk, K., Field, S., Suzuki, H., Soriano-Ferrer, M., Scheunemann, A., Guerrant, M., Clinton, A., & Burneo, C. A. (2018). Geographically representative scholarship and internationalization in school and educational psychology: A bibliometric analysis of eight journals from 2002-2016. *Journal of School Psychology, 70*, 44-63. https://doi.org/10.1016/j.jsp.2018.07.001

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), 407-425. https://doi.org/10.1037/a0021524

Bem, D. J., & Honorton, C. (1994). Does psi exist? Evidence for an anomalous process of information transfer. *Psychological Bulletin, 115*(1), 4-18. https://doi.org/10.1037/0033-2909.115.1.4

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., …, Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6-10. https://doi.org/10.1038/s41562-017-0189-z

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology, 23*(3), 399-406. https://doi.org/10.1076/jcen.23.3.399.1181

Bolles, R. C. (1962). The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports, 11*(3), 639-645. https://doi.org/10.2466/pr0.1962.11.3.639

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57*(1), 49-64. https://doi.org/10.1037/h0041412

Borenstein, M. (2022). In a meta-analysis, the I-squared statistic does not tell us how much the effect size varies. *Journal of Clinical Epidemiology, 152*, 281-284. https://doi.org/10.1016/j.jclinepi.2022.10.003

Bowlby, J. (1984). Psychoanalysis as an natural science. *Psychoanalytic Psychology, 1*(1), 7-21. https://doi.org/10.1037/0736-9735.1.1.7

Boyce, V., Mathur, M., & Frank, M. C. (2023). Eleven years of student replication projects provide evidence on the correlates of replicability in psychology. *PsyArXiv.* https://doi.org/10.31234/osf.io/dpyn6

Brand, A., Bradley, M. T., Best, L. A., & Stoica, G. (2008). Accuracy of effect size estimates from published psychological research. *Perceptual and Motor Skills, 106*(2), 645-649. https://doi.org/10.2466/PMS.106.2.645-649

Brewer, J. K. (1972). On the power of statistical tests in the "American Educational Research Journal". *American Educational Research Journal, 9*(3), 391-401. https://doi.org/10.2307/1161755

Brodeur, A., Cook, N. M., Hartley, J. S., & Heyes, A. (2022). Do pre-registration and pre-analysis plans reduce p-hacking and publication bias. *MetaArXiv.* https://doi.org/10.31222/osf.io/uxf39

Broers, N. J., & Otgaar, H. (2021). Toward a simplified justification of the chosen sample. *PsyArXiv*. https://doi.org/10.31234/osf.io/kmx75

Brunner, J., & Schimmack, U. (2020). Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology, 4*, MP.2018.874. https://doi.org/10.15626/MP.2018.874

Brydges, C. R. (2018). Evaluation of publication bias and statistical power in gerontologial psychology. *PsyArXiv.* https://doi.org/10.31234/osf.io/ruwxt

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1*(1), 9. https://doi.org/10.5334/joc.10

Buchner, E. F. (1912). Psychological progress in 1911. *Psychological Bulletin, 9*(1), 1-10. https://doi.org/10.1037/h0072778

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

Buzbas, E. O., Devezer, B., & Baumgaertner, B. (2023). The logical structure of experiments lays the foundation for a theory of reproducibility. *Royal Society Open Science, 10*, 221042. https://doi.org/10.1098/rsos.221042

Bühler, K. (1926). Die Krise der Psychologie. *Kant-Studien, 31*, 455-526. https://doi.org/10.1515/kant.1926.31.1-3.455

Callaway, E. (2011). Report finds massive fraud at Dutch universities. *Nature, 479*(7371), 15. https://doi.org/10.1038/479015a

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

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science, 21*(10), 1363-1368. https://doi.org/10.1177/0956797610383437

Cartwright, N. (1991). Replicability, reproducibility, and robustness: Comments on Harry Collins. *History of Political Economy, 23*(1), 143-155. URL: https://EconPapers.repec.org/RePEc:hop:hopeec:v:23:y:1991:i:1:p:143-155

Cascio, W. F., Valenzi, E. R., & Silbey, V. (1978). Validation and statistical power: Implications for applied research. *Journal of Applied Psychology, 63*(5), 589-595. https://doi.org/10.1037/0021-9010.63.5.589

Case, C. M. (1928). Scholarship in sociology. *Sociology and Social Research, 12*(4), 323-340.

Cesana, B. M. (2018). What p-value must be used as the statistical significance threshold? P<0.005, P<0.01, P<0.05 or no value at all? *Biomedical Journal of Scientific & Technical Research, 6*(3), MS.ID.001359. https://doi.org/10.26717/BJSTR.2018.06.001359

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Champely, S. (2020). pwr: Basic functions for power analysis. R package version 1.3-0. URL: https://CRAN.R-project.org/package=pwr

Chang, H. (2004). *Inventing temperatue: Measurement and scientific progress.* Oxford University Press. https://doi.org/10.1093/0195171276.001.0001

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*, 1-8. https://doi.org/10.1016/j.jebo.2011.08.009

Chase, L. J., & Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology, 61*(2), 234-237. https://doi.org/10.1037/0021-9010.61.2.234

Chase, L. J., & Tucker, R. K. (1976). Statistical power: Derivation, development, and data-analytic implications. *The Psychological Record, 26*, 473-486. https://doi.org/10.1007/BF03394413

Clark, R. G. (2009). Sampling of subpopulation in two-stage surveys. *Statistics in Medicine, 28*, 3697-3717. https://doi.org/10.1002/sim.3723

Chopik, W. J., Bremner, R. H., Defever, A. M., & Keller, V. N. (2018). How (and whether) to teach undergraduates about the replication crisis in psychological science. *Teaching of Psychology, 45*(2), 158-163. https://doi.org/10.1177/0098628318762900

Cochrane, R., & Duffy, J. (1974). Psychology and scientific method. *Bulletin of the British Psychological Society, 27*(95), 117-121. URL: https://psycnet.apa.org/record/1975-25359-001

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

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Academic Press.

Cohen, J. (1973). Brief notes: Statistical power analysis and research results. *American Educational Research Journal, 10*(3), 225-230. https://doi.org/10.3102/00028312010003225

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cohen, J. (1990). What I have learned (so far). *American Psychologist, 45*(12), 1304-1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. https://doi.org/10.1037/0033-2909.112.1.155

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

Colling, L. J., & Szűcs, D. (2021). Statistical inference and the replication crisis. *Review of Philosophy and Psychology, 12*, 121-147. https://doi.org/10.1007/s13164-018-0421-4

Collins, E., & Watt, R. (2021). Using and understanding power in psychological research: A survey study. *Collabra: Psychology, 7*(1), 28250. https://doi.org/10.1525/collabra.28250

Cooper, H., & Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. *Personality and Social Psychology Bulletin, 8*(1), 168-173. https://doi.org/10.1177/014616728281026

Copeland, M. A. (1930). Psychology and the natural-science point of view. *The Psychological Review, 37*(6), 461-487. https://doi.org/10.1037/h0072054

Coyne, J. C. (2009). Are most positive findings in health psychology false…. or at least somewhat exaggerated? *The European Health Psychologist, 11*, 49-51. URL: https://www.ehps.net/ehp/index.php/contents/article/view/ehp.v11.i3.p49

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology, 66*, 93-99. https://doi.org/10.1016/j.jesp.2015.10.002

Cronbach, L. J. (1957). The two disciplines of scientific psychology. In E. R. Hilgard (Ed.), *American psychology in historical perspective* (pp. 435-458). American Psychological Association. https://doi.org/10.1037/10049-022

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. https://doi.org/10.1037/h0040957

Cuddy, A. J. C., Schultz, S. J., & Fosse, N. E. (2018). *P*-curving a more comprehensive body of research on postural feedback reveals clear evidential value for power-posing effects: Reply to Simmons and Simonsohn (2017). *Psychological Science, 29*(6), 656-666. https://doi.org/10.1177/0956797617746749

Cummins, R. (2000). “How does it work?” vs. “What are the laws?”: Two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 117-145). MIT Press.

Danziger, K. (1985). The methodological imperative in psychology. *Philosophy of the Social Sciences, 15*(1), 1-13. https://doi.org/10.1177/004839318501500101

Danziger, K. (1990). *Constructing the subject.* Cambridge University Press.

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist, 42*(2), 145-151. https://doi.org/10.1037/0003-066X.42.2.145

De Houwer, J. (2011). Why the cognitive approach in psychology would profit from a functional approach and vice versa. *Perspectives on Psychological Science, 6*(2), 202-209. https://doi.org/10.1177/1745691611400238

Derksen, M., & Morawski, J. (2022). Kinds of replication: Examining the meanings of "conceptual replication" and "direct replication". *Perspectives on Psychological Science, 17*(5), 1490-1505. https://doi.org/10.1177/17456916211041116

De Rond, M., & Miller, A. N. (2005). Publish or perish: Bane or boon of academic life? *Journal of Management Inquiry, 14*(4), 321-329. https://doi.org/10.1177/1056492605276850

Deutsch, M. (1976). Theorizing in social psychology. *Personality and Social Psychology Bulletin, 2*(2), 134-141. https://doi.org/10.1177/014616727600200214

Devezer, B., & Buzbas, E. O. (2022). Minimum viable experiment to replicate. *PhilSci Archive.* URL: http://philsci-archive.pitt.edu/id/eprint/21475

Devezer, B., Nardin, L. G., Baumgaertner, B., & Burbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLoS ONE, 14*(5), e0216125. https://doi.org/10.1371/journal.pone.0216125

de Vries, Y. A., Schoevers, R. A., Higgins, J. P. T., Munafò, M. R., & Bastiaansen, J. A. (2022). Statistical power in clinical trials of interventions for mood, anxiety, and psychotic disorders. *Psychological Medicine, 53*(10), 4499-4506. https://doi.org/10.1017/S0033291722001362

Dingledine, R. (2018). Why is it so hard to do good science? *eNeuro, 5*(5), e0188-18.2018. https://doi.org/10.1523/ENEURO.0188-18.2018

Doke, S. K., & Dhawale, S. C. (2013). Alternatives to animal testing: A review. *Saudi Pharmaceutical Journal, 23*(3), 223-229. https://doi.org/10.1016/j.jsps.2013.11.002

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

Driesch, H. (1925). *The crisis in psychology.* Princeton University Press.

Dukes, W. F. (1965). N = 1. *Psychological Bulletin, 64*(1), 74-79. https://doi.org/10.1007/978-3-319-24612-3_301651

Dunlap, K. (1925). The experimental methods of psychology. *The Pedagogical Seminary and Journal of Genetic Psychology, 32*(3), 502-522. https://doi.org/10.1080/08856559.1925.10532333

Dunlap, J. W. (1933). Comparable tests and reliability. *Journal of Educational Psychology, 24*(6), 442-453. https://doi.org/10.1037/h0075324

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

Eddy, M. (2011, May 5). Controversy as journal refuses to publish studies that fail to support precognition. *The Mary Sue.* https://www.themarysue.com/precognition-studies-journal/

Eerland, A., Sherrill, A. M., Magliano, J. P., & Zwaan, R. A. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11*(1), 158-171. https://doi.org/10.1177/1745691615605826

Eissler, K. R. (1950). The Chicago Institute of Psychoanalysis and the sixth period of the development of psychoanalytic technique. *Journal of General Psychology, 43*, 103-157. https://doi.org/10.1080/00221309.1950.9920150

Elms, A. C. (1975). The crisis of confidence in social psychology. *American Psychologist, 30*(10), 967-976. https://doi.org/10.1037/0003-066X.30.10.967

Epstein, S. (1980). The stability of behavior. II. Impliciations for psychological research. *American Psychologist, 35*(9), 790-806. https://doi.org/10.1037/0003-066X.35.9.790

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*(1), 1-11. https://doi.org/10.3758/BF03203630

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science, 16*(4), 779-788. https://doi.org/10.1177/1745691620970586

Estabrooks, G. H. (1929). The enigma of telepathy. *The North American Review, 227*(2), 201-211. URL: https://www.jstor.org/stable/25110685

Ethical principles in the conduct of research with human participants. (1973). *American Psychologist, 28*(1), 79-80. https://doi.org/10.1037/h0038067

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology, 66*, 68-80. https://doi.org/10.1016/j.jesp.2015.07.009

Fagley, N. S. (1985). Applied statistical power analysis and the interpretation of nonsignificant results by research consumers. *Journal of Counseling Psychology, 32*(3), 391-396. https://doi.org/10.1037/0022-0167.32.3.391

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

Fanelli, D. (2010). “Positive” results increase down the hierarchy of the sciences. *PLoS ONE, 5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Farrell, B. A. (1978). The progress in psychology. *British Journal of Psychology, 69*(1), 1-8. https://doi.org/10.1111/j.2044-8295.1978.tb01626.x

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science, 7*(1), 45-52. https://doi.org/10.1177/1948550615612150

Feest, U. (2019). Why replication is overrated. *Philosophy of Science, 86*, 895-905. https://doi.org/10.1086/705451

Feng, C., Thompson, W. K., & Paulus, M. P. (2021). Effect sizes of associations between neuroimaging measures and affective symptoms: A meta-analysis. *Depression and Anxiety, 39*(1), 19-25. https://doi.org/10.1002/da.23215

Ferguson, C. J. (2007). Evidence for publication bias in video game violence effects literature: A meta-analytic review. *Aggression and Violent Behavior, 12*, 470-482. https://doi.org/10.1016/j.avb.2007.01.001

Fidler, F. (2005). From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology. *Thesis Commons.* https://doi.org/10.31237/osf.io/7xdpq

Finkel, E. J., & Baumeister, R. F. (2019). Social psychology: Crisis and renaissance. In R. F. Baumeister & E. J. Finkel (Eds.), *Advanced social psychology: The state of the science* (pp. 1-8). Oxford University Press.

Firebaugh, G., Warner, C., & Massoglia, M. (2013). Fixed effects, random effects, and hybrid models for causal analysis. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 113-132). Springer.

Fisher, R. A. (1933). The contributions of Rothamsted to the development of the science of statistics. *Annual Report of the Rothamsted Experimental Station,* 43-50. URL: https://hdl.handle.net/2440/15213

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological), 17*(1), 69-78. URL: https://www.jstor.org/stable/2983785

Fisher, R. A. (1958). Cancer and smoking. *Nature, 182*, 596. https://doi.org/10.1038/182596a0

Fisher, R. A. (1973). *Statistical methods and scientific inference* (3rd Ed.). Hafner.

Fishman, D. B., & Neigher, W. D. (1982). American psychology in the eighties: Who will buy? *American Psychologist, 37*(5), 533-546. https://doi.org/10.1037/0003-066X.37.5.533

Fiske, S. T. (2004). Mind the gap: In praise of informal sources of formal theory. *Personality and Social Psychology Review, 8*(2), 132-137. https://doi.org/10.1207/s15327957pspr0802_6

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245-260). Russell Sage Foundation.

Flis, I. (2019). Psychologists psychologizing scientific psychology: An epistemological reading of the replication crisis. *Theory & Psychology, 29*(2), 158-181. https://doi.org/10.1177/0959354319835322

Fox, N. W., Honeycutt, N., & Jussim, L. (2018). How many psychologists use questionable research practices? Estimating the population size of current QRP users. *PsyArXiv.* https://doi.org/10.31234/osf.io/3v7hx

Fraley, R. C., Chong, J. Y., Baacke, K. A., Greco, A. J., Guan, H., & Vazire, S. (2022). Journal n-pact factors from 2011 to 2019: Evaluating the quality of social/personality journals with respect to sample size and statistical power. *Advances in Methods and Practices in Psychological Science, 5*(4). https://doi.org/10.1177/25152459221120217

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19*, 975-991. https://doi.org/10.3758/s13423-012-0322-y

Francis, G. (2014). The frequency of excess success for articles in *Psychological Science*. *Psychological Bulletin & Review, 21*, 1180-1187. https://doi.org/10.3758/s13423-014-0601-x

Franz, D. J. (2022). "Are psychological attributes quantitative?" is not an empirical question: Conceptual confusions in the measurement debate. *Theory & Psychology, 32*(1), 131-150. https://doi.org/10.1177/09593543211045340

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS ONE, 13*(7), e0200303. https://doi.org/10.1371/journal.pone.0200303

Friese, M., & Frankenbach, J. (2020). *p*-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods, 25*(4), 456-471. https://doi.org/10.1037/met0000246

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2018). Is ego depletion real? An analysis of arguments. *Personality and Social Psychology Review, 23*(2), 107-131. https://doi.org/10.1177/1088868318762183

Frith, U., & Frith, C. (2014, May 28). A question of trust: Fixing the replication crisis. *The Guardian.* URL: https://www.theguardian.com/science/occams-corner/2014/may/28/question-trust-fixing-replication-crisis-experimenter-reputation

Fritz, A., Scherndl, T., & Kühberger, A. (2012). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? *Theory & Psychology, 23*(1), 98-122. https://doi.org/10.1177/0959354312436870

Frost, J. (2019). Low power tests exaggerate effect sizes. *Statistics By Jim.* URL: https://statisticsbyjim.com/hypothesis-testing/low-power-studies/

Furchtgott, E. (1984). Replicate, again and again. *American Psychologist, 39*(11), 1315-1316. https://doi.org/10.1037/0003-066X.39.11.1315.b

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology, 103*(6), 933-948. https://doi.org/10.1037/a0029709

Gawronski, B., & Bodenhausen, G. V. (2015). Social-cognitive theories. In B. Gawronski & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 65-83). Guilford Press.

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

Giner-Sorolla, R., Montoya, A. K., Aberson, C. L., Carpenter, T., Lewis, N. A., Bostyn, D. H., Conrique, B. G., Ng, B. W., Reifman, A., Schoemann, A. M., & Soderberg, C. (2023). Power to detect what? Considerations for planning and evaluating sample size. *PsyArXiv*. https://doi.org/10.31234/osf.io/rv3kw

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology, 8*(2), 195-204. https://doi.org/10.1177/0959354398082006

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science, 1*(2), 198-218. https://doi.org/10.1177/2515245918771329

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and individual differences, 102*, 74-78. https://doi.org/10.1016/j.paid.2016.06.069

Giorgi, A. (1987). The crisis of humanistic psychology. *The Humanistic Psychologist, 15*(1), 5-20. https://doi.org/10.1080/08873267.1987.9976779

Girden, E. (1962). A review of psychokinesis (PK). *Psychological Bulletin, 39*(5), 353-388. https://doi.org/10.1037/h0048209

Goertzen, J. R. (2008). On the possibility of unification: The reality and nature of the crisis in psychology. *Theory & Psychology, 18*(6), 829-852. https://doi.org/10.1177/0959354308097260

Goldman, A. I. (1988). Strong and weak justification. *Philosophical Perspectives,* vol. 2, 51-69. https://doi.org/10.2307/2214068

Goldwater, B. C. (1972). Psychological significance of pupillary movements. *Psychological Bulletin, 77*(5), 340-355. https://doi.org/10.1037/h0032456

Gopalakrishna, G., ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLoS ONE, 17*(2), e0263023. https://doi.org/10.1371/journal.pone.0263023

Gottfredson, L., & Saklofske, D. H. (2009). Intelligence: Foundations and issues in assessment. *Canadian Psychology, 50*(3), 183-195. https://doi.org/10.1037/a0016641

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*, 493-498. https://doi.org/10.1111/2041-210X.12504

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*(1), 1-20. https://doi.org/10.1037/h0076157

Griffin, J. W. (2021). Calculating statistical power for meta-analysis using metapower. *The Quantitative Methods for Psychology, 17*(1), 24-39. https://doi.org/10.20982/tqmp.17.1.p024

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science, 16*(4), 789-802. https://doi.org/10.1177/1745691620970585

Guilford, J. P. (1956). *Fundamental statistics in psychology and education* (3rd ed.). McGraw-Hill.

Guttinger, S. (2018). Replications everywhere. *BioEssays, 40*(7), 1800055. https://doi.org/10.1002/bies.201800055

Guttinger, S. (2019). A new account of replication in the experimental life sciences. *Philosophy of Science, 86*, 453-471. https://doi.org/10.1086/703555

Guy, R. K. (1988). The strong law of small numbers. *The American Mathematician Monthly, 95*(8), 697-712. https://doi.org/10.2307/2322249

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology, 29*(1), 58-65. https://doi.org/10.1037/0022-0167.29.1.58

Hacking, I. (1965). *Logic of statistical inference.* Cambridge University Press.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546-573. https://doi.org/10.1177/1745691616652873

Haig, B. D. (2021). Understanding replication in a way that is true to science. *Review of General Psychology, 26*(2), 224-240. https://doi.org/10.1177/10892680211046514

Hansson, S. O. (2021). Science and pseudo-science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. URL: https://plato.stanford.edu/archives/fall2021/entries/pseudo-science/

Hart, W., & Albarracín, D. (2011). Learning about what others were doing: Verb aspect and attribution of mundane and criminal intent for past actions. *Psychological Science, 22*(2), 261-266. https://doi.org/10.1177/0956797610395393

Hartgerink, C. H., J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Distributions of *p*-values smaller than .05 in psychology: What is going on? *PeerJ, 4*, e1935. https://doi.org/10.7717/peerj.1935

Hartgerink, C. H. J., & Wicherts, J. M. (2016). Research practices and assessment of research misconduct. *ScienceOpen Research, 0*(0), 1-10. https://doi.org/10.14293/S2199-1006.1.SOR-SOCSCI.ARYSBI.v1

Hayasaka, S., Peiffer, A. M., Hugenschmidt, C. E., & Laurienti, P. (2007). Power and sample size calculation for neuroimaging studies by non-central random field theory. *NeuroImage, 37*, 721-730. https://doi.org/10.1016/j.neuroimage.2007.06.009

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486-504. https://doi.org/10.1037/1082-989X.3.4.486

Heinlein, P. C., & Heinlein, J. H. (1938). Critique of the premises and statistical methodology of parapsychology. *The Journal of Psychology, 5*(1), 135-148. https://doi.org/10.1080/00223980.1938.9917558

Helwegen, K., Libedinsky, I., & van den Heuvel, M. P. (2023). Statistical power in network neuroscience. *Trends in Cognitive Sciences, 27*(3), 282-301. https://doi.org/10.1016/j.tics.2022.12.011

Hempel, C. G. (1952). *Fundamentals of concept formation in empirical science*. University of Chicago Press.

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61-83. https://doi.org/10.1017/S0140525X0999152X

Hensel, W. M. (2020). Double trouble? The communication dimension of the reproducibility crisis in experiment psychology and neuroscience. *European Journal for Philosophy of Science, 10*, 44. https://doi.org/10.1007/s13194-020-00317-6

Heyde, C. C. (2006). Central limit theorem. *Encyclopedia of Actuarial Science*. https://doi.org/10.1002/9780470012505.tac019

Hida, R. M., Begeny, J. C., Oluokun, H. O., Bancroft, T. E., Fields-Turner, F., Ford, B. D., Jones, C. K., Ratliff, C. B., & Smith, A. Y. (2020). Internationalization and geographically representative scholarship in journals devoted to behavior analysis: An assessment of 10 journals across 15 years. *Scientometrics, 122*, 719-740. https://doi.org/10.1007/s11192-019-03289-4

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558. https://doi.org/10.1002/sim.1186

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*(1), 19-24. https://doi.org/10.1198/000313001300339897

Holmes, C. B. (1979). Sample size in psychological research. *Perceptual and Motor Skills, 49*(1), 283-288. https://doi.org/10.2466/pms.1979.49.1.283

Holmes, C. B. (1983). Sample size in four areas of psychological research. *Transactions of the Kansas Academy of Science (1903-), 86*(2/3), 76-80. https://doi.org/10.2307/3627914

Holmes, C. B., Holmes, J. R., & Fanning, J. J. (1981). Sample size in non-APA journals. *The Journal of Psychology, 108*, 263-266. https://doi.org/10.1080/00223980.1981.9915273

Hubbard, R., & Armstrong, J. S. (1997). Publication bias against null results. *Psychological Reports, 80*(1), 337-338. https://doi.org/10.2466/pr0.1997.80.1.337

Hubbard, R., & Ryan, P. (2000). The historical growth of statistical significance testing in psychology—and its future prospects. *Educational and Psychological Measurement, 60*(5), 661-681. https://doi.org/10.1177/00131640021970808

Hudson, R. (2023). Explicating exact versus conceptual replication. *Erkenntnis, 88*, 2493-2514. https://doi.org/10.1007/s10670-021-00464-z

Hughes, P. (1930). Forms of generalization, and their causes. *The Journal of Philosophy, 27*(11), 281-287. https://doi.org/10.2307/2015454

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology, 66*, 81-92. https://doi.org/10.1016/j.jesp.2015.09.009

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

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics reseach and randomized trials. *Journal of Clinical Epidemiology, 58*, 543-549. https://doi.org/10.1016/j.jclinepi.2004.10.019

Irvine, E. (2021). The role of replication studies in theory building. *Perspectives on psychological science, 16*(4), 844-853. https://doi.org/10.1177/1745691620970558

James, W. (1892). A plea for psychology as a 'natural science'. *The Philosophical Review, 1*(2), 146-153. https://doi.org/10.2307/2175743

John, I. D. (1992). Statistics as rhetoric in psychology. *Australian Psychologist, 27*(3), 144-149. https://doi.org/10.1080/00050069208257601

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

Jost, J. T. (2018). A quarter century of system justification theory: Questions, answers, criticisms, and societal applications. *British Journal of Social Psychology, 58*(2), 263-314. https://doi.org/10.1111/bjso.12297

Kaiser, M., Drivdal, L., Hjellbrekke, J., Ingierd, H., & Rekdal, O. B. (2022). Questionable research practices and misconduct among Norwegian researchers. *Science and Engineering Ethics, 28*, 2. https://doi.org/10.1007/s11948-021-00351-4

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217. https://doi.org/10.1207/s15327957pspr0203_4

Kekecs, Z., Palfi, B., Szaszi, B., Szecsi, P., Zrubka, M., Kovacs, M., Bakos, B. E., Cousineau, D., Tressoldi, P., Schmidt, K., Grassi, M., Evans, T. R., Yamada, Y., Miller, J. K., Liu, H., Yonemitsu, F., Dubrov, D., Röer, J. P., Becker, M., ..., & Aczel, B. (2023). Raising the value of research studies in psychological science by increasing the credibility of research reports: The transparent psi project. *Royal Society Open Science, 10*, 191375. https://doi.org/10.1098/rsos.191375

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*(2), 137-152. https://doi.org/10.1037/a0028086

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in organization sciences. *Organizational Research Methods, 15*(4), 624-662. https://doi.org/10.1177/1094428112452760

Kepes, S., Banks, G. C., & Oh, I.-S. (2012). Avoiding bias in publication bias research: The value of “null” findings. *Journal of Business and Psychology, 29*(2), 183-203. https://doi.org/10.1007/s10869-012-9279-0

Kim, U. (1999). After the “crisis” in social psychology: The development of the transactional model of science. *Asian Journal of Social Psychology, 2*(1), 1-19. https://doi.org/10.1111/1467-839X.00023

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology, 24*(3), 326-338. https://doi.org/10.1177/0959354314529616

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

Koch, S. (1951). Theoretical psychology, 1950: An overview. *Psychological Review, 58*(4), 295-301. https://doi.org/10.1037/h0055768

Kostyleff, R. (1911). *La crise de la psychologie expérimentale.* Alcan.

Krishna, A., & Peter, S. M. (2018). Questionable research practices in student final theses—prevalence, attitudes, and the role of the supervisor's perceived attitudes. *PLoS ONE, 13*(8), e0203470. https://doi.org/10.1371/journal.pone.0203470

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. https://doi.org/10.3758/s13423-016-1221-4

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist, 43*(8), 635-642. https://doi.org/10.1037//0003-066x.43.8.635

Ladd, G. T. (1892). Psychology as so-called 'natural science'. *The Philosophical Review, 1*(1), 24-53. https://doi.org/10.2307/2175528

Lakatos, I. (1978). Falsification and the methodology of scientific research programmes. In J. Worrall & G. Currie (Eds.), *The methodology of scientific research programmes: Philosophical papers* (pp. 8-101). Cambridge University Press. https://doi.org/10.1017/CBO9780511621123.003

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review, 62*(3), 221-230. https://doi.org/10.24602/sjpr.62.3_221

Lakens, D. (2020). Effect sizes and power for interactions in ANOVA designs. *The 20% Statistician*. URL: https://daniellakens.blogspot.com/2020/03/effect-sizes-and-power-for-interactions.html

Lakens, D. (2021, November 20). Why p-values should be interpreted as p-values and not as measures of evidence. *The 20% Statistician*. URL: https://daniellakens.blogspot.com/2021/11/why-p-values-should-be-interpreted-as-p.html

Lakens, D. (2023). Concerns about replicability, theorizing, applicability, generalizability, and methodology across two crises in social psychology. *PsyArXiv*. https://doi.org/10.31234/osf.io/dtvs7

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., …, Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour, 2*, 168-171. https://doi.org/10.1038/s41562-018-0311-x

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*(4), 371-379. https://doi.org/10.1037/a0025172

Ledgerwood, A., & Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science, 7*(1), 60-66. https://doi.org/10.1177/1745691611427304

Levi, I. (1967). *Gambling with truth: An essay on induction and the aims of science.* MIT Press.

Lewin, M. A. (1977). Kurt Lewin's view of social psychology: The crisis of 1977 and the crisis of 1927. *Personality and Social Psychology Bulletin, 3*(2), 159-172. https://doi.org/10.1177/014616727700300203

Linder, C., & Farahbakhsh, S. (2020). Unfolding the black box of questionable research practices: Where is the line between acceptable and unacceptable practices? *Business Ethics Quarterly, 30*(3), 335-360. https://doi.org/10.1017/beq.2019.52

Lindsay, R. M., & Ehrenberg, A. S. C. (1993). The design of replicated studies. *The American Statistician, 47*(3), 217-228. https://doi.org/10.2307/2684982

Line, W. (1931). Three recent attacks on associationism. *The Journal of General Psychology, 5*(4), 495-513. https://doi.org/10.1080/00221309.1931.9918419

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Sage.

Lubin, A. (1957). Replicability as a publication criterion. *American Psychologist, 12*(8), 519-520. https://doi.org/10.1037/h0039746

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

Lynch, J. G., Bradlow, E. T., Huber, J. C., & Lehmann, D. R. (2015). Reflections on the replication corner: In praise of conceptual replications. *International Journal of Research in Marketing, 32*(4), 333-342. https://doi.org/10.1016/j.ijresmar.2015.09.006

Machery, E. (2021). The alpha war. *Review of Philosophy and Psychology, 12*, 75-99. https://doi.org/10.1007/s13164-019-00440-1

Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language, 91*, 5-27. https://doi.org/10.1016/j.jml.2016.03.009

Maier, M., Bartoš, F., Stanley, T. D., Shanks, D. R., Harris, A. J. L., & Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *PNAS, 119*(31), e2200300119. https://doi.org/10.1073/pnas.2200300119

Manderscheid, L. V. (1965). Significance levels. 0.05, 0.01, or? *Journal of Farm Economics, 47*(5), 1381-1385. https://doi.org/10.2307/1236396

Maraun, M. D. (1998). Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory & Psychology, 8*(4), 435-461. https://doi.org/10.1177/0959354398084001

Marquis, D. G. (1948). Research planning at the frontiers of science. *American Psychologist, 3*(10), 430-438. https://doi.org/10.1037/h0056696

Marr, D. (1982). *Vision*. W. H. Freeman.

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

Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). Scientists behaving badly. *Nature, 435*(7043), 737-738. https://doi.org/10.1038/435737a

Mason, S., Merga, M. K., Canché, M. S. G., & Roni, S. M. (2021). The internationality of published higher education scholarship: How do the 'top' journals compare? *Journal of Informetrics, 15*, 101155. https://doi.org/10.1016/j.joi.2021.101155

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537-563. https://doi.org/10.1146/annurev.psych.59.103006.093735

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

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology, 59*, 927-953. https://doi.org/10.1111/j.1744-6570.2006.00059.x

McGrath, R. E. (2004). The making of meaning: Comments on Hofstee and Ten Berge. *Journal of Personality Assessment, 83*(2), 128-130. https://doi.org/10.1207/s15327752jpa8302_05

McNemar, Q. (1960). At random: Sense and nonsense. *American Psychologist, 15*(5), 295-300. https://doi.org/10.1037/h0049193

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2020). Average power: A cautionary note. *Advances in Methods and Practices in Psychological Science, 3*(2), 185-199. https://doi.org/10.1177/2515245920902370

Mead, C. D. (1917). Results in silent versus oral reading. *Journal of Educational Psychology, 8*(6), 367-368. https://doi.org/10.1037/h0067774

Mede, N. G., Schäfer, M. S., Ziegler, R., & Weißkopf, M. (2020). The "replication crisis" in the public eye: Germans' awareness and perceptions of the (ir)reproducibility of scientific research. *Public Understanding of Science, 30*(1), 91-102. https://doi.org/10.1177/0963662520954370

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103-115. https://doi.org/10.1086/288135

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

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*(1), 195-244. https://doi.org/10.2466/PR0.66.1.195-244

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin, 117*(3), 363-386. https://doi.org/10.1037/0033-2909.117.3.363

Mills, T. M. (1979). Changing paradigms for studying human groups. *Journal of Applied Behavioral Science, 15*(3), 407-423. https://doi.org/10.1177/002188637901500313

Moors, A. (2009). Theories of emotion causation: A review. *Cognition and Emotion, 23*(4), 625-662. https://doi.org/10.1080/02699930802645739

Moors, A. (2012). Comparison of affect program theories, appraisal theories, and psychological construction theories. In P. Zachar & R. D. Ellis (Eds.), *Categorical versus dimensional models of affect: A seminar on the theories of Panksepp and Russell* (pp. 257-278). John Benjamins.

Moran, C., Richard, A., Wilson, K., Twomey, R., & Coroiu, A. (2022). I know it's bad, but I have been pressured into it: Questionable research practices among psychology students in Canada. *Canadian Psychology/Psychologie Canadienne, 64*(1), 12-24. https://doi.org/10.1037/cap0000326

Morawski, J. (2019). The replication crisis: How might philosophy and theory of psychology be of use? *Journal of Theoretical and Philosophical Psychology, 39*(4), 218-238. https://doi.org/10.1037/teo0000129

Morawski, J. (2020). Psychologists' psychologies of psychologists in a time of crisis. *History of Psychology, 23*(2), 176-198. https://doi.org/10.1037/hop0000140

Morrison, D. E., & Henkel, R. E. (1969). Significance tests reconsidered. *The American Sociologist, 4*(2), 131-140. URL: https://www.jstor.org/stable/27701482

Moss, S., & Butler, D. C. (1978). The scientific credibility of ESP. *Perceptual and Motor Skills, 46*(3_suppl), 1063-1079. https://doi.org/10.2466/pms.1978.46.3c.1063

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Prims, J. P., Sun, J., Washburn, A. N., Wong, K. M., Yantis, C., & Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology, 113*(1), 34-58. https://doi.org/10.1037/pspa0000084

Muchinsky, P. M. (1979). Some changes in the characteristics of articles published in the *Journal of Applied Psychology* over the past 20 years. *Journal of Applied Psychology, 64*(4), 455-459. https://doi.org/10.1037/0021-9010.64.4.455

Mulder, J. D. (2023). Power analysis for the random intercept cross-lagged panel model using the powRICLPM R-package. *Structural Equation Modeling: A Multidisciplinary Journal, 30*(4), 645-658. https://doi.org/10.1080/10705511.2022.2122467

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

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and replicability in science*. The National Academies Press. https://doi.org/10.17226/25303

Nederhof, A. J., & Zwier, A. G. (1983). The 'crisis' in social psychology, an empirical approach. *European Journal of Social Psychology, 13*(3), 255-280. https://doi.org/10.1002/ejsp.2420130305

Neher, A. (1967). Probability pyramiding: Research error and the need for independent replication. *The Psychological Record, 17*, 257-262. https://doi.org/10.1007/BF03393713

Neyman, J., & Pearson, E. S. (1933a). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society, 29*(4), 492-510. https://doi.org/10.1017/S030500410001152X

Neyman, J., & Pearson, E. S. (1933b). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*, 289-337. URL: https://www.jstor.org/stable/91247

Neyman, J. (1950). *First course in probability and statistics.* Henry Holt and Company, Inc.

Nguyen, P. H., Engel, S. M., & Herring, A. H. (2022). mpower: An R package for power analysis via simulation for correlated data. *arXiv.* https://doi.org/10.48550/arXiv.2209.08036

Nieuwenstein, M., & van Rijn, H. (2012). The unconscious thought advantage: Further replication failures from a search for confirmatory evidence. *Judgment and Decision Making, 7*(6), 779-798. https://doi.org/10.1017/S1930297500003338

Niiniluoto, I. (2014). Scientific progress as increasing verisimilitude. *Studies in History and Philosophy of Science Part A, 46*, 73-77. https://doi.org/10.1016/j.shpsa.2014.02.002

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

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods, 48*(4), 1205-1226. https://doi.org/10.3758/s13428-015-0664-2

Nuijten, M. B., van Assen, M. A. L. M., Augusteijn, H. E. M., Crompvoets, E. A. V., & Wicherts, J. M. (2020). Effect sizes, power, and biases in intelligence research: A meta-meta-analysis. *Journal of Intelligence, 8*(4), 36. https://doi.org/10.3390/jintelligence8040036

Oakes, M. W. (1986). *Statistical inference: A commentary for the social and behavioural sciences* (1st ed.). Wiley.

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

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive *t* tests: Lay use of statistical information. *Psychonomic Bulletin & Review, 14*(6), 1147-1152. https://doi.org/10.3758/BF03193104

Olsen, J., Mosen, J., Voracek, M, & Kirchler, E. (2019). Research practices and statistical reporting quality in 250 economic psychology master's theses: A meta-research investigation. *Royal Society Open Science, 6*, 190738. https://doi.org/10.1098/rsos.190738

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4715. https://doi.org/10.1126/science.aac4716

Osborne, J. W. (2008). Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology, 28*(2), 151-160. https://doi.org/10.1080/01443410701491718

Palus, S. (2015, December 8). Diederik Stapel now has 58 retractions. *Retraction Watch.* URL: https://retractionwatch.com/2015/12/08/diederik-stapel-now-has-58-retractions/

Parker, I. (2007). Critical psychology: What it is and what it is not. *Social and Personality Psychology Compass, 1*(1), 1-15. https://doi.org/10.1111/j.1751-9004.2007.00008.x

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments. *PLoS ONE, 7*(8), e42510. https://doi.org/10.1371/journal.pone.0042510

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*(6), 531-536. https://doi.org/10.1177/1745691612463401

Peels, R. (2019). Replicability and replication in the humanities. *Research Integrity and Peer Review, 4*, 2. https://doi.org/10.1186/s41073-018-0060-4

Pek, J., Hoisington-Shaw, K. J., & Wegener, D. T. (2022). Avoiding questionable research practices surrounding statistical power analysis. In W. O'Donohue, A. Masuda, & S.

Lilienfeld (Eds.), *Avoiding questionable research practices in applied psychology*. Springer. https://doi.org/10.1007/978-3-031-04968-2_11

Pereboom, A. C. (1971). Some fundamental problems in experimental psychology: An overview. *Psychological Reports, 28*(2), 439-455. https://doi.org/10.2466/pr0.1971.28.2.439

Perugini, M., Gallucci, M., & Constantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9*(3), 319-332. https://doi.org/10.1177/1745691614528519

Peters, C. C. (1938). An example of replication of an experiment for increased reliability. *Journal of Educational Research, 32*(1), 3-9. URL: https://www.jstor.org/stable/27526484

Piaget, J., & Kamii, C. (1978). What is psychology? *American Psychologist, 33*(7), 648-652. https://doi.org/10.1037/0003-066X.33.7.648

Popper, K. R. (1959). *The logic of scientific discovery.* Basic Books.

Pupovac, V., Prijić-Samaržija, S., & Petrovečki, M. (2017). Research misconduct in the Croatian scientific community: A survey assessing the forms and characteristics of research misconduct. *Science and Engineering Ethics, 23*(1), 165-181. https://doi.org/10.1007/s11948-016-9767-0

Quay, P. M. (1974). Progress as a demarcation criterion for the sciences. *Philosophy of Science, 41*(2), 154-170. URL: https://www.jstor.org/stable/186865

Radder, H. (1992). Experimental reproducibility and the experimenters' regress. *Proceedings of the Biennial Meeting of the Philosophy of Science Association,* Volume One: Contributed Papers, 63-73. URL: https://www.jstor.org/stable/192744

Radzikhovskii, L. A. (1992). The historical meaning of the crisis in psychology. *Russian Social Science Review, 33*(1), 70-93. https://doi.org/10.2753/RSS1061-1428330170

Rajah-Kanagasabai, C. J., & Roberts, L. D. (2015). Predicting self-reported research misconduct and questionable research practices in university students using an augmented Theory of Planned Behavior. *Frontiers in Psychology, 6*, 535. https://doi.org/10.3389/fpsyg.2015.00535

Rastle, K., Chan, J., Cleary, A., Pexman, P., & Staub, A. (2023). Beware influential findings that have not been replicated. *Journal of Memory and Language, 129*, 104390. https://doi.org/10.1016/j.jml.2022.104390

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. URL: https://www.R-project.org/

Reardon, K. W., Smack, A. J., Herzhoff, K., & Tacken, J. L. (2019). An *N*-pact factor for clinical psychological research. *Journal of Abnormal Psychology, 128*(6), 493-499. https://doi.org/10.1037/abn0000435

Reber, A. S., & Alcock, J. E. (2020). Searching for the impossible: Parapsychology's elusive quest. *American Psychologist, 75*(3), 391-399. https://doi.org/10.1037/amp0000486

Reed, H. B. (1917). A repetition of Ebert and Meumann's practice experiment on memory. *Journal of Experimental Psychology, 11*(5), 315-346. https://doi.org/10.1037/h0073769

Rees, M. J. (1980). Gravitational collapse and cosmology. *Contemporary Physics, 21*(2), 99-120. https://doi.org/10.1080/00107518008210948

Reinero, D. A., Wills, J. A., Brady, W. J., Mende-Siedlecki, P., Crawford, J. T., & Van Bavel, J. J. (2020). Is the political slant of psychology research related to scientific replicability? *Perspectives on Psychological Science, 15*(6), 1310-1328. https://doi.org/10.1177/1745691620924463

Remmers, H. H., & Whisler, L. (1938). Test reliability as a function of method of computation. *The Journal of Educational Psychology, 29*(2), 81-92. https://doi.org/10.1037/H0056123

Renkewitz, F., & Heene, M. (2019). The replication crisis and open science in psychology: Methodological challenges and developments. *Zeitschrift für Psychologie, 227*(4), 233-236. https://doi.org/10.1027/2151-2604/a000389

Renkewitz, F., & Keiner, M. (2019). How to detect publication bias in psychological research: A comparative evaluation of six statistical methods. *Zeitschrift für Psychologie, 227*'(4), 261-279. https://doi.org/10.1027/2151-2604/a000386

Resnick, B. (2016, March 25). What psychology's crisis means for the future of science. *Vox*. https://www.vox.com/2016/3/14/11219446/psychology-replication-crisis

Rhine, J. E. (1934). *Extrasensory perception.* Society for Psychic Research.

Rich, P., de Haan, R., Wareham, T., & van Rooij, I. (2021). How hard is cognitive science? *Proceedings of the Annual Meeting of the Cognitive Science Society,* 43. URL: https://escholarship.org/uc/item/8cr8x1c4

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331-363. https://doi.org/10.1037/1089-2680.7.4.331

Ring, K. (1967). Experimental social psychology: Some sober questions about some frivolous values. *Journal of Experimental Social Psychology, 3*(2), 113-123. https://doi.org/10.1016/0022-1031(67)90016-9

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS ONE, 7*(3), e33423. https://doi.org/10.1371/journal.pone.0033423

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science, 16*(4), 725-743. https://doi.org/10.1177/1745691620974697

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass, 14*, e12633. https://doi.org/10.1111/phc3.12633

Rosenstreich, D., & Wooliscroft, B. (2006). How international are the top academic journals? The case of marketing. *European Business Review, 18*(6), 422-436. https://doi.org/10.1108/09555340610711067

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641. https://doi.org/10.1037/0033-2909.86.3.638

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). Russell Sage Foundation.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*(5), 646-656. https://doi.org/10.1037/0022-006X.58.5.646

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18*, 682-689. https://doi.org/10.3758/s13423-011-0088-7

Rubin, M. (2022). The costs of HARKing. *The British Journal for the Philosophy of Science, 73*(2), 535-560. https://doi.org/10.1093/bjps/axz050

Rubin, M. (2023). Questionable metascience practices. *Journal of Trial & Error.* https://doi.org/10.36850/mr4

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton University Press. https://doi.org/10.2307/j.ctv173f2gh

Salzinger, K. (1996). How many new discoveries do we need to avoid a crisis in psychology? *Joursal of Social Distress and the Homeless, 5*(4), 353-357. https://doi.org/10.1007/BF02092912

Schäfer, T. (2023). On the use and misuse of standardized effect sizes in psychological research. *Open Science Framework*. https://doi.org/10.31219/osf.io/x8n3h

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology, 10*, 813. https://doi.org/10.3389/fpsyg.2019.00813

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science, 4*(2), 1-12. https://doi.org/10.1177/25152459211007467

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science, 16*(4), 744-755. https://doi.org/10.1177/1745691620966795

Schikore, J. (2011). What does history matter to philosophy of science? The concept of replication and the methodology of experiments. *Journal of the Philosophy of History, 5*, 513-532. https://doi.org/10.1163/187226311X599934

Schimmack, U. (2019). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science, 16*(2), 396-414. https://doi.org/10.1177/1745691619863798

Schimmack, U., & Brunner, J. (2017). Z-curve: A method for estimating replicability based on test statistics in original studies. URL: http://datacolada.org/wp-content/uploads/2017/12/5777-Schimmack-Brunner-Z-Curve.pdf

Schlosberg, H. (1951). Repeating fundamental experiments. *American Psychologist, 6*(5), 177. https://doi.org/10.1037/h0056148

Schmidt, F., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*(4), 901-912. https://doi.org/10.1111/j.1744-6570.2000.tb02422.x

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90-100. https://doi.org/10.1037/a0015108

Schoenegger, P., & Pils, R. (2023). Social sciences in crisis: On the proposed elimination of the discussion section. *Synthese, 202*, 54. https://doi.org/10.1007/s11229-023-04267-3

Schwarzkopf, D. S. (2014). We should have seen this coming. *Frontiers in Human Neuroscience, 8*, article 332. https://doi.org/10.3389/fnhum.2014.00332

Sedgwick, P. (2015). Meta-analyses: What is heterogeneity? *BMJ, 350*, h1435. https://doi.org/10.1136/bmj.h1435

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

Senn, S. J. (2002). Power is indeed irrelevant in interpreting completed studies. *British Medical Journal, 325*(7375), 1304. https://doi.org/10.1136/bmj.325.7375.1304

Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (2011). Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology, 96*(5), 1055-1064. https://doi.org/10.1037/a0023322

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*(1), 487-510. https://doi.org/10.1146/annurev-psych-122216-011845

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., & Simonsohn, U. (2017). Power posing: *P*-curving the evidence. *Psychological Science, 28*(5), 687-693. https://doi.org/10.1177/0956797616658563

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76-80. https://doi.org/10.1177/1745691613514755

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559-569. https://doi.org/10.1177/0956797614567341

Singleton Thorn, F., Fidler, F., & Dudgeon, P. (2019). The statistical power of psychological research: A systematic review and meta-analysis. *Open Science Framework*. https://doi.org/10.17605/OSF.IO/H8U9W

Smaldino, P. E. (2019). Better methods can't make up for mediocre theory. *Nature, 575*(7781), 9. https://doi.org/10.1038/d41586-019-03350-5

Smaldino, P. E. (2020). How to translate a verbal theory into a formal model. *Social Psychology, 51*(4), 207-218. https://doi.org/10.1027/1864-9335/a000425

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-*N* design. *Psychonomic Bulletin & Review, 25*, 2083-2101. https://doi.org/10.3758/s13423-018-1451-8

Sotola, L. K., & Credé, M. (2021). On the predicted replicability of two decades of experimental research on system justification: A Z-curve analysis. *European Journal of Social Psychology, 52*(5-6), 895-909. https://doi.org/10.1002/ejsp.2858

Spearman, C. (1913). Correlations of sums or differences. *British Journal of Psychology, 5*(4), 417-426. https://doi.org/10.1111/J.2044-8295.1913.TB00072.X

Stamm, K., Christidis, P., & Lin, L. (2017). How much federal funding is directed to research in psychology. *Monitor on Psychology, 48*(4). URL: https://www.apa.org/monitor/2017/04/datapoint

Stanley, T. D. (2005). Beyond publication bias. *Journal of Economic Surveys, 19*(3), 309-345. https://doi.org/10.1111/j.0950-0804.2005.00250.x

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144*(12), 1325-1346. https://doi.org/10.1037/bul0000169

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of *p*-hacking strategies. *Royal Society Open Science, 10*, 220346. https://doi.org/10.1098/rsos.220346

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Psychology, 227*(4), 280-292. https://doi.org/10.1027/2151-2604/a000385

Steinle, F. (2016). Stability and replication of experimental results: A historical perspective. In H. Atmans-pacher & S. Maasen (Eds.), *Reproducibility: Principles, problems, practices, and prospects* (pp. 39-63). John Wiley & Sons, Inc.

Sterrett, J. M. (1909). The proper affiliation of psychology—with philosophy or with the natural sciences? *Psychological Review, 16*(2), 85-106. https://doi.org/10.1037/h0073595

Stevens, S. S. (1939). Psychology and the science of science. *Psychological Bulletin, 36*(4), 221-263. https://doi.org/10.1037/h0056886

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*(3), 768-777. https://doi.org/10.1037/0022-3514.54.5.768

Strahan, R. F. (1982). Multivariate analysis and the problem of type I error. *Journal of Counseling Psychology, 29*(2), 175-179. https://doi.org/10.1037/0022-0167.29.2.175

Strick, M., Dijksterhuis, A., Bos, M. W., Sjoerdsma, A., van Baaren, R. B., & Nordgren, L. F. (2011). A meta-analysis on unconscious thought effects. *Social Cognition, 29*(6), 738-762. https://doi.org/10.1521/soco.2011.29.6.738

Stricker, J., & Günther, A. (2019). Scientific misconduct in psychology: A systematic review of prevalence estimates and new empirical data. *Zeitschrift für Psychologie, 227*(1), 53-63. https://doi.org/10.1027/2151-2604/a000356

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*(1), 59-71. https://doi.org/10.1177/1745691613514450

Student (1908). The probable error of a mean. *Biometrika, 6*(1), 1-25. https://doi.org/10.2307/2331554

Sturm, T., & Mülberger, A. (2012). Crisis discussions in psychology—New historical and philosophical perspectives. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(2), 425-433. https://doi.org/10.1016/j.shpsc.2011.11.001

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 102*(4), 989-1004. https://doi.org/10.1037/a0019507

Suter, W. N. (2020). Questionable research practices: How to recognize and avoid them. *Home Health Care Management & Practice, 32*(4), 183-190. https://doi.org/10.1177/1084822320934468

Świątkowski, W., & Dompnier, B. (2017). Replicability crisis in social psychology: Looking at the past to find new pathways for the future. *International Review of Social Psychology, 30*(1), 111-124. https://doi.org/10.5334/irsp.66

Symonds, P. M. (1928). Factors influencing test reliability. *The Journal of Educational Psychology, 14*(2), 73-87. https://doi.org/10.1037/h0071867

Tein, J.-Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(4), 640-657. https://doi.org/10.1080/10705511.2013.824781

Thagard, P. (2007). Coherence, truth, and the development of scientific knowledge. *Philosophy of Science, 74*(1), 28-47. https://doi.org/10.1086/520941

Thaler, R. H., & Sunstein, C. R. (2021). *Nudge: The final edition* (revised ed.). Penguin Putnam Inc.

Thoemmes, F., MacKinnon, D. P., & Reiser, M. R. (2010). Power analysis for complex mediational designs using Monte Carlo methods. *Structural Equation Modeling: A Multidisciplinary Journal, 17*(3), 510-534. https://doi.org/10.1080/10705511.2010.489379

Thompson, B. (2004). The "significance" crisis in psychology and education. *The Journal of Socio-Economics, 33*(5), 607-613. https://doi.org/10.1016/j.socec.2004.09.034

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology, 53*, 207-216. https://doi.org/10.1016/S0895-4356(99)00161-4

Tichý, P. (1976). Verisimilitude redefined. *The British Journal for the Philosophy of Science, 27*(1), 25-42. URL: https://www.jstor.org/stable/686376

Tijdink, J. K., Verbeke, R., & Smulders, Y. M. (2014). Publication pressure and scientific misconduct in medical scientists. *Journal of Empirical Research on Human Research Ethics, 9*(5), 64-71. https://doi.org/10.1177/1556264614552421

Timmons, A. C., & Preacher, K. J. (2015). The importance of temporal design: How do measurement intervals affect the accuracy and efficiency of parameter estimates in longitudinal research? *Multivariate Behavioral Research, 50*(1), 41-55. https://doi.org/10.1080/00273171.2014.961056

Tinkler, J. E. (2012). Controversies in implicit race bias research. *Sociology Compass, 6*(12), 987-997. https://doi.org/10.1111/soc4.12001

Tonneau, F. (2011). Associationism. In N. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 326-329). Springer. https://doi.org/10.1007/978-1-4419-1428-6_505

Trafimow, D. (2018). An a priori solution to the replication crisis. *Philosophical Psychology, 31*(8), 1188-1214. https://doi.org/10.1080/09515089.2018.1490707

Tressoldi, P. E., & Giofré, D. (2015). The pervasive avoidance of prospective statistical power: Major consequences and practical solutions. *Frontiers in Psychology, 6*, 726. https://doi.org/10.3389/fpsyg.2015.00726

Tressoldi, P. E., Giofré, D., Sella, F., & Cumming, G. (2013). High impact = high statistical standards? Not necessarily so. *PLoS ONE, 8*(2), e56180. https://doi.org/10.1371/journal.pone.0056180

Tsang, E. W. K., & Kwan, K.-M. (1999). Replication and theory development in organizational science: A critical realist perspective. *The Academy of Management Review, 24*(4), 759-780. https://doi.org/10.2307/259353

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105-110. https://doi.org/10.1037/h0031322

Tucker, J. (2016, March 9). Does social science have a replication crisis. *The Washington Post.* https://www.washingtonpost.com/news/monkey-cage/wp/2016/03/09/does-social-science-have-a-replication-crisis/

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

Ulrich, R., & Miller, J. (2020). Meta-research: Questionable research practices may have little effect on replicability. *eLife, 9*, e58237. https://doi.org/10.7554/eLife.58237

Utts, J. (1991). Replication and meta-analysis in parapsychology. *Statistical Science, 6*(4), 363-403. https://doi.org/10.1214/ss/1177011577

Vankov, I., Bowers, J., & Munafò, M. R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology, 67*(5), 1037-1040. https://doi.org/10.1080/17470218.2014.885986

van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories: A tutorial by dialogue. *Social Psychology, 51*(5), 285-298. https://doi.org/10.1027/1864-9335/a000428

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science, 16*(4), 682-697. https://doi.org/10.1177/1745691620970604

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*(3), 274-290. https://doi.org/10.1111/j.1745-6924.2009.01125.x

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., & Gronau, Q. F. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11*(6), 917-928. https://doi.org/10.1177/1745691616674458

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

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

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics, 10*(4), 299-326. URL: https://www.jstor.org/stable/2235609

Washburn, A. N., Hanson, B. E., Motyl, M., Skitka, L. J., Yantis, C., Wong, K. M., Sun, J., Prims, J. P., Mueller, A. B., Melton, Z. J., & Carsel, T. S. (2018). Why do some psychology researchers resist adopting proposed reforms to research practices? A description of researchers' rationales. *Advances in Methods and Practices in Psychological Science, 1*(2), 166-173. https://doi.org/10.1177/2515245918757427

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*(2), 158-177. https://doi.org/10.1037/h0074428

Weisberg, M. (2006). Water is *not* H_2O. In D. Baird, E. Scerri, & L. McIntyre (Eds.), *Philosophy of chemistry: Synthesis of a new discipline* (pp. 337-345). Springer.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Wieser, M. (2016). Psychology's "crisis" and the need for reflection. A plea for modesty in psychological theorizing. *Integrative Psychological and Behavioral Science, 50*, 359-367. https://doi.org/10.1007/s12124-016-9343-9

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview of theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology, 39*(4), 202-217. https://doi.org/10.1037/teo0000137

Willy, R. (1899). *Die Krisis in der Psychologie.* O. R. Reisland.

Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. In D. T. Campbell, M. B., Brewer, & B. E. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 123-162). Jossey-Bass.

Witt, J. K. (2019). Insights into criteria for statistical significance from signal detection analysis. *Meta-Psychology,* vol. 3. https://doi.org/10.15626/MP.2018.871

Whewell, W. (1858). *Novum organon renovatum: Being the second part of the philosophy of the inductive sciences.* John W. Parker & Son

Wundt, W. (1907). Über Ausfrageexperimente über die Methoden zur Psychologie des Denkens. In W. Wundt (Ed.), *Psychologische Studien 3* (pp. 301-360). Verlag von Wilhelm Engelmann.

Yates, F. (1964). Sir Ronald Fisher and the design of experiments. *Biometrics, 20*(2), 307-321. https://doi.org/10.2307/2528399

Yong, E. (2012, October 3). Nobel laureate challenges psychologists to clean up their act. *Nature.* https://doi.org/10.1038/nature.2012.11535

Yong, E. (2013, November 26). Welcome to the era of big replication. *National Geographic.* https://www.nationalgeographic.com/science/article/welcome-to-the-era-of-big-replication

Yong, E. (2016, March 4). Psychology's replication crisis can't be wished away. *The Atlantic.* https://www.theatlantic.com/science/archive/2016/03/psychologys-replication-crisis-cant-be-wished-away/472272/

Zarahn, E., & Slifstein, M. (2001). A reference effect approach for power analysis in fMRI. *NeuroImage, 14*, 768-779. https://doi.org/10.1006/nimg.2001.0852

Zhang, X., Astivia, O. L. O., Kroc, E., & Zumbo, B. D. (2021). How to think clearly about the central limit theorem. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000448

Zhang, Z., & Mai, Y. (2021). WebPower: Basic and advanced statistical power analysis. R package version 0.6. URL: https://CRAN.R-project.org/package=WebPower

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, e120. https://doi.org/10.1017/S0140525X17001972