Univerzita Karlova
Pedagogická fakulta
Katedra psychologie

BAKALÁŘSKÁ PRÁCE

# Struktura incentivů v psychologickém výzkumu a integrita dat v randomizovaných studiích psychologických intervencí

# The incentive structure of psychological science and the integrity of reported data in randomised controlled trials of psychological interventions

Benjamin Šimsa

| | |
|---|---|
| Vedoucí práce: | Mgr. Ivan Ropovik, PhD. |
| Studijní program: | Psychologie |
| Studijní obor: | Psychologie s rozšířením o speciální pedagogiku |

2021

Odevzdáním této bakalářské práce na téma *Struktura incentivů v psychologickém výzkumu a integrita dat v randomizovaných studiích psychologických intervencí* potvrzuji, že jsem ji vypracoval pod vedením vedoucího práce samostatně za použití v práci uvedených pramenů a literatury. Dále potvrzuji, že tato práce nebyla využita k získání jiného nebo stejného titulu.

V Praze dne 18. dubna 2021 Benjamin Šimsa

……

I would like to thank Dr. Ivan Ropovik for introducing me to the world of meta-research as well as for his guidance and patience during the thesis-writing process.

This thesis (as well as most of my other research endeavors) would not have been possible without Alexandra Elbakyan's work. Thank you for making science truly open.

**ABSTRAKT**

Problémy spojené s nedostatečnou replikovatelností a reliabilitou výzkumných poznatků se v rámci akademické psychologie během poslední dekády staly široce diskutovaným tématem. Do popředí diskurzu psychologické metodologie se mimo jiné dostalo take téma diskutabilních výzkumných praktik *(questionable research practices)*. Tato bakalářská práce podává přehled dostupných poznatků o aktuálních metodologických problémech psychologického výzkumu spojených s používání statistického přístupu *NHST* (testování signifikance nulových hypotéz) s důrazem na diskutabilní výzkumné praktiky ohrožující integritu výzkumných dat. Mezi tyto praktiky patří p-hacking (zneužívání analytické flexibility), vytváření post-hoc hypotéz nebo falzifikace a fabrikace výzkumných dat. Práce dále pojednává o faktorech asociovaných s těmito problematickými praktikami jak na úrovni individuálních výzkumníků, tak na strukturální úrovni. Zvláštní důraz je v teoretické části kladen na strukturu incentivů v psychologickém výzkumu. Těchto poznatků pak je využito v praktické části práce. Cílem empirické části je odhad přítomnosti problémů s integritou reportovaných pretestových dat v randomizovaných kontrolovaných studiích intervencí psychologického charakteru pomocí statistických metod (Carlislova metoda, Stoufferova a Fisherova metoda, Monte Carlo simulace, GRIM, GRIMMER). Výsledky praktické části naznačují přítomnost systematických nekonzistencí a nepravidelností v těchto reportovaných datech v psychologických studiích. Dále jsou diskutovány možné příčiny těchto problémů a nedostatků, včetně problémů v procesu randomizace a diskutabilních výzkumných praktik.

**KLÍČOVÁ SLOVA**

psychologická metodologie, metavýzkum, replikační krize v psychologii, diskutabilní výzkumné praktiky, integrita reportovaných dat

**ABSTRACT**

The recent replication crisis in psychology has sparked a broader debate about the threats that the limitations of the null hypothesis significance testing approach and the use of questionable research practices by research psychologists pose to the reliability and replicability of psychological findings. The present thesis summarizes available evidence on the factors associated with questionable research practices on both the level of individual researchers and on the structural level. Special emphasis is put on the role of the incentive structure of psychological research in promoting different forms of scientific misconduct. The goal of the empirical part is to estimate the presence of statistical inconsistencies in reported baseline data in randomized controlled trials of psychological interventions using several statistical techniques (Carlisle's, Stouffer and Fisher's methods, Monte Carlo simulations, GRIM, GRIMMER). The results of the present thesis indicate a presence of systematical inconsistencies and other problems with the integrity of reported data in the literature. Moreover, possible causes of these inconsistencies, including problems with randomization and the presence of the use of questionable research practices.

**KEYWORDS**

psychological methodology, meta-research, replication crisis in psychology, questionable research practices, data integrity

# Contents

## Introduction

The concerns about the replicability of published findings (or lack thereof) started to permeate the mainstream discourse of academic psychology in the eventful years of 2011 and 2012. In autumn 2011, the psychological research community was shook by the case of Diederik Stapel, a Dutch social psychologist who, unbeknownst to his students and collaborators, fabricated data for more than 50 of his first-authored articles (some of them published in prestigious outlets such as *Nature* or *JPSP*; Markowitz & Hancock, 2014).

Another high-profile event that triggered this shift of focus was the publication of Daryl Bem's article *Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect* in the prestigious *Journal of Personality and Social Psychology.* In this paper, Bem claimed to have found evidence for human precognition in 8 out of 9 of the included experiments, while using methodological procedures and analytical methods that were widely used and accepted in psychology at the time. A wave of skepticism and controversy ensued. In reaction to Bem's study, Simmons, Nelson and Simonsohn (2011) published a paper warning about the easiness of getting false positive results by exploiting analytical flexibility inherent in behavioral data analysis. Parallel to these attempts to shine light on the inadequacies of psychological research methods, large multi-centric replication efforts such as *The Reproducibility Project: Psychology* emerged, with their results suggesting that the replicability of the effects found in psychological studies might be concerningly low[1] (Open Science Collaboration, 2015).

Following these events, the last decade in psychological research saw topics such as open science, meta-research (the scientific inquiry into the problems related with research itself), pre-registration, questionable research practices, and publication bias, as well as various crisis narratives (e.g., replication crisis, theory crisis or generalizability crisis) being brought into the spotlight.

[1] Out of the 100 effects replicated in the *Reproducibility Project: Psychology,* 97 were significant in the original studies. Only 36 of these replicated successfully in the pre-registered, adequately powered replication studies. It is important to note, though, that it is always problematic to draw conclusions from single studies, however methodologically sound they are.

Regardless of their relative novelty in the mainstream discourse, the concerns about replicability and substandard research practices are not exactly new in psychology (and the broader fields of social/life sciences). In 2005, John Ioannidis famously warned that, given the flexibility in research designs and statistical analysis, publication bias, insufficient statistical power, and overall lesser true effect sizes, a majority of published research findings in some scientific subfields might be based on false-positive results. Sterling warned about the inflated false-positive rates caused by the selective publication of significant results and lack of replication studies as early as 1959. In psychology, similar points about publication bias and the aversion to the null were raised by Greenwald (1975) and Rosenthal (1979). In similar fashion, Cohen (1962) pointed out the lack of sufficient statistical power in psychological research and its detrimental effect on scientific inferences. Paul Meehl had been warning psychologists about various methodological drawbacks since the fifties, and the discussion about the problems related with the use of null hypothesis significance testing (NHST), which is predominant in psychology until these days, has been around for as long as NHST itself.

Even though these problems had been discussed for decades, their influence on the research practices that were actually used in psychological research was very limited. The discussion remained on the fringe of psychological discourse, only to be brought to the spotlight in the last decade. Even then, the pace of the adoption of more robust and reliable research practices has proved to be quite slow, bringing the question of the incentive structure in academia into the picture. However problematic questionable research practices are for scientific process (i.e., drawing reliable conclusions about the outside world) itself, it might well be the case that in the increasingly competitive ecosystem in academia, in which the merits of one's work are measured by scientometric indices (such as the number of citations and publications in impacted journals), the use of suboptimal research methods may in fact be a rational, reward-maximizing choice for researchers to make. On the other hand, strictly using reliable research practices might put honest researchers at a competitive disadvantage. There is a growing body of evidence (from both qualitative and quantitative studies, as well as principal-agent models) that this is indeed the case. In order to improve research practices and prevent strategic game-playing in psychological research, wider, more

structural changes of incentives (as opposed to the emphasis on individual integrity and better education and training in psychological research methods) might be needed.

In Chapter 1 of the present thesis, I am going to summarize the evidence about the character and prevalence of questionable research practices (QRPs) in psychology and about the broader category of research misconduct, including scientific fraud. In order to do so, a brief inquiry into the process of NHST and its possible drawbacks will be included as well. The chapter will conclude with a section focused on the factors associated with participation in research misconduct and the use of questionable research practices on the level of individual researchers.

Chapter 2 is going to focus on the incentive structure of psychological research – the structural factors that directly influence researchers' behavior and might be among the main causes of scientific misconduct. I will present evidence on publication bias and the psychology's aversion to the null, the increasingly competitive nature of psychological research, along with other elements that constitute the ecosystem of academic research. Agent-based models together with qualitative studies that conceptualize research misconduct as rational reward-maximizing strategy will be presented.

In the empirical part of this thesis, I am going to implement several methods of estimating the prevalence of inconsistencies and irregularities in data reported in randomized controlled studies of psychological interventions. Along with other possible causes, certain patterns of such inconsistencies might be indicative of illegitimate data manipulation or the use of questionable research practices. A statistical procedure for assessing the integrity of reported baseline data in randomized controlled trials, first used by John Carlisle (2017) will be employed as the primary method to assess those inconsistencies.

The empirical part of the present thesis is a pilot study for a larger research project carried out at the Center for meta-research in education, Charles University, led by Dr. Ivan Ropovik.

# I Theoretical part

# 1 Scientific misconduct and questionable research practices

In the first chapter I am going to introduce the basic concepts related to questionable research practices and scientific misconduct as primary causes of reproducibility problems in psychological science, along with the factors that are associated with such problematic research behaviors on the level of individual researchers. Furthermore, I will introduce some of the more widespread statistical methods used to assess the prevalence of questionable research practices from reported data and the distributions of *p*-values.

As the current discourse on questionable research practices in psychology are usually tied with the wider discussion about the use of Null hypothesis significance testing, I will begin the chapter with a brief discussion on the basic principles and drawbacks of this statistical paradigm.

## 1.1 Null hypothesis significance testing

The use of null hypothesis significance testing (NHST) in psychology is quite controversial and often criticized, with some of its critics going as far as to assert its use in the "soft" areas of psychology to be "a terrible mistake, scientifically unsound . . . and one of the worst things that ever happened in the history of psychology" (Meehl, 1978). Despite these criticisms, NHST still remains the dominant paradigm in behavioral data analysis (Nickerson, 2000). The NHST approach borrows terminology and concepts from two frequentist statistical frameworks – Fisher's *significance testing* and Neyman and Pearson's *hypothesis testing* (Perezgonzales, 2015). It is important to point out that the criticisms of NHST are usually directed at the inconsistencies that arise from NHST being a hybrid of Fisher's and Neyman-Pearson's approach, not at these two distinct paradigms themselves.

While the concept of *p*-value as a continuous inductive measure of the strength of evidence against the null hypothesis was introduced by Ronald Fisher, Neyman and Pearson focused on the minimization of errors in categorical inferential decision-making. To

understand the NHST procedure, a brief introduction into the concepts of *p*-values and Type I / II errors is needed.

### 1.1.1 *P*-value and its behavior under null and alternative hypothesis

*P*-value is defined as $P(D|H_0)$, i.e., the conditional probability of the obtained *and* more extreme results given that the null hypothesis is true. When the null hypothesis of a specific statistical test is true, the distribution of possible *p*-values follows a uniform (0, 1) distribution. This applies notwithstanding the sample size or statistical power of the given statistical test (Hung et al., 1997). Under the alternative hypothesis, the distribution of the random variable from which the *p*-values for individual studies are drawn is right-skewed. The skewness of the distribution increases with increasing statistical power and effect size (Hung et al., 1997).

### 1.1.2 Type I error and the α level

Type I error is the rejection of the null hypothesis when it is, in fact, true (Neyman and Pearson, 1933). In the Neyman-Pearson approach, which focuses on the control of error rates in the long run, the maximum probability of Type I error is denoted by α. In the NHST framework, the α level also stands for the level of significance (i.e., the critical threshold for the maximum *P*-value needed to reject the null hypothesis) (Perezgonzalez, 2015).

Although the α level is arbitrary and should only be set after considering the needs of the particular study (Lakens et al., 2018), α levels of 0.05 and 0.01 are the most common significance thresholds used in psychological research (Lavrakas, 2008). Theoretically, with α set to 0.05, *at most* 5% of the positive results are actually false positives. This only holds true under the assumption of the absence of *p-hacking*, described in more detail in subchapter 2.2.1.

### 1.1.3 Type II error and statistical power

Type II error is the rejection of a true alternative hypothesis (Neyman and Pearson, 1933). The maximum probability of Type II error in the long run is usually denoted as β. The probability of rejecting the null hypothesis when there is a true effect present (i.e., concluding a true positive) is thus 1 – β, also known as *statistical power*. Statistical power

is primarily dependent on the (unknown) true effect size, sample size of the given study and the α level set by the researchers (Cohen, 1992).

Although the consequences of false negative findings are not usually considered to be as harmful as the effects of drawing conclusions from Type I errors, there is a growing awareness about them in psychological research. Despite the efforts to warn about the insufficient statistical power in psychological science, starting with Cohen (1962), the overall power in psychology remains to be quite low in the present. Cohen (1962) estimated the average power of 70 articles featured in the *Journal of Abnormal and Social Psychology* to be less than 0.5 for medium effects and less than 0.2 for detecting small effects. Thus, for moderate effect sizes, there was less than a 50-50 chance of obtaining a positive result when there truly was an effect present. Almost 40 years later, Bezeau and Graves (2001) found a very similar (0.5) average statistical power for medium effect sizes in their analysis of 66 research papers in the field of neuropsychology, indicating that Cohen's warnings did not lead to significant changes in research practices. Hartgerink et al. (2017) offer evidence indicating that about two in three published psychological research papers contain at least one false negative finding.

## 1.2 The taxonomy of scientific misconduct

Steneck (2007) distinguishes between research fraud and questionable research practices (QRPs). While the severely stigmatized behaviors that comprise research fraud – falsification, fabrication and plagiarism[2] – are hardly ever committed unintentionally and without knowledge of their consequences, questionable research practices (first classified as a form of scientific misconduct in the Netherlands Code of Conduct for Research integrity) are often more ethically ambiguous. They usually arise from the combination of analytical flexibility and researcher confirmation bias. Researchers often use QRPs without any bad intentions, and sometimes QRPs are even sanctioned or encouraged in research training

[2] U.S. Office of Science defines fabrication as „making up data or results", falsification as „manipulating research materials, … changing or omitting data or results so that the research is not accurately represented in the research record" and plagiarism as „the appropriation of another person's ideas … without giving appropriate credit".

(Kerr, 1998). Nonetheless, the use of QRPs in psychological research leads to outcomes that are harmful for the progress of scientific knowledge, such as inflation of false-positive results (Simmons et al., 2011). This can in turn result in major waste of researchers' time and resources, as well as undermining of the credibility of whole scientific fields.

Two main categories of questionable research practices can be derived from the literature: *p*-hacking (together with selective reporting of analyses and variables) and post-hoc hypothesizing (also known as HARKing, or Hypothesizing After the Results are Known). Each of these two categories represents a violation of a different part of the process of hypothetico-deductive scientific inference, although the lines between them are often fuzzy (e.g., HARKing can be considered a form of selective reporting).

### 1.2.1 P-hacking

Despite its arbitrariness, the 0.05 significance threshold α is ubiquitous in psychological research. It often serves as a decision criterion not only for inferences regarding whether the studied effect is present or not, but also for decisions about whether a given study will be published (for discussion on publication bias and psychology's aversion to the null see chapter 3.1). In addition to these pressures to publish, some researchers might be influenced by confirmation bias. Researchers usually hold prior beliefs about the truthfulness of their theories and hypotheses, which have led them to decide to study the particular phenomena in the first place (Bishop, 2020). Considering these two factors (pressures to publish and confirmation bias), it is not far-fetched to suggest that researchers might be inclined to disregard the analyses that lead to insignificant results as irrelevant. As Simmons et al. (2011) point out, there is a great amount of flexibility (researcher degrees of freedom) present at various stages of behavioral data analysis. P-hacking is an umbrella term used to describe a set of analytical decisions used by researchers, whether intentionally or not, to exploit the researchers' degrees of freedom to obtain statistically significant results from their data.

#### Optional stopping (data peeking)

In the NHST framework, the probability of mistakenly rejecting the null hypothesis approaches 1 with increasing sample size when the researchers carry on with testing statistical significance after adding new observations, only stopping the data collection when

the *p*-value falls below the significance threshold (Mayo, 2018). Even when done without bad intentions and sporadically (e.g., after each 20 new observations), this form of data peeking leads to significant inflation of Type I errors. To prevent this, Simmons et al. (2011) recommend for authors to "decide the rule for terminating data collection before data collection begins and report this rule in the article". In specific cases, though, analyzing the data continuously and stopping after reaching significance can be considered a time- and cost-effective way of conducting research. Using sequential analyses is advised for researchers that want to reap those benefits while controlling for Type I errors (Lakens, 2014).

**Selective reporting of dependent variables and experimental conditions**

Simmons et al. (2011) warn about the increased probability of obtaining a false positive result caused by the selective reporting of dependent variables or whole experimental conditions in psychological research. According to their simulation study, the false-positive rate doubles (at 0.05 significance level) when two correlated dependent variables are used for analyses. In a similar manner, selective reporting of experimental conditions is deemed to be problematic in their paper. Additionally, the probability of obtaining false-positive results increases to an alarming extent when these forms of exploiting analytical flexibility are combined with optional stopping.

### 1.2.2 HARKing

Falsification of a priori hypotheses by data is an essential part of the hypothetico-deductive approach adopted in the majority of psychological research (Murphy and Aguinis, 2019). Hypothesizing after results are known (HARKing), i.e., presenting post-hoc hypotheses based on the observation of data as a priori hypotheses, is antithetical to the principle of falsification. Hypotheses derived from already available observations cannot be possibly disproved by the same observations. Therefore, HARKing can lead to the reification of Type I errors into scientific theories (Kerr, 1998). As such, HARKing is believed to be one of the driving factors of the replication crisis (Rubin, 2017), as well as the theory crisis, the perceived lack of formalized theories in psychology (Oberauer and Lewandowsky, 2019).

In their survey among American research psychologists, John et al. (2012) report that 30% of their sample admitted to "reporting an unexpected finding as having been predicted from the start" at least once. The discussion about HARKing in psychology gave rise to the debate about differentiating between confirmatory and exploratory analyses in psychological science (Wagenmakers et al., 2012).

Rubin (2017) offers a more nuanced classification of HARKing. He distinguishes between *constructing hypotheses after results are known*, i.e., proposing hypotheses designed to accommodate the observed results, and *suppressing hypotheses after results are known*, i.e., not reporting those a priori hypotheses that were not supported by the results. *Constructing hypotheses after results are known* can lead to the production of ungeneralizable hypotheses that are overfitted to the peculiarities of the dataset itself (instead of the population effect). *Suppressing hypotheses after results are known* can cause bias in estimation of effect sizes in meta-analytic studies. Moreover, Rubin recognizes a distinct category of *retrieving hypotheses after results are known* (RHARKing): testing hypotheses devised by other researchers in the literature after seeing the results. He claims that RHARKing does not pose a threat to science (as compared to the other forms of HARKing). Rubin also notes that HARKing can take on a passive form when it is carried out as a reaction to requests arising in the review process (for example, due to the editorial emphasis on clean narrative in scientific papers).

## 1.3 The prevalence of questionable research practices in psychology

Given its delicate nature, measuring the prevalence of questionable research practices in scientific subfields is a difficult undertaking. Nevertheless, there is a wealth of studies that aimed to evaluate the frequency of this phenomenon using self-report questionnaires administered to the researchers themselves.

In one of the first attempts to assess the prevalence of QRPs in psychology using a self-report questionnaire, John et al. (2012) surveyed over 2,000 US research psychologists about their self-admitted rate of QRP use and their beliefs about the prevalence of QRPs in their respective fields. Questions about ten different QRPs were included in their survey,

ranging from failing to report all examined conditions or dependent measures in a study to downright fabrication of entire datasets. More than 90% of the researchers admitted to having used QRPs at least once in their career. Majority of the respondents admitted engaging in selective reporting of studies and dependent measures, optional stopping and post-hoc exclusion of data, with about 1% admitting to having falsified their datasets for at least one research article (John et al., 2012). This led the authors to conclude that some QRPs have become a scientific norm, sometimes even necessary to thrive in the ecosystem of academia. Furthermore, those respondents who admitted to using QRPs were more likely to believe that the use thereof was ethically defensible. Both the self-admission rate (40%) and the defensibility ratings were the highest in the sub-group of social psychology researchers. Researchers in the field of clinical psychology had the lowest self-admission rate (27%).

Research misconduct self-admission rates very similar to those of John et al. were found in a survey study of Italian research psychologists (Agnoli et al., 2017), hence suggesting that the use of questionable research practices might be of concern in academia outside of the US as well.

The use of survey methods for measuring prevalence of questionable research practices is not without limitations, especially when we consider the severity of claims proposed by John et al. On the one hand, reliance on self-report of socially stigmatized behavior possibly leads to under-reporting of true prevalence. The surveys about QRPs that contain the words “fabrication” or “falsification” usually lead to lower self-admission rates (Fanelli, 2009). On the other hand, the inferences made by John et al. (2012) are considered to be too far-reaching by some, potentially overestimating the rates of QRP use in psychology.

According to Fiedler and Schwarz (2016), deriving literature-wide prevalence of QRPs from the proportion of researchers who admitted to having participated in such practices *at least once* is problematic and leads to wrong conclusions. Fiedler and Schwarz also voiced their concerns about the wording of some items used by John et al. The formulation of some of these items can lead researchers to “admitting” using some research practices while not actually having participated in their problematic forms. This can cause further overestimation of QRP prevalence rates.

The lower bound of prevalence of more severe forms of scientific misconduct, such as data fabrication, can also be estimated by considering the number of retracted papers in a given scientific field. Although the retraction rates in psychology do increase over time, with first retractions appearing in the late 1990's (Stricker and Gunther, 2019), this increase can be explained by the advances in the methods of scientific misconduct detection and in retraction procedures (Fanelli, 2013). According to Stricker and Gunther, 0.82 per 10,000 psychological journal articles were retracted because of scientific misconduct. Social psychology emerged as the sub-discipline with most (3.80) retractions per 10,000 papers (Stricker and Gunther 2019). This might be partially due to the high-profile cases of data fabrication (such as that of Diederik Stapel) in social psychology.

## 1.4 Detection of scientific misconduct from patterns of *p*-values

The most reliable way of detecting scientific misconduct or the presence of questionable research practices in individual studies is the inspection of full datasets used in the original studies. In some circumstances, though, such data are not available. Several methods of estimating the presence of questionable research practices or scientific fraud from reported data were developed. One such approach, used quite widely in the recent decade, is based on examining the reported *p*-values in specific subsets of literature.

Given the patterns of *p*-values under the null and alternative hypothesis (as described in subchapter 2.1.1), if there were no QRPs used in the analyses, the distribution of observed *p*-values across a given subset of literature is a monotonically decreasing combination of uniform and right-skewed distributions (Hartgerink et al., 2016). When the assumption of the absence of QRPs (such as optional stopping) is violated, a left skew just below the significance threshold (usually 0.05) can emerge (Hartgerink, 2015). The aim of this chapter is to introduce the discussion about whether such "bump" in *p*-values is present in psychological literature and about its possible interpretations and possible explanations.

Masicampo and Lalande (2012) examined more than three thousand *p*-values between .01 and .10 extracted from 36 issues of three psychological journals. Consistent with previous research, the distribution was right-skewed. Potentially more surprisingly, they

found a bump / local left-skew in the "barely significant" region of .045 – .05. They proposed publication bias and specific forms of p-hacking (repeated data peeking and optional stopping) as possible causes of this pattern.

Other researchers followed the suit, with Head et al. (2015) using text-mining techniques to obtain a very large (n = 285,050) set of reported *p*-values from diverse sets of scientific disciplines. Their results seemed to confirm those of Masicampo and Lalande. There was a left skew just below the 0.05 cut-off (used as a significance level in most of the analyzed papers). The authors consider the „bump“ to be direct evidence of p-hacking, which potentially also endangers the validity of meta-analytic outcomes. They argue that while the use of *p*-hacking techniques in individual papers is not likely to change the qualitative conclusions (i.e., the binary decisions about the existence of the studied effect) in most meta-analyses, it might result in the inflation of estimated effect size.

Hartherink (2015) reanalyzed the data used by Head et al., stating that the gravity of their conclusions demands for the results to be robust. Considering that Head et al. used text-mining procedure to extract the reported *p*-values, the *p*-values could not be recalculated from other reported data (such as test statistics or degrees of freedom). Hence, a large portion of the analyzed *p*-values were only reported as two-decimals, potentially biasing the analysis. Taking the reporting tendencies into account, Hartgerink used different bin sizes more suited for the character of the data. Contrary to the results reported by Head et al., in Hargerink’s more fine-grained reanalysis no evidence for a left-skew that could indicate the use of p-hacking was found.

On a similar note, after comparing the *p*-values reported in two psychological journals in 1965 and 2005, Legett et al. (2013) concluded that the prevalence of *p*-values just below 0.05 (as compared to *p*-values above 0.05) has increased over time. They attributed this tendency to the relative straightforwardness of statistical analyses in 2005 (as compared to the hand-computed *p*-values used in 1965), which can potentially facilitate p-hacking, as well as to the “shifting research climate”.

According to Lakens (2015b), the change of ratio of *p*-values just under and just above the significance level found by Legett et al. (2013) is caused by a relative decrease of the

frequency of *p*-values just above 0.05, rather than by increase of barely significant *p*-values, as suggested by Legett et al.

Lakens proposes the increase of publication bias over time and change in average statistical power as sources of the disparity that are more probable than the rise of p-hacking. Echoing the reservations stated by Gelman and O'Rourke, Lakens concludes that most of the attempts to find evidence for increase in QRPs over time using large sets of reported *p*-values are rather futile, as his proposed explanations are at least as plausible.

In an attempt to avoid some of the limitations present in the previous studies, Hartgerink et al. (2016) analyzed a set of more than 250,000 statistical test results in psychology. They distinguished between the *p*-values reported in the original articles and the *p*-values recalculated from reported test statistics. Although the bump caused by barely significant *p*-values was present when only analyzing the reported *p*-values, it disappeared after using the recalculated *p*-values. Perhaps more interestingly, strong evidence for incorrect rounding of *p*-values was found after comparing the reported and recalculated *p*-values. 67.45% of the *p*-values reported as exactly 0.05 were found to be larger than 0.05 after recalculation (Hartgerink et al. 2016). Thus, incorrect reporting, rather than other QRPs such as repeated peeking, might account for the increase of barely significant reported *p*-values.

## 1.5 Factors associated with scientific misconduct

The incentive structure of academic research, set to disproportionally reward novel positive results over null (albeit often more methodologically sound) findings is, together with publication pressure, one of the more talked about factors associated with QRPs and scientific misconduct. These factors and their possible association with research misconduct and QRPs will be described in more detail in the following chapters. This subchapter aims to examine the literature for the possible predictors of research misconduct on the level of individual differences between researchers.

### 1.5.1 Gender

Fang et al. (2013) found male researchers to be overrepresented (disproportionately to the ratio of men/women in given scientific subfields) in scientific fraud cases in American life sciences. This can be ascribed either to the differences in cultural norms for men and women or, potentially, to a smaller chance of scientific misconduct attempted by a woman being detected (Kaatz et al., 2013). Conversely, in a study of factors associated with fraudulent image duplication in biology, women were found to be more likely to partake in this particular sort of scientific misconduct scientists (or more likely to be found out).

No gender differences in Research Misbehavior Severity Score were found in a survey study examining a sample of PhD researchers, conducted by Holm and Hofmann (2018), nor was gender associated with retractions of scientific papers due to misconduct (Fanelli, 2015).

### 1.5.2 Personality factors and personal motivations

Machiavellianism, one of the Dark Triad personality traits, was found to be positively associated with self-reported research misbehavior and QRP use in a sample of Dutch biomedical scientists (Tijdink et al., 2016). No associations were found between research misbehavior and the other measured personality traits (Narcissism, Psychopathy and Self-esteem) in the same study.

On the contrary, Janke et al. (2019) argue that the engagement in QRPs can be better explained by fluctuating and situational influences, such as personal motivations, rather than stable personality factors. In a sample of research psychologist, they found the desire for outward displays of competence *(appearance goals)* to positively predict the use of QRPs. On the other hand, the desire to develop own competencies *(learning goals)* predicted the researchers' participation in QRPs negatively. Janke et al. thus propose promoting learning goals and the aspirations to develop own competencies as a possible way of reducing the prevalence of QRP use.

### 1.5.3 Age and career stage

A negative association between age and scientific misconduct score was found in a survey study of health professions education researchers (Maggio et al. 2019). Relatedly,

early-career, non-tenured scientists were more likely to admit scientific misconduct than their more established counterparts in the same study. The authors provide two alternative explanations for this difference: either the junior scientists' careers are more dependent on publishing positive findings which makes the use of QRPs more appealing to them, or they are not sufficiently familiar with responsible research conduct.

In a survey study carried out by Tijdjink et al. (2016), higher academic positions were associated with higher Research Misconduct Severity Score. A difference in research misconduct rates was found between tenured professors and postgraduate students, while there was no significant difference between professors and postdoctoral researchers.

### 1.5.4 Publication rate

In a study by Maggio et al. (2019), the scientific misconduct scores were higher for authors who published more papers during their lifetimes (as compared to their less prolific peers). On the contrary, Fanelli et al. (2015) found the authors with more published papers to be less likely to have a paper retracted. Only one retracted paper per author was randomly selected for the study to eliminate the prolific retractor bias (i.e., the bias induced by the influence of researchers with large numbers of retracted paper) in their study.

### 1.5.5 Personal attitudes towards scientific misconduct

According to John et al. (2012), the psychologists who rated certain QRPs as ethically defensible reported higher self-admitted QRP rates. Likewise, positive attitudes toward research integrity were negatively associated with research misconduct in doctoral students in Holm and Bofmann's (2018) study.

### 1.5.6 Geographic region

The researchers' country of residence was identified as a predictor for retraction. Papers authored by researchers from Australia, China, South Korea and Turkey were more likely to be retracted than papers from American authors (Fanelli et al., 2015). Maggio et al. (2019) found a similar pattern in their study, with researchers in Asian countries scoring higher on the Research Misbehavior Severity scale than their North American colleagues.

### 1.5.7 Limitations

As obvious from the previous paragraphs, the question of what factors are associated with research misconduct (and what the directions of such associations are) is far from resolved. Moreover, it must be noted that the operationalizations of scientific misconduct differed notably across the aforementioned studies, ranging from the self-admitted rate of QRP use to considering the already detected cases of scientific fraud. Both the self-report method and using retracted papers or detected cases of scientific misconduct come with their own problems. In the former case, it is possible that we are correlating honesty about QRP use (rather than the QRP use itself) with the potential predictors. Using the latter method will potentially inform us about the correlates or predictors of the more careless forms of research misconduct (as the more sophisticated attempts at research fraud are more likely to go undetected).

Despite the limitations of the studies both on the prevalence and the associated factors of questionable research practices, it remains clear that advancing the knowledge about the systemic and individual factors of researcher behavior is necessary for both the prevention of QRPs and the promotion of more robust research methods (Wass et al., 2019).

# 2 Structural factors associated with scientific misconduct

In the previous chapter, I have summarized the current discussion on questionable research practices in psychology and described the factors associated with research misconduct on the level of individual researchers. Although these individual factors are important to understand in order to promote more robust and reproducible research practices, they cannot be grasped properly without examining the structural factors that influence researchers' behavior significantly – the incentive structure of academia.

In this chapter, I am going to describe how institutional and career pressures and incentives impact research behavior and methodological choices. Special emphasis will be put on pressures to publish, publication bias and the competitive character of scientific research in psychology as potential causes of irresponsible research practices that drive the ongoing crises in psychological science. In the second part of the chapter I am going to focus on the way in which the wider incentive structure can render irreproducible research practices a rational, reward-maximizing strategy.

## 2.1 Publication bias

Many of the reproducibility- and generalizability-related problems facing psychological research can be tracked down to the existence of publication bias – the selective publication of significant or novel results which is influences researchers' decisions at all stages of research process. First, I am going to put publication bias into a broader context of pressures to publish and psychology's aversion to the null. Then I will proceed to propose a definition of publication bias itself, explore its different types and summarize the evidence on the associations between publication bias and questionable research practices.

### 2.1.1 Pressures to publish

The academic job market is a very precarious one. Research job positions are scarce, with the number of PhD graduates far exceeding the number of available academic job openings (Larson et al., 2014). Scientific productivity is one of the most influential criteria for selection for research-related job openings. It is usually assessed by quantitative indices mostly based (in one form or another) on the number of research papers published by a given

researcher, number of citations these papers received (or a combination of both, which is the case of the widely-used *h-index*), and the impact factor of the journals they were published in (Kun, 2018) – which is, in turn, also based on the number of citations received by papers published in the particular journal (Garfield, 1999).

Thus, research scientists and their careers benefit directly from publishing, and are endangered when a certain number of publications is not reached, as reflected in the well-known adage *"publish or perish"*. Moreover, the advantage gained early in scientists' careers by publishing in journals with high impact factor is cumulative (Allison et al., 1974), as the initial edge later translates into greater chance of getting and keeping an academic job or receiving research grants (Cole and Cole, 1967).

In that manner, it is not surprising that researchers perceive pressures to publish. In their worldwide survey of demographers, van Dalen and Henkens (2012) found that a majority (ranging from 52% in developing countries to 74% in the US) of research scientists in that field regard the institutional pressures to publish as high. Similar results were obtained in a survey of Dutch biomedical scientists (Tijdink et al., 2013), with 54% agreeing that "the pressure to publish has become excessive", and 38% agreeing that "publication pressure causes serious doubts regarding the validity of research results".

### 2.1.2 Definition of publication bias

The process of deciding which papers are going to be published is influenced by publication bias in both the pre-submission and the peer-review stage. Put simply, publication bias is the disproportionality in the chance of getting published between research papers, based on criteria of significance (i.e., whether the results supported the researchers' hypotheses; Rosenthal, 1979), novelty, and other factors.

Publication bias is present on both the institutional (journal) level, and on the level of individual researchers. On the journal level, non-significant (null) findings and replication studies are less likely to be accepted for publication. This can be in part attributed to problems with interpretation of non-significant results in the NHST framework (Cohen, 1994). Additionally, as the journal impact factor depends on citations received by individual articles and papers with null results are more likely to be cited (Duyx et al., 2017), journal editors might be hesitant to publish papers in which the hypotheses were not supported by

the data. On the individual researcher level, the scientists can be less likely to write up and submit their negative results for review, either because they presume that they will not be accepted, or because the results potentially disprove their own theories (Ferguson and Heene, 2012).

### 2.1.3 Bias towards positive results and the aversion to the null in psychology

The proportion of reported positive results (i. e. papers in which the main null hypothesis was rejected) in the literature has been steadily increasing in most scientific disciplines in the last few decades (Fanelli, 2012). There is evidence suggesting that psychology is the field with the highest ratios (up to 91%) of positive results literature-wide (Fanelli, 2010, but see Pautasso, 2010). While the relative increase of positive results in the literature can be explained by an increase in statistical power in papers across scientific fields, the evidence about the average statistical power in social sciences and psychology (see, for example, Szucs & Ioannidis, 2017) suggests otherwise. According to Fanelli (2012), the most plausible explanation is that null findings, in fact, have a lesser likelihood of being submitted and accepted for publication. This can lead to null results either ending up in the file-drawer (as described by Rosenthal, 1979) or being transformed into significant results via p-hacking or HARKing.

Sure enough, negative results from methodologically sound and adequately powered studies are as important for the process of incremental science as positive findings are (Munafo and Neill, 2016). The rising appreciation of the importance of null results led to the introduction of new publication formats, such as Registered reports (Chambers et al. 2015), allowing for good quality papers to be published regardless of the significance of the results.

Additionally, the publication bias against negative results and replication studies can protect theories (which can possibly be based on inferences from false-positive results) from severe tests and possible falsification, rendering said theories practically invincible. Ferguson and Heene (2012) describe how publication bias can help theories to survive attempts at their falsification, allowing for them to stay well-regarded in the public discourse despite the evidence disproving them, and eventually accumulating in a “vast graveyard of undead theories”.

### 2.1.4 Other forms of publication bias

Although publication bias is often thought to be dependent on statistical significance only, it can be influenced by other study characteristics. The probability of a study being published can be positively affected by the novelty of the research methods / hypotheses, or, on the other hand, the study's conformity to prevailing social preferences.

**Publication bias towards novel findings**

Novelty is an important factor in publication decisions (Chambers, 2017, but see Wang et al., 2017). Seemingly groundbreaking results are more likely to be published than results of replication studies, with some journals going as far as including the requirement of novelty in their official policies (Chambers, 2017). This demand for originality is reflected in the increase in the use of words describing novelty in scientific papers – the relative frequency of words like "innovative" or "unprecedented" in abstracts in the PubMed database has increased literally by several orders of magnitude in the 40 years between 1974 and 2014 (Vinkers et al., 2015). The construction of scientific knowledge relies (or is thought to rely) on replications and other forms of incremental research. Under bias towards novelty, replication studies have a higher chance of staying in the file-drawer. The emphasis put on novelty of results is thus detrimental to the accumulation of scientific knowledge and possible self-correction of science (Chambers, 2017).

**Conformity publication bias**

Conformity publication bias is defined as selective publishing of studies that are aligned with dominant views or social preferences held in the given scientific discipline at the time (and selection against the studies that dispute these widely-held views, regardless their statistical significance; Coburn and Vevea, 2015). In a way, conformity publication bias can be viewed as an opposing force to publication bias towards novelty (as the former favors counterintuitive theories and results, while the latter puts results that support the epistemological status quo at advantage).

### 2.1.5 Publication bias and questionable research practices

The publication bias against null results, in synergy with pressures to publish, can lead some researchers to exploit the analytical flexibility inherent in (non-preregistered) behavioral data analysis (Simmons et al., 2011) to transform their negative results into significant ones (Agnoli et al., 2017; Fanelli, 2012). Necker (2014) found a positive association between perceived publication pressure and the self-admitted use of questionable research practices such as selective reporting or optional stopping (with the use of QRPs operationalized as the researchers having *ever* used the given QRPs). Similarly, publication bias and pressure to publish (together with the aesthetic demand of clean narrative in the paper) were the two most frequent justifications for participating in QRPs in a survey study conducted by Fraser et al. (2014).

### 2.1.6 Consequences of publication bias

Even in ideal conditions (sufficient statistical power, sound methodology), it is usually advised against drawing inferences from single studies. Given that the chance of finding a significant result (when there indeed is some real effect) is equal to correctly predicting a result of a coin toss when statistical power equals 0.50, and the average statistical power is estimated to be closer to 0.35 in psychology (Bakker et al., 2012), researchers, practitioners and policy-makers often turn to evidence synthesis methods such as literature reviews and meta-analyses. When publication bias is present in the literature (that is, when only a biased subset of all studies conducted on a given topic is available to meta-analysts, when null results are less likely to be published and replications are discouraged), no sound conclusions can be drawn even from meticulous meta-analyses (Amrhein et al., 2017). Extreme, unprecedented results are usually preferred for publication in journals. At the same time, the results that suggest large effect sizes are the most likely to overestimate the real effect (or to indicate the existence of an effect when there is none). The consequences of this phenomenon, coined as *winner's curse* by Young et al. (2008), are further enhanced by the fact that effect sizes are associated negatively with sample sizes in most meta-analyses (Levine et al., 2009). The parameter estimates obtained from studies with smaller sample sizes are naturally less precise, therefore exhibiting greater variance.

Consequently, the estimated effect size can drastically underestimate or overestimate the magnitude of the real effect. Under publication bias, only the studies that overestimate the real effect are published, therefore leading to the inflation of effect size estimations.

As such, publication bias might render the majority of meta-analytic effect size estimations uninterpretable (Ferguson and Heene, 2012), and make the decisions and policies derived from meta-analyses costly mistakes based on Type I errors. Although different techniques to assess and correct the influence of publication bias in meta-analyses were developed (for an overview, see Marks-Anglin and Chen (2020)), they usually suffer from low statistical power to detect it (Ferguson and Brannick, 2012). Additionally, efforts to include unpublished studies and gray literature in meta-analyses are being made. The extent of unpublished studies that can be tracked down and included is very limited, which causes further bias in meta-analyses that include them (in comparison to meta-analyses that included published papers only; Ferguson and Brannick, 2012). In Tsuji et al.'s (2020) analysis of 20 meta-analyses in psychology, 70% included unpublished data, with an average of only 11% of datapoints coming from unpublished papers and gray literature. It should be noted, though, that even the inclusion of all unpublished bias will not make it possible to correct for publication bias entirely. As researchers are usually mindful about the presence of publication bias, they might be disincentivized to even write up their null results into a research paper – or, alternatively, after considering the influence of publication bias, the researchers can be inclined to exploit their analytical flexibility to transform their null findings into positive ones (as described in subchapter 2.1.5).

## 2.2 Competition in psychological research

Scientific research is an inherently competitive undertaking. As such, competition is one of the dominant features of the incentive structure of scientific research. Prestige, research funding and job positions are scarce, and a disproportionate amount of credit is being given to researchers that are the first to make certain discoveries. This leads Fang and Casadevall (2015) to describe the economics of scientific research as *winner-take-all*. Obviously, apart from its numerous downsides, there are many benefits of competition in psychological research as well. Competition can motivate research scientists to perform

better in order to outperform their rivals. Theoretically, competition might also motivate researchers to be more critical to their colleagues' work, thus accelerating the process of self-correction in science (Fang and Casadevall, 2015).

Goodstein (2002) describes the shift in the character of competition of science from "purely intellectual competition" to "an intense struggle for scarce resources" and asserts that this change will probably have negative effects on researchers' ethical behavior, including research fraud and dishonest peer-review. Goodstein's concerns are supported by empirical findings. In their focus-group interviews with fifty-one American researchers, Anderson et al. (2007) found six negative consequences of competition to emerge in the researchers' responses. The respondents reported being less inclined to openly share their data or methods as a consequence of competition in science; they spoke about being concerned about scientists sabotaging others' work and interfering with the peer-review process; deformation of relationships, questionable research conduct and strategic game-playing.

The strategic game-playing view of scientific research is a very important one. However idealistic the individual researchers might be about the purpose of their research undertakings, the institutional and systemic pressures might lead them to focus on the maximizing of rewards in adherence to the prevailing incentive structures of psychological research.

## 2.3 Models of researchers' behavior

### 2.3.1 Principal-agent models of scientists' behavior: economic approach

Following the arguments stated in the previous sub-chapter, academia can be (in an admittedly simplified way) conceptualized as an ecosystem in which scholars compete with each other for scarce resources (published papers and the job positions, grant money and prestige that publications can bring; Higginson and Munafo, 2016). Although it is crucial to not downplay researchers' intrinsic motives (such as the pursuit of reliable knowledge about the outside world), even the most idealistic scholars need a job and research funding to be able to carry on with their noble goals.

Some of the main reasons of the replication crisis in psychology – namely deliberately conducting low-powered studies and using questionable research practices – can be attributed to researchers using their limited resources (time and money) in a strategic way (Bakker et al., 2012). In the incentive structure characterized by the competition resulting from the emphasis on high publication output, scientists strive to maximize the rewards obtainable by publishing a large number of positive findings. While conducting studies with high statistical power is crucial for scientific progress in psychology, high-powered research designs are also more time-demanding and costly (as statistical power increases with sample size). Hence, using limited resources to carry out few well-powered, large-sample studies might not be the best way to obtain a large number of positive findings (and, in turn, survive in the publish-or-perish ecosystem of academia). The choice of small sample sizes (and resulting overall small statistical power in whole bodies of literature) can thus be viewed as economically rational behavior (Braganza, 2019) – a result of scientists responding and adapting to structural incentives (Vankov et al., 2014).

In their simulation study, Bakker et al. (2012) show that performing several small, underpowered studies, pays off (in terms of "positive" results) more than performing one large study with adequate statistical power. This discrepancy is further accentuated when the researcher uses QRPs in the process of analyzing the underpowered studies. Similar results were found in Higginson and Munafo's study that identified conducting studies with statistical power between 10% and 40% to be the optimal strategy for obtaining as many positive results as possible (be they false or true positives). It is worth mentioning that the adaptation of research strategy to maximize the reward can happen non-consciously, without any intentions to deliberately cheat or use insufficient methods (Smaldino and McElreath, 2016).

This contrast between what is good for scientific progress (finding out true things about the outside world) and what benefits the careers of individual researchers (publishing as many positive results as possible, be they true-positives or not) leads some to describe the incentive system in academia as *perverse* (Edwards and Roy, 2017).

### 2.3.2 Principal-agent models of scientists' behavior: ethical approach

In contrast to the economic approach which views scientists as reward-maximizing rational agents adapting to incentive structures, the ethical approach to modeling researchers' behavior offers a different perspective. It conceptualizes researchers as agents who are invested in values that are aligned with the goals of science at large (such as honesty and reliability), whether adhering to these values is optimal in the current incentive system or not (Desmond, 2021). Apart from it being somewhat cynical, the economic approach is unable to explain certain phenomena occurring in the academic ecosystem – in particular, the default trust between scientists, the a priori good-faith attitude of scientists towards the motives of their peers. Default trust "cannot be justified in a culture of credit-maximization" (Desmond, 2021).

## 2.4 Qualitative evidence on incentive structure in academia

The principal-agent models paint a bleak picture of academia as a hypercompetitive ecosystem, with *publish or perish* pressures being slowly replaced by a system that makes individual researchers choose between *cheating or perishing.* However, the qualitative evidence based on interviews and focus groups with researchers themselves gives us reasons to be hopeful, at least in regard to individual scientists' motives.

Hangel and Schmidt-Pfister (2017) distinguish between epistemic, pragmatic, and personal motivations to publish. While epistemic motivations represent the (rather idealistic) focus on production of valid knowledge about the outside world, pragmatic motives are connected with the pressures of academia and involve strategic game-playing. Hangel and Schmidt-Pfister identify four common motivational aspects in the qualitative analysis of answers from researchers across scientific fields and career stages: communicating interesting results to the scientific community; gaining personal reputation; enjoying the process of research itself; obtaining funding. They conclude that epistemic, pragmatic and personal motivations appear to be intertwined together in the researchers' responses, thus providing a more nuanced view about the conflict between idealistic approach to science and strategic game-playing.

The results from a qualitative study conducted by Bruton et al. (2020) offer a slightly different view on the relationships between researchers' personal motivations and the incentive structures in academia. The most often mentioned topic in the researchers' responses were publishing practices. The respondents mentioned the publication bias towards novel and positive findings as major obstacles in improving the reliability of science. Career-oriented pressures, paired with the inflation of publication expectations and the negative effect of the academic incentive structure on scientific integrity were mentioned often as well. One of the respondents stated that "when someone's whole future is dependent upon such publications, if QRPs will help produce those publications, I guarantee that those QRPs will continue." (Bruton et al., 2020), thus supporting the conclusions derived from the principal agent-based models and the notion that for the reliability of psychological science to improve, the incentive structure must change first.

# II Empirical part

## 3 Research questions

In the theoretical part, I have described the negative effect of questionable research practices on the reliability and credibility of psychological science. Additionally, I have explored the broader context of research misconduct in psychology and the associated individual and structural factors, focusing on the incentive structure of academic research.

Drawing from the evidence presented in the previous chapters, the aim of the present study was to assess the presence of statistical inconsistencies and irregularities in a subset of psychological studies using several data-forensic techniques (described in more detail in Chapter 5). For this purpose, I used an automatic screening procedure to obtain a set of more than 19,000 full-text psychological research articles published in the last 50 years. I manually coded a random sample of 4,094 articles obtained by this screening procedure. I then extracted and analyzed the reported baseline data from a set of 530 research papers that fulfilled the selection criteria of the present study and contained data from randomized controlled trials of psychological interventions.

The present study examined the following four research questions:

*RQ 1: Distribution of p-values from baseline balance tests:*

> Does the distribution of baseline balance *p*-values recomputed from reported pretest data follow a uniform distribution that can be expected under the assumption of simple random allocation of participants into groups?
> If not, what are the improbable patterns of the deviations from the expected distribution and what problems in the research process do they indicate?

*RQ 2: Insufficient variability in reported variance estimates:*

> Do the reported baseline data exhibit the natural degree of variability expected under random allocation?

*RQ 3: Variability of variance estimates combined with the distribution of p-values from baseline balance tests:*

> How probable is the combination of the reported baseline means and the degree of variability from each of the included studies under the null hypotheses of random allocation of participants into groups?

*RQ 4: The presence of statistically inconsistent reported means and standard deviations in discrete variables*

> To what degree are mathematically impossible reported data (i.e., reported means and standard deviations that cannot possibly occur in discrete variables with given sample size and number of items) present in the included articles?

I am going to describe the statistical hypotheses and assumptions of these specific research questions and methods in Chapter 5.

The present empirical part is a pilot study for a larger research project led by Dr. Ivan Ropovik at the Center for meta-research in education, Charles University.

# 4 Sample

## 4.1 Sampling procedure

A random sample of randomized controlled studies of psychological interventions written in English and published in peer-reviewed journals indexed in the Web of Science between 1970 and 2020 was used in the present study. The studies were identified using a pre-piloted Boolean search string (see below) in the Web of Science database.

The full Boolean search string used for the search in the Web of Science database: *(TI=((intervention OR treatment OR training OR therapy) AND (random* OR RCT OR controlled near/3 stud* OR controlled near/3 trial* OR controlled near/3 experiment* OR controlled near/3 design OR placebo OR "true experiment" OR "parallel group" OR CONSORT OR superiority OR non-inferiority OR noninferiority OR "equivalence trial" OR "equivalence study" OR allocat* near/3 conceal* OR allocat* near/2 blind* OR "by chance") NOT ({mg} OR 0mg OR 1mg OR 2mg OR 3mg OR 4mg OR 5mg OR 6mg OR 7mg OR 8mg OR 9mg OR quasi-experiment* OR quasiexperiment* OR "not random" OR "not at random" OR "not randomly" OR non-random* OR nonrandom* OR match* near/2 pair* OR MP-RCT OR matched-pair* OR permut* near/3 block* OR block* near/2 random* OR adaptive near/2 random* OR permuted-block OR stratifi* near/2 random* OR minimization OR minimisation OR restricted near/2 random*)) OR AB=((intervention OR treatment OR training OR therapy) AND (random* OR RCT OR controlled near/3 stud* OR controlled near/3 trial* OR controlled near/3 experiment* OR controlled near/3 design OR placebo OR "true experiment" OR "parallel group" OR CONSORT OR superiority OR non-inferiority OR noninferiority OR "equivalence trial" OR "equivalence study" OR allocat* near/3 conceal* OR allocat* near/2 blind* OR "by chance") NOT ({mg} OR 0mg OR 1mg OR 2mg OR 3mg OR 4mg OR 5mg OR 6mg OR 7mg OR 8mg OR 9mg OR quasi-experiment* OR quasiexperiment* OR "not random" OR "not at random" OR "not randomly" OR non-random* OR nonrandom* OR match* near/2 pair* OR MP-RCT OR matched-pair* OR permut* near/3 block* OR block* near/2 random* OR adaptive near/2 random* OR permuted-block OR stratifi* near/2 random* OR minimization OR minimisation OR restricted near/2 random*))) AND SU=psychology*

The first search yielded 26,846 articles. Using an automated, machine-based procedure, full-text versions of these were downloaded. To further select eligible articles containing randomized controlled studies, the full-texts of the downloaded articles were filtered using another Boolean search string:

*(pre-test OR "pre test" OR pre-intervention OR pre-intervention OR pre-treatment OR pretreatment OR pre-training OR baseline OR "base-line") AND (post-test OR "post test" OR post-intervention OR postintervention OR post-treatment OR posttreatment OR post-training OR baseline OR "base-line")*

After the second screening, 18,862 full-text articles remained as potentially eligible. Out of this sample, 4,094 randomly selected full-text articles were manually checked for inclusion into the study following pre-specified inclusion criteria. 531 of the articles were included for further coding, while 3,563 did not meet the inclusion criteria.

## 4.2 Inclusion criteria

The following inclusion criteria were used to decide whether a given article was included in the study:

1) The article was written in English and published in a peer-reviewed journal indexed in the Web of Science database.
2) The article examined an effect of a psychological intervention in a randomized controlled trial.
3) The study explicitly reported a simple random allocation of participants into groups. Studies that employed any sort of restricted randomization procedure were excluded from the present study.
4) The study reported means, variability measures and group sizes for baseline demographic variables (such as age) or pre-test data of psychological/behavioral variables. The studies that did not report these data for the full randomized sample (e.g., only reporting baseline data only for the participants who completed the treatment) were excluded from the present study.

## 4.3 Coding procedure

Following the sampling procedure and the assessment of eligibility for inclusion, the following data from all 530 eligible studies were manually coded for each reported baseline/pre-test variable: Reported means, standard deviations or standard errors, and the number of participants in each randomized group.

Reported data for baseline variables that were explicitly dependent on other reported variables (e.g., total scores when subscores were reported in the paper) were excluded. Additionally, the reported *p*-values for baseline balance tests, the number of randomization clusters, and the number of items for discrete variables were coded where available. In total, 9,067 reported group datapoints from 530 studies were extracted and analyzed.

# 5 Analysis

All statistical analyses were performed using R version 4.0.5 (R Core Team, 2021). The following packages were used for the analyses or data manipulation/vizualization: *tidyverse, magrittr, overlapping, entropy, metap, poolr, MBESS, CarletonStats, goftest, qqtes, dqsample, beepr, drat, ggplot2.*

## 5.1 Distribution of *p*-values from baseline balance tests

### 5.1.1 Carlisle's method

The primary method used to examine the inconsistencies in the baseline balance data (namely, the excessive similarity or dissimilarity in means) in the present paper is a modification of a statistical method first used by Carlisle (2017). The basic rationale of Carlisle's study is as follows: Under simple randomization of participants into groups, the distribution of p-values for baseline balance tests is a uniform (0, 1) one. This is due to the fact that when testing for baseline balance between randomized groups, the data-generating process is known – any possible differences between the groups are purely random (if simple randomization was indeed employed and it was used correctly). As the null hypothesis (no difference between the randomized groups) is always true in the long run, any significant test result is a Type I error. This can, of course, be expected from the definition of Type I error in NHST.

As follows from the previous paragraph, testing for baseline differences in randomized controlled trials is a pointless practice (Senn, 1994). Regardless of these criticisms, it is still widespread in various scientific fields, including psychology, and often even encouraged by journal editors. It is commonly (and erroneously) believed that testing for baseline differences is important to show whether the randomization truly "worked" (Senn, 1994). Consequently, researchers might be incentivized to demonstrate that the randomized groups were balanced, i.e., that there were no significant differences between the groups.

Carlisle used the fact that the distribution of *p*-values under *H0* is uniform in his 2017 study to assess the integrity of baseline data in more than 5,000 randomized controlled trials in the field of anesthesiology. He employed t-tests, ANOVAs and Monte Carlo simulations to recompute the *p*-values for baseline balance tests from reported data.

To combine the *p*-values from each trial (i.e., to compute the probability of observing the same or higher *p*-values under the null hypothesis), Carlisle (2017) used Stouffer's method. The overall trial-level *p*-value is computed using the following three steps:

1. The $Z_k$ score for each variable is computed (with ϕ representing the standard normal cumulative distribution and $p_k$ representing the *p*-value for k[th] baseline balance test in a given trial):

$$Z_k = \phi^{-1}(1 - p_k)$$

2. The overall trial-level Z-score, $Z_{tot}$ is calculated (with *b* representing the number of *p*-values that are being combined):

$$Z_{tot} = \sum_{k=1}^{b} Z_k \div \sqrt{b}$$

3. The overall study-level *p*-value is obtained by subtracting $Z_{tot}$ from 1 (Carlisle et al., 2015).

The distribution of overall (trial-level) *p*-values analyzed by Carlisle was found to be left-skewed at the higher end. This indicates an excessive similarity of means in the included articles (as compared to the expected cumulative uniform distribution). Additionally, a significant difference was found between the cumulative distributions of *p*-values from baseline balance tests reported in non-retracted papers and a sample of papers retracted due to scientific misconduct (Carlisle, 2017).

There are several limitations to Carlisle's method (apart from the possible errors that might have arisen during the coding process). While the baseline variables in the same sample might be intercorrelated (therefore having similar *p*-values in baseline balance tests), the Stouffer's method treats them as independent. This can potentially distort the results. Additionally, the assumption of the uniform distribution of *p*-values from baseline balance tests only holds true when simple randomization is used. Carlisle did not take this into account in his paper.

### 5.1.2 Modified version of Carlisle's method used in the present study

Considering the limitations stated in the previous subchapter, a modified version of the Carlisle's method was used in the present study as a primary method of assessing the distribution of baseline balance *p*-values. Only studies that used simple random allocation of participants to groups were included for data coding. Furthermore, the possible dependencies between the baseline variables were taken into account by analyzing the recomputed *p*-values for randomly permuted sets of individual variables alongside the overall trial-level *p*-values computed using the Stouffer's and Fisher's method (as done by Carlisle). Moreover, we accounted for possible intercorrelation in the analyses of the trial-wide overall *p*-values.

**P-value recomputation from reported data**

The reported measures of variability (standard deviations or standard errors) were corrected for small sample bias using the *s.u* function from the *MBESS* R package. The *p*-value for each baseline balance test was then recomputed from the reported means and *SD*s using Monte Carlo simulations with 10,000 iterations (a procedure described by Carlisle et al., 2015). Monte Carlo simulations were used instead of more conventional ANOVA models to compensate for the lack of precision in reported means and *SD*s. Simulating noise with known properties namely aims to account for situations in which the rounded means or *SD*s are numerically identical, which would lead to mathematically impossible *p*-value of 1.

**The distribution of *p*-values from individual variables**

To avoid the problems caused by the dependencies between the individual variables in a given study (present in the original Carlisle's method), we employed a permutation procedure to randomly select one variable (within 100,000 iterations) from each trial. Consequently, we obtained 100,000 columns with 554 randomly selected *p*-values in each (one *p*-value from each study/trial). An overall sample-level *p*-value was computed using Stouffer's and Fisher's method on each of the permuted sets of *p*-values. Under the assumption of truly random allocation of participants into groups in the original studies (and therefore the uniform distribution of baseline-balance test *p*-values), the distribution of these combined (overall sample-level) *p*-values obtained by Stouffer's and Fisher's method is uniform, with a mean of 0.5. We computed the mean and SD for the distributions of sample-

level *p*-values obtained by Stouffer's and Fisher's method and used Wald's *z*-test to examine whether the computed mean estimates are statistically different from 0.5.

**The distribution of combined study-level p-values**

Following Carlisle's protocol, we examined the distribution of combined study-level *p*-values. As an important addition to Carlisle's design, we accounted for dependencies between the individual variables in each study to avoid the inflation of the combined *p*-values caused by the dependencies between the variables. To achieve this, a correlation matrix among the one-tailed *p*-values within the given trial (with assumed intercorrelation of 0.3) was constructed for each trial (i.e., set of *p*-values to be combined). Based on this correlation matrix, the combined one-tailed *p*-value for each study was estimated using Stouffer's and Fisher's method.

Similarly to the analysis described in the previous subchapter, under the assumption of proper simple randomization in the original studies and the absence of publication bias and questionable research practices or data fabrication, the distribution of the combined *p*-values is a uniform (0, 1) one. When the density of observed overall *p*-values is greater than 1 on the lower end, there is an excess of studies in which the randomized groups are too similar in baseline variables, which might be indicative of data falsification or researchers using restricted randomization methods without disclosing it. On the other hand, an excess of observed *p*-values in the high-end as compared to the expected uniform distribution would indicate an excess of studies in which the differences between randomized experimental groups were too large in baseline variables, potentially indicating a lack of randomization or problems within the randomization process.

To assess the presence of indicators of statistical inconsistencies, we compared the observed cumulative distribution function of the distribution of study-level *p*-values with a simulated uniform theoretical. Anderson-Darling test (for higher power in the tails of the distribution) and Kolmogorov-Smirnov test (for higher power in the center of the distribution) were used for this comparison.

## 5.2 Insufficient variability in reported variance estimates

As an addition to the distribution of *p*-values from baseline balance tests, the variability of variance estimates (*SD*s) was assessed using simulations, following a procedure first used by Simonsohn (2013). A certain degree of variability in variance estimates (i.e., standard deviations or standard errors of measurement) can be expected between groups of the experiment that were obtained by random allocation. This holds true even when the population standard deviations σ are identical across groups, i.e., assuming a single underlying population from which the samples were drawn. (Simonsohn, 2013). This variability of variance estimates naturally decreases with increasing sample size. In some (hopefully rare) cases, insufficient variability in reported variance estimates can be indicative of data falsification or fabrication (as humans are usually not very skilled at generating random numbers or estimating the amount of variability in genuine datasets).

Monte Carlo simulations were employed to assess the probability of the variability of variance estimates from each study occurring under random allocation. As a first step, the standard deviation of standard deviations was computed for each baseline balance test. We then simulated 10,000 random samples drawn from a normal distribution for each of the baseline balance tests, using the actual group *N*s, group means, and pooled standard deviations as parameters for the normal distributions. We assessed the presence of adequate sampling variability (or lack thereof) by counting the number of simulated distributions that exhibited identical or lower variability in standard deviations (effectively computing a permutation-based *p*-value of observed data under randomization). A low probability would indicate a lack of variability between the samples as compared to the expected degree of sampling error.

As our main target, the combined, study-level *p*-value based on the analysis of standard deviation of standard deviations (one for each variable of the experiment) was combined with the study-level *p*-value based on the analysis of the group means. Again, Stouffer's and Fisher's methods were used to compute the combined *p*-values. These combined *p*-values were used as a primary indicator of inconsistencies in the baseline data.

## 5.3 Statistically inconsistent reported means and standard deviations

For baseline variables measured on a discrete scale (e.g., age in whole years/months, number of symptoms, or variables measured by Likert scales), the statistical consistency of reported means and SDs was assessed using GRIM and GRIMMER methods, respectively.

### 5.3.1 GRIM

The granularity-related inconsistency of means method (GRIM) is a mathematical technique for evaluating the consistency (or lack thereof) of reported means of discrete variables (Brown and Heathers, 2016). The GRIM method uses the granularity of data – "the numerical separation between possible values of the summary statistics" (Brown and Heathers, 2016) to assess whether the reported summary statistics are mathematically possible. To illustrate the general principle of GRIM on an example, let us consider a Likert-based variable consisting of 2 items administered to 7 participants. The smallest possible difference between two mean scores for this variable is 0.2867 (1 / (7/2)). When the means are reported to two decimal places, only some values (such as 1, 1.29, 1.57 and so on) of the mean can occur (i.e., are mathematically possible). The rest of the values in a given range are thus deemed mathematically inconsistent, because they cannot possibly arise from the given sample size and number of discrete items. As such, GRIM is an effective method that can (under some assumptions) provide us with clear evidence about the consistency of the reported data. Although the information about the consistency provided by GRIM is unambiguous, there are several different explanations for the detected inconsistencies, ranging from typographical errors, software bugs and severe rounding errors (although GRIM does account for possible rounding errors to some degree) to the fabrication of the reported means.

It is important to note that granularity decreases with increasing sample size and number of discrete items. If the variable was measured using multiple discrete items, the analytic sample sizes are effectively the product of the number of participants times the number of items ($n * k$). Therefore, GRIM is only effective for analytic sample sizes $n * k$ less than 100 when the means are reported to two decimals, or less than 1,000 when the means are reported to three decimals (Brown and Heathers, 2016).

Only the integer variables for which both the sample size and the number of discrete items was explicitly and precisely reported were analyzed with GRIM and GRIMMER.

### 5.3.2 GRIMMER

Granularity-related inconsistency of means mapped to error repeats (GRIMMER) employs a principle very similar to GRIM to detect mathematically impossible variance estimates (Anaya, 2016). Hence, GRIMMER was used in the present study to complement GRIM results with further evidence from mathematically inconsistent variability estimates.

# 6 Results

## 6.1 Distribution of *p*-values from individual baseline balance tests

Markedly more recomputed *p*-values from baseline balance test (8.92%) were significant at the 0.05 level than the 5% that would be expected under the null hypothesis of simple random allocation of participants into groups in each of the trials.

The mean Stouffer's *p* over the iterations ($M = 0.605$, $SD = 0.269$) was significantly different ($p = 0.0001$) from the expected mean of the theoretical uniform distribution (0.5). The same applied to the mean Fisher's *p* over the iterations with even more extreme results: $M = 0.761$, $SD = 0.228$ ($p < 0.0001$). These results indicate that the observed distribution of *p*-values from the individual baseline balance tests is markedly different from the theoretical uniform distribution that can be expected under the assumption of simple randomization of participants into groups in the individual studies. The distributions of both Stouffer's and Fisher's *p*-values are markedly different from the theoretical uniform distribution (plotted as a dotted line in *Figure 1,* next page). This provides us with some initial clues about the presence and character of inconsistencies or randomization issues that might be present in the coded sample of randomized controlled trials of psychological interventions. It appears that the proportion of baseline balance tests in which the compared groups were too different was markedly larger than would be expected under the assumption of simple randomization

of participants into groups. This might indicate problems with randomization procedures (or an absence thereof) in the population of psychological RCTs.

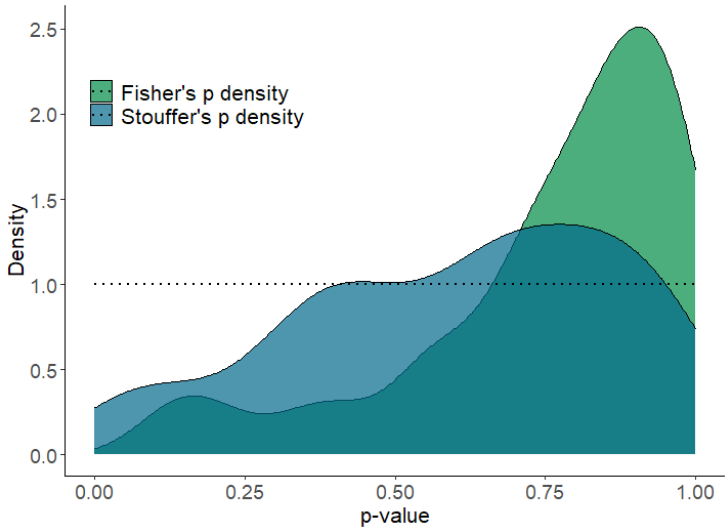


Figure 1 – the distribution of the 100,000 permuted *p*-values. Each of the *p*-values is a Stouffer's (blue) or Fisher's (green) combination of 554 randomly selected *p*-values (one *p*-value from each study/trial). High combined *p*-values represent an excess of baseline balance comparisons across iterations in which the two randomized groups were more different than expected under random allocation.

## 6.2 The distribution of combined study-level *p*-values

First, we compared the observed number of studies with the a with the probability *p* or less of the data arising from a simple randomization with the expected number of such studies from the theoretical uniform distribution. The results can be found in *Table 1*. Markedly more trials than expected under the null hypothesis (*H0*: the combined *p*-values arose from combining baseline balance tests that used simple randomization and thus have a uniform distribution) were found to have too dissimilar means. Hence, results can be interpreted as further evidence for the presence of systematic problems of randomization procedures used in the included articles.

| | *Stouffer's method* | | | | | *Fisher's method* | | | |
|---|---|---|---|---|---|---|---|---|---|
| *p-value* | < 0.0001 | < 0.001 | < 0.01 | < 0.05 | | < 0.0001 | < 0.001 | < 0.01 | < 0.05 |
| *Expected percent* | 0.01% | 0.1% | 1% | 5% | | 0.01% | 0.1% | 1% | 5% |
| *Observed percent* | 0.38% | 0.38% | 1.89% | 7.92% | | 0.19% | 0.57% | 1.51% | 6.42% |
| *Expected count (out of 530)* | 0.053 | 0.53 | 5.3 | 26.5 | | 0.053 | 0.53 | 5.3 | 26.5 |
| *Observed count* | 2 | 2 | 10 | 42 | | 1 | 3 | 8 | 34 |

Table 1

Accordingly, the distributions of combined Stouffer's and Fisher's *p*-values were found to be different from the expected distribution both by the Kolmogorov-Smirnov ($p = 9.2\times10^{-10}$ for Stouffer's *p*-values and $p = 6.7\times10^{-5}$ for Fisher's combined *p*-values) and Anderson-Darling ($p = 1.13\times10^{-6}$ for Stouffer's *p*-values and $p = 1.13\times10^{-6}$ for the distribution of Fisher's *p*-values) test.

The Q-Q plots comparing the distribution of Stouffer's *(Figure 2)* and Fisher's *(Figure 3)* p-values with the expected distribution, as well as the distribution density overlay plot *(Figure 4)* provide us with further evidence about the patterns the inconsistencies. When the full empirical distribution of *p*-values was overlaid with the theoretical uniform distribution, the overlap between these distributions was at 80.57% (estimated using a non-parametric bootstrap procedure; Pastore and Calcagnì, 2019). Again, a noticeable excess of *p*-values occurred in the higher end of the distribution (between 0.8 and 1). These irregularities in the distribution of combined *p*-value can be indicative of problems related to the randomization procedures, rather than some form of selective reporting, data fabrication or the use of questionable research practices.

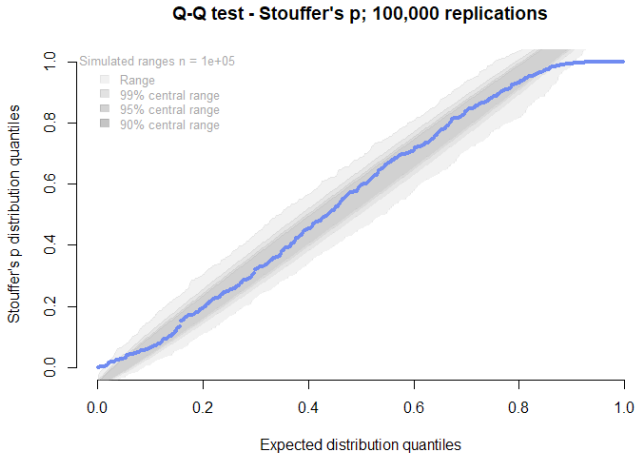


Figure 2 – Q-Q test comparing the observed distribution of Stouffer's combined *p*-values to the expected uniform distribution.

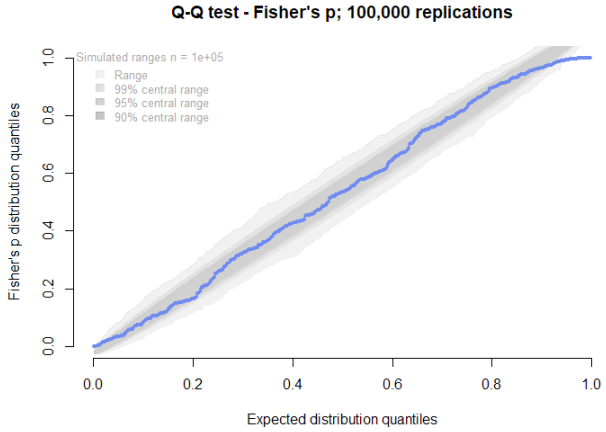


Figure 3 – Q-Q test comparing the observed distribution of Fisher's combined *p*-values to the expected uniform distribution.

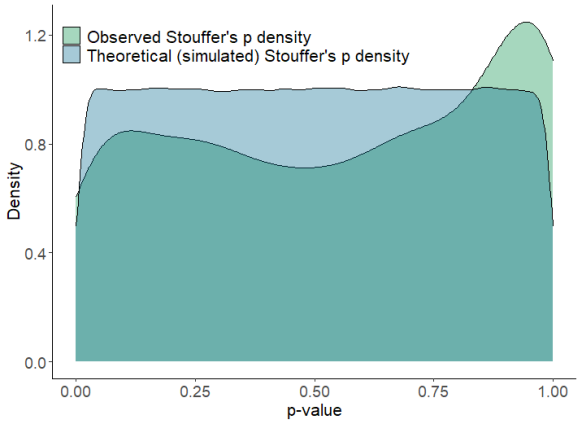


Figure 4 – Density plot overlay

## 6.3 Insufficient variability in reported variance estimates

As a target measure of data inconsistencies present in the included randomized controlled studies, we combined 1) the observed trial-level *p*-value based on the analysis of baseline means and 2) the probability of observing the reported or smaller SD) assuming random allocation. The two probabilities were combined using Stouffer's and Fisher's method for each of the included studies ($n$ = 530). The results are reported in *Table 2.* The interpretation of the results follows the same logic described in subchapter 6.2.

Both the results included in *Table 2* and the inferences drawn from the inspection of Q-Q plots *(Figure 5* and *Figure 6)* are consistent with the inferences stated in the previous subchapter. However, after combining the data on baseline means with the data on baseline *SD*s, it is more obvious that there is an overrepresentation of studies that exhibit very low variability in their baseline data. While the observed density of papers with insufficient variability is many times greater than the expected one in the < 0.0001 and < 0.001 intervals, studies with such lack of variability are still quite rare in the coded sample (there were two identified by Stouffer's method and five identified using Fisher's method at $\alpha < 0.0001$),

| | *Stouffer's method* | | | | | *Fisher's method* | | | |
|---|---|---|---|---|---|---|---|---|---|
| *p-value* | < 0.0001 | < 0.001 | < 0.01 | < 0.05 | | < 0.0001 | < 0.001 | < 0.01 | < 0.05 |
| *Expected percent* | 0.01% | 0.1% | 1% | 5% | | 0.01% | 0.1% | 1% | 5% |
| *Observed percent* | 0.38% | 1.89% | 5.1% | 11.1% | | 0.94% | 2.45% | 6.6% | 12.64% |
| *Expected count (out of 530)* | 0.053 | 0.53 | 5.3 | 26.5 | | 0.053 | 0.53 | 5.3 | 26.5 |
| *Observed count* | 2 | 10 | 27 | 58.8 | | 5 | 13 | 35 | 67 |

Table 2

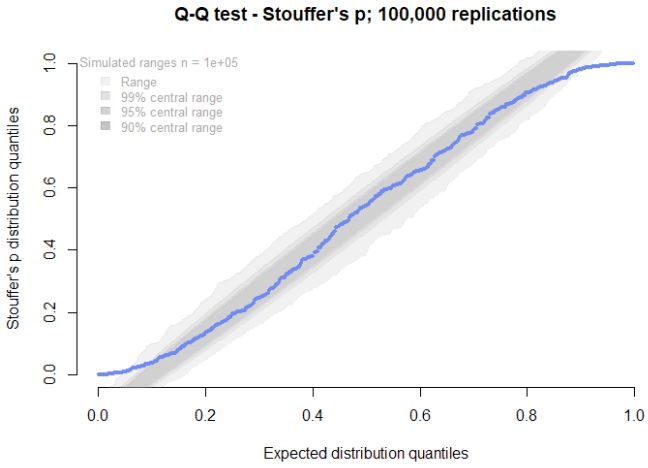


Figure 5 – Q-Q test comparing the observed distribution of Stouffer's study-level combined *p*-values (combining *p*-values from baseline balance trials and variance estimate simulations) to the expected uniform distribution.

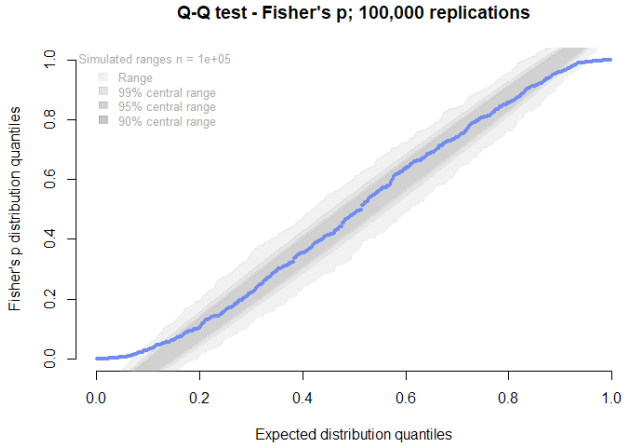


Figure 6 – Q-Q test comparing the observed distribution of Fisher's study-level combined *p*-values (combining *p*-values from baseline balance trials and variance estimate simulations) to the expected uniform distribution.

## 6.4 Statistically inconsistent reported means and standard deviations

For 2,294 out of the total 9,067 analyzed means and SDs, the variable was measured on a discrete scale and the number of items was explicitly reported. Out of the 2,294 reported means, 104 (4.53%) were detected as mathematically impossible by the GRIM test. Additionally, 58 (2.53%) reported standard deviations were found to be inconsistent by GRIMMER. In total, 162 reported datapoints (7.06% of 2,294) were contained mathematically inconsistent reported means and/or standard deviations.

Given that the sensitivity of GRIM and GRIMMER decreases with increasing sample size and number of discrete items in the variables from which the means and SDs were computed, these results can be considered to be the lower bound estimate of problems with the mathematical consistency of the reported data.

# Discussion and conclusion

## Discussion of the results

In his analysis of more than 5,000 randomized controlled studies in the field of anesthesiology, Carlisle (2017) found an excess of trials in which the compared groups were too similar. He interpreted this deviation from expected uniform distribution as a potential indicator of data fabrication or undisclosed non-random allocation. In our attempt to examine whether similar tendencies occur in psychological RCTs, we found rather strong evidence for an opposite pattern. The distributions of both the baseline balance comparisons for individual variables and combined study-level *p*-values indicated an excess of studies in which the randomized groups were too different from each other. Similarly to Carlisle (2017), we also found a moderate evidence for an excess of baseline balance tests in which the compared groups were too similar in the population of randomized controlled trials of psychological interventions. Moreover, similarly to Carlisle, we also found strong evidence for a lack of variability within the reported means or standard deviations. At 0.001 significance level, around 1 in 50 of the included studies exhibited concerningly low degree of variability. These inconsistencies can be attributed to several causes, ranging from simple reporting or coding errors, to undisclosed use of non-randomization allocation, to scientific misconduct and data fabrication. Either way, the discrepancies between the expected and observed distributions of *p*-values from baseline balance tests (whether individual or combined) were remarkably large, indicating the presence of problems in the randomization process which can possibly endanger the validity of inferences drawn from randomized controlled trials of psychological interventions.

As a whole, the overall distribution of the reported baseline means and standard deviations in randomized controlled trials of psychological interventions was found to be inconsistent with random allocation of participants into groups. While these results are in some ways analogical to those obtained by Carlisle (2017) in anesthesiology, we found the psychological studies to contain an excess of baseline means that were too different from each other, potentially suggesting problems with the randomization process. As no similar studies assessing the distributions of baseline data in psychology are available in the present,

further comparison with other evidence about this topic specifically in psychology is not possible at present.

The prevalence of mathematically impossible reported means and standard deviations is concerning as well – especially considering that given the limitations of GRIM and GRIMMER methods, the proportion of detected inconsistencies (7.06%) is just a lower bound estimate of the true prevalence of inconsistent reported data. Although the proportion of detected mathematically inconsistent reported data found in the present study is smaller than the one found by Brown and Heathers (2016) in a similar random sample of psychological studies, it is still problematic.

**Limits of the present study and further research**

As most statistical analyses were conducted using simulation methods with large numbers of iterations, the results should be quite robust. Nonetheless, given that a limited sample of the whole population of psychological randomized controlled trials was used in the present study, the results are only modestly informative. In the following large-scale study, far larger part of the entire population of available randomized controlled trials of psychological interventions (around two or three thousand articles) will be coded and analyzed. Moreover, further analyses will be conducted, including analyses of associations of data inconsistencies with factor such as first author gender and career stage, number of authors and number of study citations.

It should also be mentioned that in the process of manually coding thousands of variables, various mistakes could have occurred. Hopefully, the extent of such random mistakes was as little as possible.

**Conclusion**

In the present thesis, I have attempted to provide a synthesis of evidence on the individual and structural factors associated with scientific misconduct and the negative influence of questionable research practices on the reproducibility and reliability of psychological findings.

The distribution of baseline balance *p*-values was found to be inconsistent with random allocation of participants, indicating possible problems with the randomization

process. Additionally, when examining the degree of variability of the reported data, we found an excess of studies that considerably lack variability in reported means and standard deviations. This can indicate the presence of mistakes in reporting, data fabrication, or undisclosed non-random allocation (which can itself be considered a questionable research practice). The study protocol has proven to be feasible and will be implemented in a large-scale research project, which will hopefully broaden the scope of the inferences made in the present study and provide us with more conclusive evidence about the studied phenomena.

## References


Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS ONE*, 12(3), 1–17. https://doi.org/10.1371/journal.pone.0172792

Allison, P.D.; Stewart, J.A (1974). Productivity differences among scientists: Evidence for accumulative advantage. *American Sociological Review*, 39, 596–606.

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat (p > 0:05): Significance thresholds and the crisis of unreplicable research. *PeerJ*, 2017(7), 1–40. https://doi.org/10.7717/peerj.3544

Anaya, J. (2016). The GRIMMER test: A method for testing the validity of reported measures of variability. *PeerJ Preprints*, 4, e2400v1.

Anderson, M. S., Ronning, E. A., De Vries, R., & Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics*, 13(4), 437–461. https://doi.org/10.1007/s11948-007-9042-5

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, 7(6), 543–554. https://doi.org/10.1177/1745691612459060

Bezeau, Scott; Graves, Roger (2001). Statistical Power and Effect Sizes of Clinical Neuropsychology Research. *Journal of Clinical and Experimental Neuropsychology* (Neuropsychology, Development and Cognition: Section A), 23(3), 399–406. doi:10.1076/jcen.23.3.399.1181

Bishop, D. V. M. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology*, 73(1), 1–19. https://doi.org/10.1177/1747021819886519

Braganza O (2020) A simple model suggesting economically rational sample-size choice drives irreproducibility. *PLoS ONE* 15(3): e0229615. https://doi.org/10.1371/journal.pone.0229615

Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology. *Social Psychological and Personality Science,* 8(4), 363–369. https://doi.org/10.1177/1948550616673876

Bruton, S. V., Medlin, M., Brown, M., & Sacco, D. F. (2020). Personal Motivations and Systemic Incentives: Scientists on Questionable Research Practices. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00182-9

Carlisle, J. B., Dexter, F., Pandit, J. J., Shafer, S. L., & Yentis, S. M. (2015). Calculating the probability of random sampling for continuous variables in submitted or published randomised controlled trials. *Anaesthesia*, 70(7), 848-858.

Carlisle, J. B. (2017). Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia*, 72(8), 944–952. https://doi.org/10.1111/anae.13938

Chambers, Chris (2019). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice.* Princeton University Press.

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: realigning incentives in scientific publishing. *Cortex*, 66, A1-A2.

Coburn, K. M., & Vevea, J. L. (2015). Publication bias as a function of study characteristics. *Psychological Methods*, 20(3), 310–330. https://doi.org/10.1037/met0000046

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

Cohen J. (1992). *Statistical Power Analysis*. Current Directions in Psychological Science. 1(3):98-101. doi:10.1111/1467-8721.ep10768783

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, 49, 997–1003.

Cole, S.; Cole, J.R (1967). Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review*. 32, 377–390.

van Dalen, H. P., & Henkens, K. (2012). Intended and Unintended Consequences of a Publish-or-Perish Culture: A Worldwide Survey. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1983205

Desmond, H. (2021). Incentivizing Replication is Insufficient to Safeguard Default Trust. Forthcoming, *Philosophy of Science*.

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

Edwards, M. A., & Roy, S. (2017). Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. In *Environmental Engineering Science* (Vol. 34, Issue 1, pp. 51–61). Mary Ann Liebert Inc. https://doi.org/10.1089/ees.2016.0223

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

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

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

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Fanelli, D. (2013). Why growing retractions are (mostly) a good sign. *PLoS Medicine*, 10, e1001563. https://doi.org/10.1371/ journal.pmed.1001563

Fanelli, D., Costas, R., & Larivière, V. (2015). Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity. *PLoS ONE*, 10(6). https://doi.org/10.1371/journal.pone.0127556

Fanelli, D., Costas, R., Fang, F. C., Casadevall, A., & Bik, E. M. (2019). Testing Hypotheses on Risk Factors for Scientific Misconduct via Matched-Control Analysis of Papers Containing Problematic Image Duplications. *Science and Engineering Ethics*, 25(3), 771–789. https://doi.org/10.1007/s11948-018-0023-7

Fanelli D (2020) Pressures to publish: what effects do we see?. in: Mario Biagioli & Alexandra Lippman eds. *Gaming the Metrics*. MIT press.

Fang, F. C., Bennett, J. W., & Casadevall, A. (2013). Males are overrepresented among life science researchers committing scientific misconduct. *MBio*, 4(1), 1–3. https://doi.org/10.1128/mBio.00640-12

Fang, F. C., & Casadevall, A. (2015). Competitive science: Is competition ruining science? *Infection and Immunity*, 83(4), 1229–1233. https://doi.org/10.1128/IAI.02939-14

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, 17(1), 120–128. https://doi.org/10.1037/a0024445

Ferguson, C. J., & Heene, M. (2012). A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null. Perspectives on Psychological *Science*, 7(6), 555–561. https://doi.org/10.1177/1745691612459059

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

Garfield, E. (1999). Journal impact factor: a brief review. *Cmaj*, 161(8), 979-980

Gelman, A., & O'Rourke, K. (2014). Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values. *Biostatistics*, 15(1), 18–22. https://doi.org/10.1093/biostatistics/kxt034

Goodstein, D. (2002). Scientific misconduct. *Academe*, 88, 28–31.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82(1), 1.

Hangel, N., & Schmidt-Pfister, D. (2017). Why do you publish? On the tensions between generating scientific knowledge and publication pressure. *Aslib Journal of Information Management*, 69(5), 529–544. https://doi.org/10.1108/AJIM-01-2017-0019

Hartgerink, C. H. J. (2017). Reanalyzing Head et al. (2015): Investigating the robustness of widespread p-hacking. *PeerJ*, 2017(3). https://doi.org/10.7717/peerj.3068

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*, 2016(4). https://doi.org/10.7717/peerj.1935

C. H. J. Hartgerink, J. M. Wicherts, M. A. L. M. van Assen (2017). Too Good to be False: Nonsignificant Results Revisited. *Collabra: Psychology*, 3 (1): 9. https://doi.org/10.1525/collabra.71

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). https://doi.org/10.1371/journal.pbio.1002106

Holm, S., & Hofmann, B. (2018). Associations between attitudes towards scientific misconduct and self-reported behavior. *Accountability in Research*, 25(5), 290–300. https://doi.org/10.1080/08989621.2018.1485493

H. M. James Hung, O'Neill, R., Bauer, P., & Kohne, K. (1997). The Behavior of the P-Value When the Alternative Hypothesis is True. *Biometrics*, 53(1), 11-22. doi:10.2307/2533093

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.

Janke, S., Daumiller, M., & Rudert, S. C. (2019). Dark Pathways to Achievement in Science: Researchers' Achievement Goals Predict Engagement in Questionable Research Practices. *Social Psychological and Personality Science*, 10(6), 783–791. https://doi.org/10.1177/1948550618790227

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

Kaatz, A., Vogelman, P. N., & Carnes, M. (2013). Are men more likely than women to commit scientific misconduct? Maybe, maybe not. *MBio*, 4(2). https://doi.org/10.1128/mBio.00156-13

Kepes, S., Banks, G. C., & Oh, I. S. (2014). Avoiding bias in publication bias research: The value of "null" findings. *Journal of Business and Psychology*, 29, 183–203. http://dx.doi.org/10.1007/s10869-012-9279-0

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

Kun, Á. (2018). Publish and who should perish: You or science? *Publications*, 6(2). https://doi.org/10.3390/publications6020018

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7), 701-710.

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

Lakens, D. (2015b). On the challenges of drawing conclusions from p-values just below 0.05. *PeerJ*, 2015(7). https://doi.org/10.7717/peerj.1142

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(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Larson, R. C., Ghaffarzadegan, N., & Xue, Y. (2014). Too many PhD graduates or too few academic job openings: The basic reproductive number R0 in academia. *Systems Research and Behavioral Science*, 31(6), 745-750.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods (Vols. 1-0).* Thousand Oaks, CA: Sage Publications, Inc. doi: 10.4135/9781412963947

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. R. (2013). The life of p: "Just significant" results are on the rise. *Quarterly Journal of Experimental Psychology*, 66(12), 2303–2309. https://doi.org/10.1080/17470218.2013.863371

Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs*, 76(3), 286–302. https://doi.org/10.1080/03637750903074685

Maggio, L., Dong, T., Driessen, E., & Artino, A. (2019). Factors associated with scientific misconduct and questionable research practices in health professions education.

*Perspectives on Medical Education*, 8(2), 74–82. https://doi.org/10.1007/s40037-019-0501-x

Markowitz, D. M., & Hancock, J. T. (2014). Linguistic traces of a scientific fraud: The case of Diederik Stapel. *PLoS ONE*, 9(8), e105937.

Marks-Anglin, A., & Chen, Y. (2020). A historical review of publication bias. *Research Synthesis Methods*, 11(6), 725–742. https://doi.org/10.1002/jrsm.1452

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*, 65(11), 2271–2279.

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

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.

Munafò, M., & Neill, J. (2016). Null is beautiful: On the importance of publishing null results. *Journal of Psychopharmacology*, 30(7), 585. https://doi.org/10.1177/0269881116638813

Murphy, K. R., & Aguinis, H. (2019). HARKing: How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results? *Journal of Business and Psychology*, 34(1), 1–17. https://doi.org/10.1007/s10869-017-9524-7

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine (2009). *On Being a Scientist: A Guide to Responsible Conduct in Research: Third Edition.* Washington, DC: The National Academies Press. https://doi.org/10.17226/12192.

Neyman, J., & Pearson, E. (1933). 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.

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

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

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

Pastore, M., & Calcagnì, A. (2019). Measuring distribution similarities between samples: a distribution-free overlapping index. *Frontiers in psychology*, 10, 1089.

Pautasso, M. (2010). Worsening file-drawer problem in the abstracts of natural, medical and social science databases. *Scientometrics*, 85(1), 193–202. https://doi.org/10.1007/s11192-010-0233-5

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

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ropovik, I., Adamkovic, M., & Greger, D. (2019). Neglect of publication bias compromises meta-analyses of educational research. *MetArXiv*, https://doi.org/10.31222/osf.io/z7drq

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

Rubin, M. (2017). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology*, 21(4), 308–320. https://doi.org/10.1037/gpr0000128

Senn, S. (1994). Testing for baseline balance in clinical trials. *Statistics in Medicine*, 13(17), 1715–1726. https://doi.org/10.1002/sim.4780131703

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

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

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

Steneck, N. H. (2007). Fostering Integrity in Research: Definitions, Current Knowledge, and Future Directions. *Science and Engineering Ethics*, 13(1), 69–82. https://doi.org/10.1007/s11948-006-0006-y

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American statistical association*, 54(285), 30-34.

Sterne, J. A., Becker, B. J., & Egger, M. (2005). The funnel plot. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, 75-98.

Stricker, J., & Günther, A. (2019). Scientific Misconduct in Psychology: A Systematic Review of Prevalence Estimates and New Empirical Data. In *Zeitschrift fur Psychologie / Journal of Psychology* (Vol. 227, Issue 1, pp. 53–63). Hogrefe Verlag GmbH & Co. KG. https://doi.org/10.1027/2151-2604/a000356

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, 15(3), e2000797.

Tijdink, J. K., Bouter, L. M., Veldkamp, C. L. S., Van De Ven, P. M., Wicherts, J. M., & Smulders, Y. M. (2016). Personality traits are associated with research misbehavior in Dutch scientists: A cross-sectional study. *PLoS ONE*, 11(9). https://doi.org/10.1371/journal.pone.0163251

Vankov, I., Bowers, J., & Munafò, M. R. (2014). Article commentary: On the persistence of low power in psychological science. *Quarterly Journal of Experimental Psychology*, 67(5), 1037-1040.

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

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8), 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006

Wass, M. N., Ray, L., & Michaelis, M. (2019). Understanding of researcher behavior is required to improve data reliability. *GigaScience*, 8(5), giz017.

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, 5(10), 1418–1422. https://doi.org/10.1371/journal.pmed.0050201

## List of abbreviations

| *Abbreviation* | *Meaning* |
|---|---|
| HARKing | Hypothesizing after the results are known |
| NHST | Null hypothesis significance testing |
| RCT | Randomized controlled trial |
| QRPs | Questionable research practices |