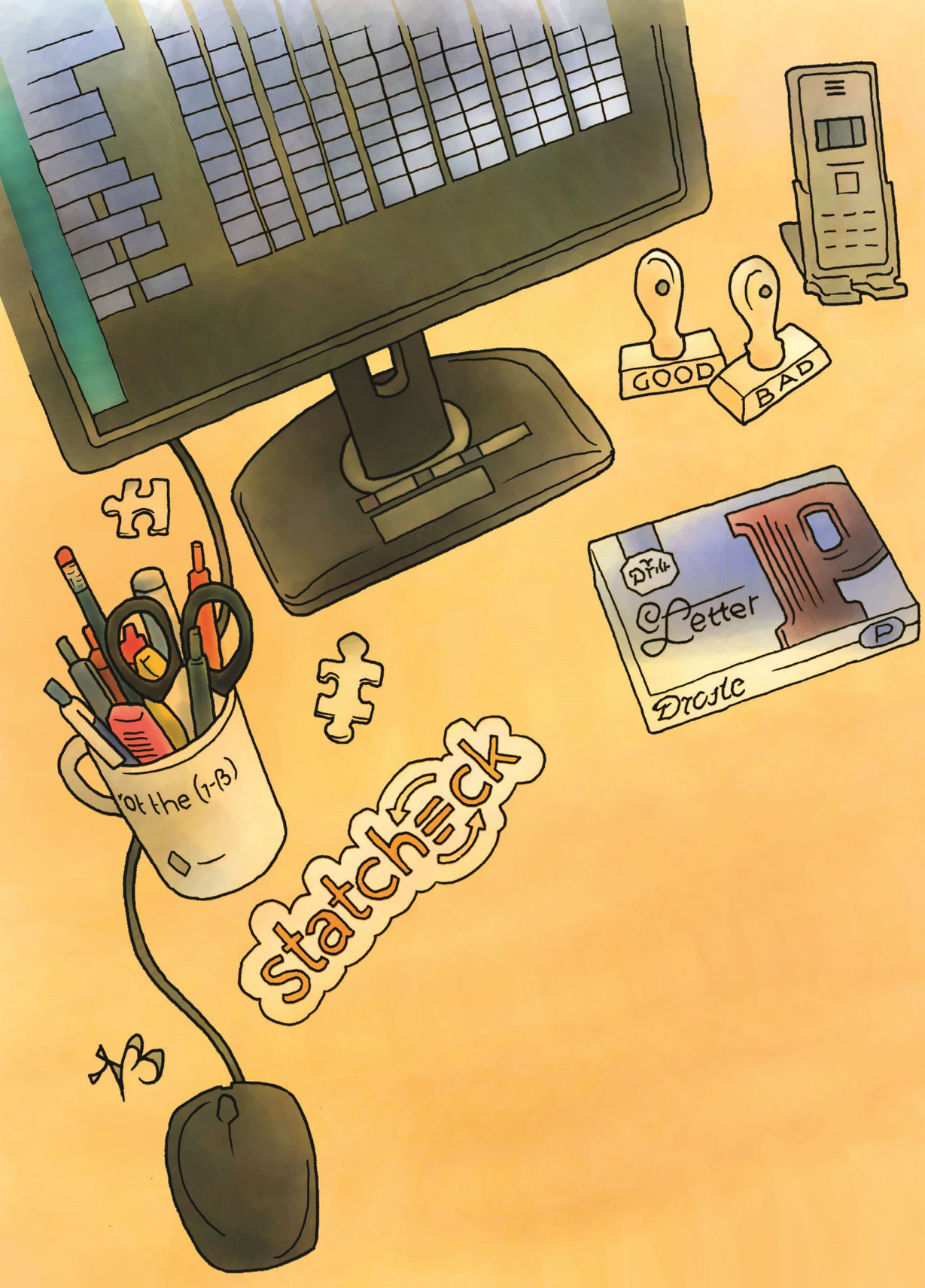
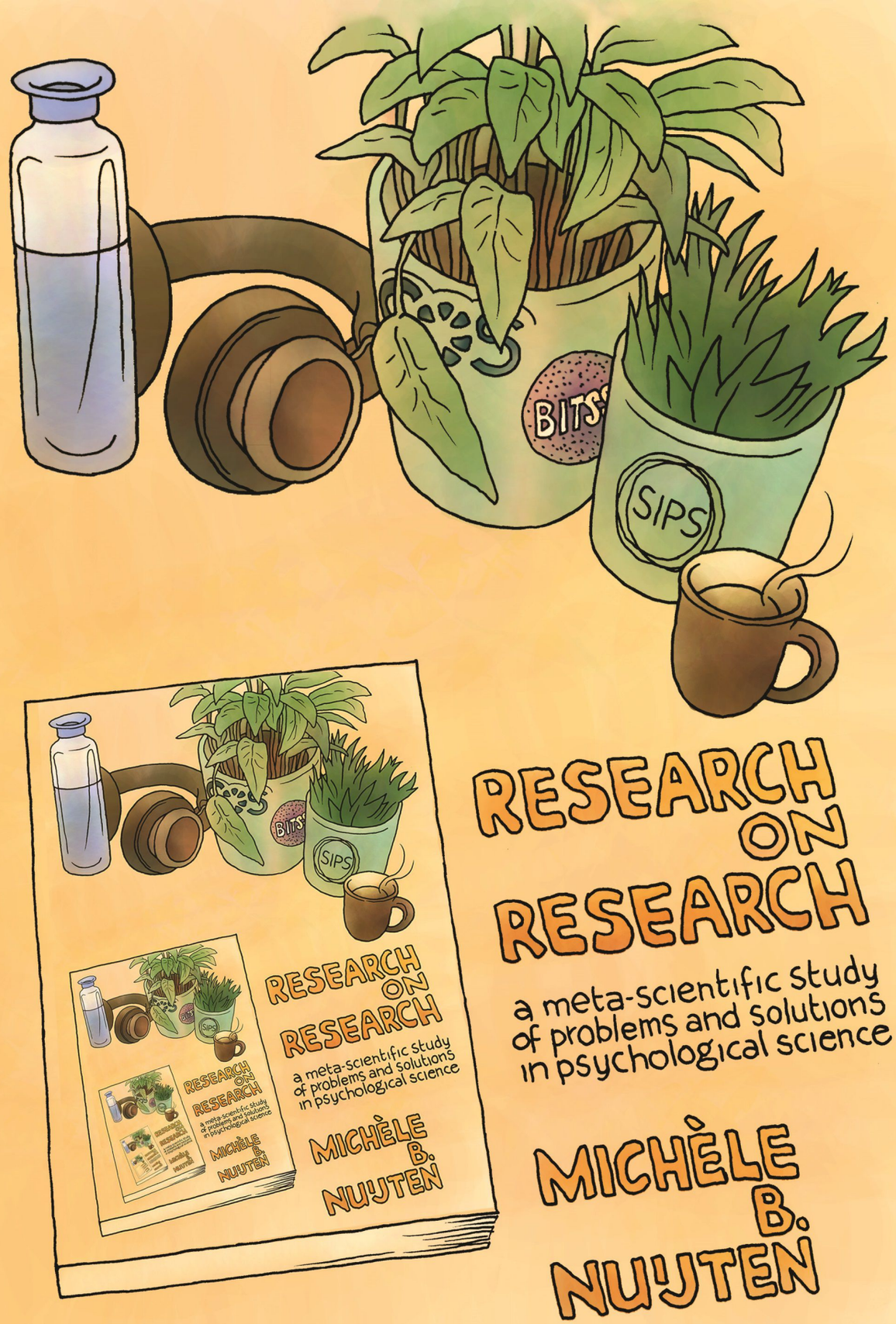


research on research

michèle b. nuijten



RESEARCH ON RESEARCH

a meta-scientific study of problems and solutions in psychological science

MICHÈLE B. NUIJTEN

Research on Research

A Meta-Scientific Study of Problems and Solutions in Psychological Science

Michèle B. Nuijten

Author: Michèle B. Nuijten
Cover design: Niels Bongers – www.nielsbongers.nl
Printed by: Gildeprint – www.gildeprint.nl
ISBN: 978-94-6233-928-6

Research on Research

A Meta-Scientific Study of Problems and Solutions

in Psychological Science

Proefschrift ter verkrijging van de graad van doctor
aan Tilburg University
op gezag van de rector magnificus,
prof. dr. E. H. L. Aarts,
in het openbaar te verdedigen
ten overstaan van
een door het college voor promoties aangewezen commissie
in de aula van de Universiteit
op woensdag 30 mei 2018 om 14:00 uur
door Michèle Bieneke Nuijten,
geboren te Utrecht.

Promotiecommissie

Promotores:	Prof. dr. J. M. Wicherts Prof. dr. M. A. L. M. van Assen
Overige leden:	Prof. dr. C. D. Chambers Prof. dr. E. J. Wagenmakers Prof. dr. R. A. Zwaan Dr. M. Bakker

Contents

Chapter 1

Introduction

Can we trust psychological research findings? This question is asked more and more, and there is growing concern that many published findings are overly optimistic (Francis, 2014; Ioannidis, 2005, 2008; John, Loewenstein, & Prelec, 2012; Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011). An increasing number of studies shows that we might have good reason to doubt the validity of published psychological findings, and researchers are even starting to speak of a "crisis of confidence" or a "replicability crisis" (Baker, 2016a; Pashler & Harris, 2012; Pashler & Wagenmakers, 2012; Spellman, 2015).

1.1 Replicability in Psychology

The growing concern about psychology's trustworthiness is fueled by the finding that a large number of published psychological findings could not be replicated in novel samples. For instance, the large-scale, collaborative Reproducibility Project: Psychology (RPP) investigated the replicability of 100 psychology studies (Open Science Collaboration, 2015). Two of the main findings in this project were that the percentage of statistically significant effects dropped from 97% in the original studies to only 36% in the replications, and that the effect sizes in the replications were only about half the size of those in the original studies. Other multi-lab initiatives also failed to replicate key findings in psychology (Alogna et al., 2014; Eerland et al., 2016; Hagger et al., 2016; Wagenmakers et al., 2016)

There are several possible explanations for the low replicability rates in psychology. One possibility is that meaningful differences between the original studies and their replications caused the differences in results (Baumeister, 2016; Dijksterhuis, 2014; Iso-Ahola, 2017; Stroebe & Strack, 2014). Indeed, there are some indications that some effects show large between-study variability, which could explain the low replicability rates (Klein et al., 2014) . Another explanation, however, is that the original studies overestimated the effects or were false positives (chance) findings.

1.2 Bias and Errors

Several research findings are in line with the notion that published effects are overestimated. For instance, the large majority of studies in psychology find support for the tested hypothesis (Fanelli, 2010; Francis, 2014; Sterling, Rosenbaum, & Weinkam, 1995). However, this is incompatible with the generally low statistical power of studies in the psychological literature (Bakker, van Dijk, & Wicherts, 2012; Button et al., 2013; Cohen, 1962; Jennions & Moller, 2003; Maxwell, 2004; Schimmack, 2012). Low power decreases the probability that a study finds a significant effect. Conversely, and perhaps counterintuitively, the lower the power, the higher the probability that a significant finding is a false positive. The large number of underpowered studies in psychology that do find significant effects therefore might indicate a problem with the trustworthiness of these findings.

The notion that many findings are overestimated also becomes clear in meta-analyses. Meta-analysis is a crucial scientific tool to quantitatively synthesize the results of different studies on the same research question (Borenstein, Hedges, Higgins, & Rothstein, 2009). The results of meta-analyses inspire policies and treatments, so it is essential that the effects reported in them are valid. However, in many fields meta-analytic effects appear to be overestimated (Ferguson & Brannick, 2012; Ioannidis, 2011; Niemeyer, Musch, & Pietrowsky, 2012, 2013; Sterne, Gavaghan, & Egger, 2000; Sutton, Duval, Tweedie, Abrams, & Jones, 2000). One of the main causes seems to be publication bias; the phenomenon that statistically significant findings have a higher probability of being published than nonsignificant findings (Greenwald, 1975).

The evidence that the field of psychology is affected by publication bias is overwhelming. Studies found that manuscripts without significant results are both less likely to be submitted and less likely to be accepted for publication (Cooper, DeNeve, & Charlton, 1997; Coursol & Wagner, 1986; Dickersin, Chan, Chalmers, Sacks, & Smith, 1987; Epstein, 1990; Franco, Malhotra, & Simonovits, 2014; Greenwald, 1975; Mahoney, 1977). Furthermore, published studies seem to have systematically larger effects than unpublished ones (Franco et al., 2014; Polanin, Tanner-Smith, & Hennessy, 2015).

The *de facto* requirement to report statistically significant results in journal articles can lead to unwanted strategic behavior in data analysis (Bakker et al., 2012). Data analysis in psychology is very flexible: there are many possible statistical analyses to answer the same research question (Gelman & Loken, 2014; Wicherts et al., 2016). It can be shown that strategic use of this flexibility will almost always result in at least one significant finding; one that is likely to be a false positive (Bakker et al., 2012; Simmons et al., 2011). This becomes even more problematic, if only the analyses that "worked" are reported and presented as if they were planned from the start (Kerr, 1998; Wagenmakers, Wetzels, Borsboom, Maas, & Kievit, 2012). Survey results show that many psychologists admit to such "questionable research practices" (QRPs; Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; John et al., 2012), and use of study registers and later disclosures by researchers provide direct evidence that indeed some of these practices are quite common (Franco, Malhotra, & Simonovits, 2016; LeBel et al., 2013).

Another example of a QRP that illustrates a strong focus on finding significant results, is wrongly rounding down *p*-values to < .05. This is a particularly surprising strategy, since this can be readily observed in published papers. If a *p*-value is wrongly rounded down, it often leads to a statistical reporting inconsistency. Statistical reporting inconsistencies occur when the test statistic, the degrees of freedom, and the *p*-value in a null hypothesis significance test (NHST) are not internally consistent. If the reported *p*-value is significant, whereas the recalculated *p*-value based on the reported degrees of freedom and test statistic is not, or vice versa, this is considered a gross inconsistency. Several studies found a high prevalence of such

reporting inconsistencies (e.g., Bakker & Wicherts, 2011; Caperos & Pardo, 2013). Even though the majority of inconsistencies seemed to be innocent typos and rounding errors, there is evidence for a systematic bias towards finding significant results, in line with the notion that some researchers may wrongly round down *p*-values in an effort to present significant results.

All these problems lead to the question: how trustworthy is psychological science? Are published findings overly optimistic? If it is true that most published findings are overestimated or even false positives (Ioannidis, 2005, 2008), the consequences are severe. It would mean that large amounts of research resources (often paid by the tax payer) are wasted by pursuing seemingly interesting research lines, that turn out to be non-replicable (Chalmers & Glasziou, 2009). Biased or erroneously reported results also lower trust in psychological science and create less useful results for society.

1.3 Meta-Research & the Focus of this Dissertation

It is important to determine if published findings in psychology are overestimated or incorrectly reported, what causes errors and overestimation, and how we can solve these problems. We can answer such (empirical) questions by doing "research on research", forming what has become known as meta-science (Ioannidis, Fanelli, Dunne, & Goodman, 2015). In this dissertation, we use a meta-scientific approach to investigate problems and solutions in psychological science.

An attempt to explain the entire replication crisis and its causes is beyond the scope of this dissertation, and arguably even beyond the scope of my entire scientific career. However, just as in any scientific field, big questions are answered by a series of small findings. In this dissertation, I specifically chose to focus on potential indicators of errors and biased effects in the published psychological literature. This means that we do not investigate the motivation or intention behind choices that researchers make. Although these are important topics and deserve a research line of their own, our focus is on the trustworthiness of published psychological research rather than on the trust we could place in individual researchers. This dissertation consists of two main parts that deal with specific problems. Part I focuses on statistical reporting inconsistencies in published articles, and Part II focuses on possible bias in effect size estimates.

1.3.1 Part I: Statistical Reporting Inconsistencies

There are several reasons why we chose to focus on statistical reporting inconsistencies. First, reporting inconsistencies are prevalent in the psychological literature; almost half of published psychology articles contain at least one inconsistent *p*-value, and in about 12-17% of the articles there is a gross inconsistency that concerns significance (Bakker & Wicherts, 2011; Caperos & Pardo, 2013). Second, investigating reporting inconsistencies might be one of the only ways to directly observe QRPs. Even though many inconsistencies

are likely to be innocent typos, self-reports show that over 20% of psychologists admit to having wrongly rounded off a *p*-value to make a result appear significant (Agnoli et al., 2017; John et al., 2012). Indeed, gross inconsistencies are often in line with researchers expectations (Bakker & Wicherts, 2011), and reporting inconsistencies are related to a reluctance to share data for verification purposes (Wicherts, Bakker, & Molenaar, 2011).

In Chapter 2, we investigate the prevalence of statistical reporting inconsistencies in over 30,000 articles from 8 prestigious psychology journals, using the R package "statcheck" (Epskamp & Nuijten, 2016). Statcheck is a tool to automatically extract statistics from articles and recalculate *p*-values. In Chapter 3, we present additional validity analyses for statcheck, based on some critiques and questions it has received. Here, we calculate statcheck's sensitivity and specificity, and investigate how it deals with statistics that are corrected for multiple testing or violations of assumptions. In Chapter 4, we use statcheck to see whether statistical reporting inconsistencies are related to journals' data sharing policies and actual data sharing practices by researchers. In Chapter 5 we make recommendations for what journal editors can do to avoid reporting inconsistencies.

We specifically do not focus on the question whether NHST is a good statistical framework in the first place (Nickerson, 2000). It has been argued that the NHST framework is inherently flawed (Krueger, 2001; Wagenmakers, 2007) and even that *p*-values should be abandoned altogether (Trafimow & Marks, 2015). Several authors have argued in favor of alternartive inferential approaches, including the use of effect size estimation and confidence intervals (Cumming, 2013), or Bayesian statistics (Kruschke, 2014; Wagenmakers, 2007). Although this is an important discussion, it is beyond the scope of this dissertation. Our aim was to document problems in the current psychological literature, and with over 90 % of articles using it, NHST is clearly dominant in this literature (Cumming et al., 2007; Hubbard & Ryan, 2000; Sterling et al., 1995).

1.3.2 Part II: Bias in Effect Sizes

Part II of this dissertation focuses on bias in effect size estimates. Previous research gave us sufficient reason to suspect that many effect sizes are overestimated (Button et al., 2013; Fanelli, 2010; Fanelli, Costas, & Ioannidis, 2017; Song et al., 2010). A big problem is that it is hard to determine for an individual study whether it contains an overestimated effect, and if so, how much it is overestimated. And if we do suspect a study contains an overestimated effect, it is hard, if not impossible to determine if that is simply because of random sampling variation, or because of problems such as publication bias and QRPs. What we can do, however, is look for patterns of bias in meta-analyses (Fanelli et al., 2017; Rothstein, Sutton, & Borenstein, 2005; Song et al., 2010).

In a meta-analysis, it is possible to compare the effect of an individual study to the rest of the included studies and to the overall average effect. This enables a (systematic)

investigation of signs of publication bias and related problems across a set of studies on a particular topic. For instance, if there is publication bias based on signifcance, one would expect that smaller studies in a set of otherwise similar studies to systematically find larger effect sizes than larger studies. This is known as the "small study effect" (Sterne & Egger, 2005). This phenomenon occurs because the chance of finding a significant result for genuine effects (i.e., the power) is lower for smaller studies. In studies with low power, effects are estimated with low precision and can be strongly under- and overestimated. In a small, underpowered study, for an effect to reach statistical significance, it has to be very large. That means that if only significant studies are published, the inflation of published effects in small studies increases (Button et al., 2013; Kraemer, Gardner, Brooks, & Yesavage, 1998). Note that publication bias is only one potential cause of a small study effect. A small study effect can also arise for other reasons, for instance if researchers determine their sample size based on an a priori power analysis in combination with a correctly appraised true effect size, or if researchers by experience learn to use smaller samples when true effect sizes tend to be larger.

Bias in effect sizes is hard to directly observe, so estimating patterns in meta-analyses, such as the small study effect, is arguably the best way to look for signs of overestimation and other potential problems. In Chapters 7 to 9 we investigate circumstances in which overestimation in meta-analyses occurs and look for factors that might worsen this overestimation. We also investigate whether there are study characteristics that predict an increased risk for overestimation.

In Chapter 7, we formally show that combining published studies to obtain an overall effect size estimate can actually decrease accuracy of the estimate. This result is very counterintuitive, as we also found in a survey among psychology students, social scientists, and quantitative psychologists. However, it is caused by publication bias and happens whenever effect sizes of small studies are statistically combined with those of large studies in meta-analysis of the relevant literature. Effectively, this is what often happens in meta-analyses, so in Chapter 8 and 9 we investigated patterns of bias in large sets of meta-analyses. In Chapter 8, we reanalyze data from 82 meta-analyses, to see if we can corroborate the findings of Fanelli and Ioannidis (2013) that overestimation of effects becomes worse for studies from the US (a so-called US effect). In Chapter 9, we analyze 131 meta-analyses about intelligence research to look for possible patterns of bias. Specifically, we look for patterns indicating a small study effect and the US effect. We also study evidence in favor of the decline effect (of effects dimishing over time), early-extremes effect (of effects being more variant in the early phases of research lines), and citation bias (i.e., the pattern wherein larger effects yield more citations than smaller effects). We chose to focus on intelligence, because it represents one of the most well-known constructs in psychology and has been investigated extensively from various (sub)disciplines (e.g., behavior genetics, cognitive psychology,

neuroscience, developmental psychology), and using different methods including correlational and experimental designs. This makes intelligence research a good field to study effect sizes, power, and biases in a wide range of fields using different methods that still focused on measures of the same construct.

Part I and Part II of this thesis focus on different aspects that influence the trustworthiness of psychology. We therefore chose to end each part with a separate discussion of the main findings. However, the two Parts also have one major theme in common; they both focus on using empirical methods to investigate problems and solutions in psychological science. If we can use this meta-scientific approach to solve problems involving statistical reporting inconsistencies and bias in effect size estimation, we are already closer to more trustworthy research. However, to solve all problems that are currently threatening psychology, we need more research and big reforms. This dissertation therefore ends with an overview of current initiatives and ideas for future research to further improve psychological science.

Part I
Statistical Reporting Inconsistencies

Chapter 2

The Prevalence of Statistical Reporting Errors in Psychology (1985-2013)

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

Abstract

This study documents reporting errors in a sample of over 250,000 *p*-values reported in eight major psychology journals from 1985 until 2013, using the new R package "statcheck". Statcheck retrieved null-hypothesis significance testing (NHST) results from over half of the articles from this period. In line with earlier research, we found that half of all published psychology papers that use NHST contained at least one *p*-value that was inconsistent with its test statistic and degrees of freedom. One in eight papers contained a grossly inconsistent *p*-value that may have affected the statistical conclusion. In contrast to earlier findings, we found that the average prevalence of inconsistent *p*-values has been stable over the years or has declined. The prevalence of gross inconsistencies was higher in *p*-values reported as significant, than in *p*-values reported as nonsignificant. This could indicate a systematic bias in favor of significant results. Possible solutions for the high prevalence of reporting inconsistencies could be to encourage sharing data, to let co-authors check results in a so-called "co-pilot model", and to use statcheck to flag possible inconsistencies in one's own manuscript or during the review process.

Most conclusions in psychology are based on the results of Null Hypothesis Significance Testing (NHST; Cumming et al., 2007; Hubbard & Ryan, 2000; Sterling, 1959; Sterling et al., 1995). Therefore, it is important that NHST is performed correctly and that NHST results are reported accurately. However, there is evidence that many reported *p*-values do not match their accompanying test statistic and degrees of freedom (Bakker & Wicherts, 2011; Bakker & Wicherts, 2014; Berle & Starcevic, 2007; Caperos & Pardo, 2013; Garcia-Berthou & Alcaraz, 2004; Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, 2014; Wicherts et al., 2011). These studies highlighted that roughly half of all published empirical psychology articles using NHST contained at least one inconsistent *p*-value and that around one in seven articles contained a gross inconsistency, in which the reported *p*-value was significant and the computed *p*-value was not, or vice versa.

This alarmingly high error rate can have large consequences. Reporting inconsistencies could affect whether an effect is perceived to be significant or not, which can influence substantive conclusions. If a result is inconsistent it is often impossible (in the absence of raw data) to determine whether the test statistic, the degrees of freedom, or the *p*-value were incorrectly reported. If the test statistic is incorrect and it is used to calculate the effect size for a meta-analysis, this effect size will be incorrect as well, which could affect the outcome of the meta-analysis (Bakker & Wicherts, 2011; in fact, the misreporting of all kinds of statistics is a problem for meta-analyses; Gotzsche, Hrobjartsson, Maric, & Tendal, 2007; Levine & Hullett, 2002). Incorrect *p*-values could affect the outcome of tests that analyze the distribution of *p*-values, such as *p*-curve (Simonsohn, Nelson, & Simmons, 2014) and *p*-uniform (van Assen, van Aert, & Wicherts, 2015). Moreover, Wicherts et al. (2011) reported that a higher prevalence of reporting errors was associated with a failure to share data upon request.

Even though reporting inconsistencies can be honest mistakes, they have also been categorized as one of several fairly common questionable research practices (QRPs) in psychology (John et al., 2012). Interestingly, psychologists' responses to John et al.'s survey fitted a Guttman scale reasonably well. This suggests that a psychologist's admission to a QRP that is less often admitted to by others usually implies his or her admission to QRPs with a higher admission rate in the entire sample. Given that rounding down *p*-values close to .05 was one of the QRPs with relatively low admission rates, the frequency of misreported *p*-values could provide information on the frequency of the use of more common QRPs. The results of John et al. would therefore imply that a high prevalence of reporting errors (or more specifically, incorrect rounding down of *p*-values to be below .05) can be seen as indicator of the use of other QRPs, such as the failure to report all dependent variables, collecting of more data after seeing whether results are significant, failing to report all conditions, and stopping data collection after achieving the desired result. Contrary to many other QRPs in John et al.'s list, misreported *p*-values that bear on significance can be readily detected on the basis of the articles' text.

Previous research found a decrease in negative results (Fanelli, 2012) and an increase in reporting inconsistencies (Leggett, Thomas, Loetscher, & Nicholls, 2013) suggesting that QRPs are on the rise. On the other hand, it has been found that the number of published corrections to the literature did not change over time, suggesting no change in QRPs over time (Fanelli, 2013, 2014). Studying the prevalence of misreported *p*-values over time could shed light on possible changes in prevalence of QRPs.

Beside possible changes in QRPs over time, some evidence suggests that the prevalence of QRPs may differ between subfields of psychology. Leggett et al. (2013) recently studied reporting errors in two main psychology journals in 1965 and 2005. They found that the increase in reporting inconsistencies over the years was higher in the *Journal of Personality and Social Psychology* (JPSP), the flagship journal of social psychology, than in *Journal of Experimental Psychology: General* (JEPG). This is in line with the finding of John et al. (2012) that social psychologists admit to more QRPs, find them more applicable to their field, and find them more defensible as compared to other subgroups in psychology (but see also Fiedler & Schwarz, 2016, on this issue). However, the number of journals and test results in Leggett et al.'s study was rather limited and so it is worthwhile to consider more data before drawing conclusions with respect to differences in QRPs between subfields in psychology.

The current evidence for reporting inconsistencies is based on relatively small sample sizes of articles and *p*-values. The goal of our current study was to evaluate reporting errors in a large sample of more than a quarter million *p*-values retrieved from eight flagship journals covering the major subfields in psychology. Manually checking errors is time-consuming work, therefore we present and validate an automated procedure in the R package *statcheck* (Epskamp & Nuijten, 2015). The validation of statcheck is described in Appendix A (see also Chapter 3 of this dissertation).

We used statcheck to investigate the overall prevalence of reporting inconsistencies and compare our findings to findings in previous studies. Furthermore, we investigated whether there has been an increase in inconsistencies over the period 1985 to 2013, and, on a related note, whether there has been any increase in the number of NHST results in general and per article. We also documented any differences in the prevalence and increase of reporting errors between journals. Specifically, we studied whether articles in social psychology contain more inconsistencies than articles in other subfields of psychology.

2.1 Method

2.1.1 "statcheck"

To evaluate the prevalence of reporting errors, we used the automated procedure *statcheck* (version 1.0.1.; Epskamp & Nuijten, 2015). This freely available R package (R Core Team, 2014) extracts statistical results and recalculates *p*-values based on reported test

statistics and their degrees of freedom. Roughly, the underlying procedure executes the following four steps.

Step 1. First, statcheck converts a PDF or HTML file to a plain text file. The conversion from PDF to plain text can sometimes be problematic, because some journal publishers use images of signs such as "<", ">", or "=", instead of the actual character. These images are not converted to the text file. HTML files do not have such problems and typically render accurate plain text files.

Step 2. From the plain text file, statcheck extracts *t*, *F*, *r*, χ^2, and *Z* statistics, with accompanying degrees of freedom (*df*) and *p*-value. Since statcheck is an automated procedure, it can only search for prespecified strings of text. Therefore, we chose to let statcheck search for results that are reported completely and exactly in APA style (American Psychological Association, 2010). A general example would be "*test statistic* (df1, df2) =/</> …, *p* =/</> …". Two more specific examples are: "*t*(37) = -4.93, *p* <.001", "χ^2(1, N = 226) = 6.90, *p* <.01". Statcheck takes different spacing into account, and also reads results that are reported as nonsignificant (*ns*). On the other hand, it does not read results that deviate from the APA template. For instance, statcheck overlooks cases in which a result includes an effect size estimate in between the test statistic and the *p*-value (e.g., "*F*(2, 70) = 4.48, *MSE* = 6.61, *p* <.02") or when two results are combined into one sentence (e.g., "*F*(1, 15) = 19.9 and 5.16, *p* <.001 and *p* <.05, respectively"). These restrictions usually also imply that statcheck will not read results in tables, since these are often incompletely reported (see Appendix A for a more detailed overview of what statcheck can and cannot read).

Step 3. Statcheck uses the extracted test statistics and degrees of freedom to recalculate the *p*-value. By default all tests are assumed to be two-tailed. We compared *p*-values recalculated by statcheck in R version 3.1.2 and Microsoft Office Excel 2013 and found that the results of both programs were consistent up to the tenth decimal point. This indicates that underlying algorithms used to approximate the distributions are not specific to the R environment.

Step 4. Finally, statcheck compares the reported and recalculated *p*-value. Whenever the reported *p*-value is inconsistent with the recalculated *p*-value, the result is marked as an *inconsistency*. If the reported *p*-value is inconsistent with the recalculated *p*-value and the inconsistency changes the statistical conclusion (assuming $\alpha = .05$) the result is marked as a *gross inconsistency*. To take into account one-sided tests, statcheck scans the whole text of the article for the words "one-tailed", "one-sided", or "directional". If a result is initially marked as inconsistent, but the article mentions one of these words *and* the result would have been consistent if it were one-sided, then the result is marked as consistent. Note that statcheck does not take into account *p*-values that are adjusted for multiple testing (e.g., a Bonferroni correction). *P*-values adjusted for multiple comparisons that are higher than the recalculated *p*-value can therefore erroneously be marked as inconsistent. However, when we

automatically searched our sample of 30,717 articles, we found that only 96 articles reported the string "Bonferroni" (0.3%) and 9 articles reported the string "Huynh-Feldt" or "Huynh Feldt" (0.03%). We conclude from this that corrections for multiple testing are rarely used and will not significantly distort conclusions in our study (but see also Chapter 3 of this dissertation).

Similar to Bakker and Wicherts (2011), statcheck takes numeric rounding into account. Consider the following example: $t(28) = 2.0$, $p<.05$. The recalculated p-value that corresponds to a t-value of 2.0 with 28 degrees of freedom is .055, which appears to be inconsistent with the reported p-value of < .05. However, a reported t-value of 2.0 could correspond to any rounded value between 1.95 and 2.05, with a corresponding range of p-values between .0498 and .0613, which means that the reported p <.05 is not considered inconsistent.

Furthermore, statcheck considers p-values reported as $p = .05$ as significant. We inspected 10% of the 2,473 instances in our sample in which a result was reported as "$p = .05$" and inspected whether these p-values were interpreted as significant.[1] In the cases where multiple p-values from the same article were selected, we only included the p-value that was drawn first to avoid dependencies in the data. Our final sample consisted of 236 instances where "$p = .05$" was reported and of these p-values 94.3% was interpreted as being significant. We therefore decided to count p-values reported as "$p = .05$" as indicating that the authors presented the result as significant.

The main advantage of statcheck is that it enables searching for reporting errors in very large samples, which would be unfeasible by hand. Furthermore, manual checking is subject to human error, which statcheck eliminates. The disadvantage of statcheck is that it is not as comprehensive as a manual procedure, because it will miss results that deviate from standard reporting and results in tables, and it does not take into account adjustments on p-values. Consequently, statcheck will miss some reported results and will incorrectly earmark some correct p-values as a reporting error. Even though it is not feasible to create an automated procedure that is as accurate as a manual search in veryfying correctness of the results, it is important to exclude the possibility that statcheck yields a biased depiction of the true inconsistency rate. To avoid bias in the prevalence of reporting errors, we performed a validity study of statcheck, in which we compared statcheck's results with the results of Wicherts, Bakker, and Molenaar (2011), who performed a manual search for and verification of reporting errors in a sample of 49 articles.

The validity study showed that statcheck read 67.5% of the results that were manually extracted. Most of the results that statcheck missed were either reported with an effect size between the test statistics and the p-value (e.g., $F(2, 70) = 4.48$, MSE = 6.61, p <.02; 201

[1] For a more extensive analysis of p-values around .05 in this sample, see Hartgerink, Van Aert, Nuijten, Wicherts, and Van Assen (2016)

instances in total) or reported in a table (150 instances in total). Furthermore, Wicherts et al. found that 49 of 1148 *p*-values were inconsistent (4.3%) and 10 of 1148 *p*-values were grossly inconsistent (.9%), whereas statcheck (with automatic one-tailed test detection) found that 56 of 775 *p*-values were inconsistent (7.2%) and 8 of 775 *p*-values grossly inconsistent (1.0%). The higher inconsistency rate found by statcheck was mainly due to our decision to count $p = .000$ as incorrect (a *p*-value cannot exactly be zero), whereas this was counted correct by Wicherts et al. If we do not include these eleven inconsistencies due to $p = .000$, statcheck finds an inconsistency percentage of 5.8% (45 of 775 results), 1.5 percentage point higher than in Wicherts et al. This difference was due to the fact that statcheck did not take into account eleven corrections for multiple testing and Wicherts et al. did. The inter-rater reliability in this scenario between the manual coding in Wicherts et al. and the automatic coding in statcheck was .76 for the inconsistencies and .89 for the gross inconsistencies. Since statcheck slightly overestimated the prevalence of inconsistencies in this sample of papers, we conclude that statcheck can render slightly different inconsistency rates than a search by hand. Therefore, the results of statcheck should be interpreted with care. For details of the validity study and an explanation of all discrepancies between statcheck and Wicherts et al., see Appendix A. A further analysis of the validity of statcheck is described in Chapter 3.

2.1.2 Sample

A pilot study of social science journals in the Web of Science citation data base showed that few journals outside psychology include APA reporting style, therefore we limited our sample to psychology journals. As explained above, statcheck cannot always read results from articles in PDF due to problems in the conversion from PDF to plain text. These problems do not occur in articles in HTML format. Therefore, to obtain the most reliable statcheck results we restricted our sample to articles that were available in HTML format. The time span over which we downloaded articles depended on the year a journal started to publish articles in HTML. We collected the data in 2014, so we included articles up until 2013 to ensure complete sets of articles for an entire year. Via EBSCOhost we manually downloaded all articles in HTML from 1985 to 2013 from six flagship psychology journals that represent six main sub disciplines: *Journal of Applied Psychology* (JAP; Applied Psychology), *Journal of Consulting and Clinical Psychology* (JCCP; Clinical Psychology), *Developmental Psychology* (DP; Developmental Psychology), *Journal of Experimental Psychology: General* (JEPG; Experimental Psychology), and *Journal of Personality and Social Psychology* (JPSP; Social Psychology). These journals are published by the APA and follow the APA reporting guidelines. Furthermore, we manually downloaded all articles in HTML from two journals in general psychology: *Psychological Science* (PS; 2003-2013) and *Frontiers in Psychology* (FP; 2010-2013). In this manual download we did not include retractions, errata, and editorials. Finally, we automatically downloaded all HTML articles with the subject "psychology" from the *Public Library Of Science* (PLOS; 2000-

2013), using the rplos R package (Chamberlain, Boettiger, & Ram, 2014).[2] In this automatic process we did not exclude retractions, errata, or editorials. The final sample consisted of 30,717 articles. The number of downloaded articles per journal is given in Table 2.1. To obtain reporting error prevalences for each subfield and for psychology in total, statcheck was used on all downloaded articles.

2.1.3 Statistical analyses

Our population of interest is all APA reported NHST results in the full text of the articles from the eight selected flagship journals in psychology from 1985 until 2013. Our sample includes this entire population. We therefore made no use of inferential statistics, since inferential statistics are only necessary to draw conclusions about populations when having much smaller samples. We restricted ourselves to descriptive statistics; every documented difference or trend entails a difference between or trend in the entire population or subpopulations based on journals. For linear trends we report regression weights and percentages of variance explained to aid interpretation.

2.2 Results

We report the prevalence of reporting inconsistencies at different levels. We document general prevalence of NHST results and present percentages of articles that use NHST per journal and over the years. Because only the five APA journals provided HTMLs for all years from 1985-2013, the overall trends are reported for APA journals only, and do not include results from Psychological Science, PLOS, and Frontiers, which only cover recent years. Reporting inconsistencies are presented both at the level of article and at the level of the individual *p*-value, i.e., the percentage of articles with at least one inconsistency and the average percentage of *p*-values within an article that is inconsistent, respectively. We also describe differences between journals and trends over time.

2.2.1 Percentage of articles with NHST results

Overall, statcheck detected NHST results in 54.4% of the articles, but this percentage differed per journal. The percentage of articles with at least one detected NHST result ranged from 24.1% in PLOS to 85.1% in JPSP (see Table 2.1). This can reflect a difference in the number of null hypothesis significance tests performed, but it could also reflect a difference in the rigor with which the APA reporting standards are followed or how often tables are used to report results. Figure 2.1 shows the percentage of downloaded articles that contained NHST results over the years, averaged over all APA journals (DP, JCCP, JEPG, JPSP, and JAP; dark gray panel), and split up per journal (light gray panels for the APA journals and white panels for the

[2] We note there is a minor difference in the number of search results from the webpage and the package due to default specifications in the rplos package. See also https://github.com/ropensci/rplos/issues/75

non-APA journals). All journals showed an increase in the percentage of articles with APA reported NHST results over the years except for DP and FP, for which this rate remained constant and or declined, respectively. Appendix B lists the number of articles with NSHT results over the years per journal.

Table 2.1

Specifications of the years from which HTML articles were available, the number of downloaded articles per journal, the number of articles with APA reported NHST results, the number of APA reported NHST results, and the median number of APA reported NHST results per article.

Journal	**Subfield**	**Years included**	**# Articles**	**#Articles with NHST results**		**# NHST results**	**Median # NHST results per article with NHST results**
PLOS	General	2000-2013	10,299	2,487	(24.1%)	31,539	9
JPSP	Social	1985-2013	5,108	4,346	(85.1%)	101,621	19
JCCP	Clinical	1985-2013	3,519	2,413	(68.6%)	27,429	8
DP	Developmental	1985-2013	3,379	2,607	(77.2%)	37,658	11
JAP	Applied	1985-2013	2,782	1,638	(58.9%)	15,134	6
PS	General	2003-2013	2,307	1,681	(72.9%)	15,654	8
FP	General	2010-2013	2,139	702	(32.8%)	10,149	10
JEPG	Experimental	1985-2013	1,184	821	(69.3%)	18,921	17
Total			**30,717**	**16,695**	**(54.4%)**	**258,105**	**11**

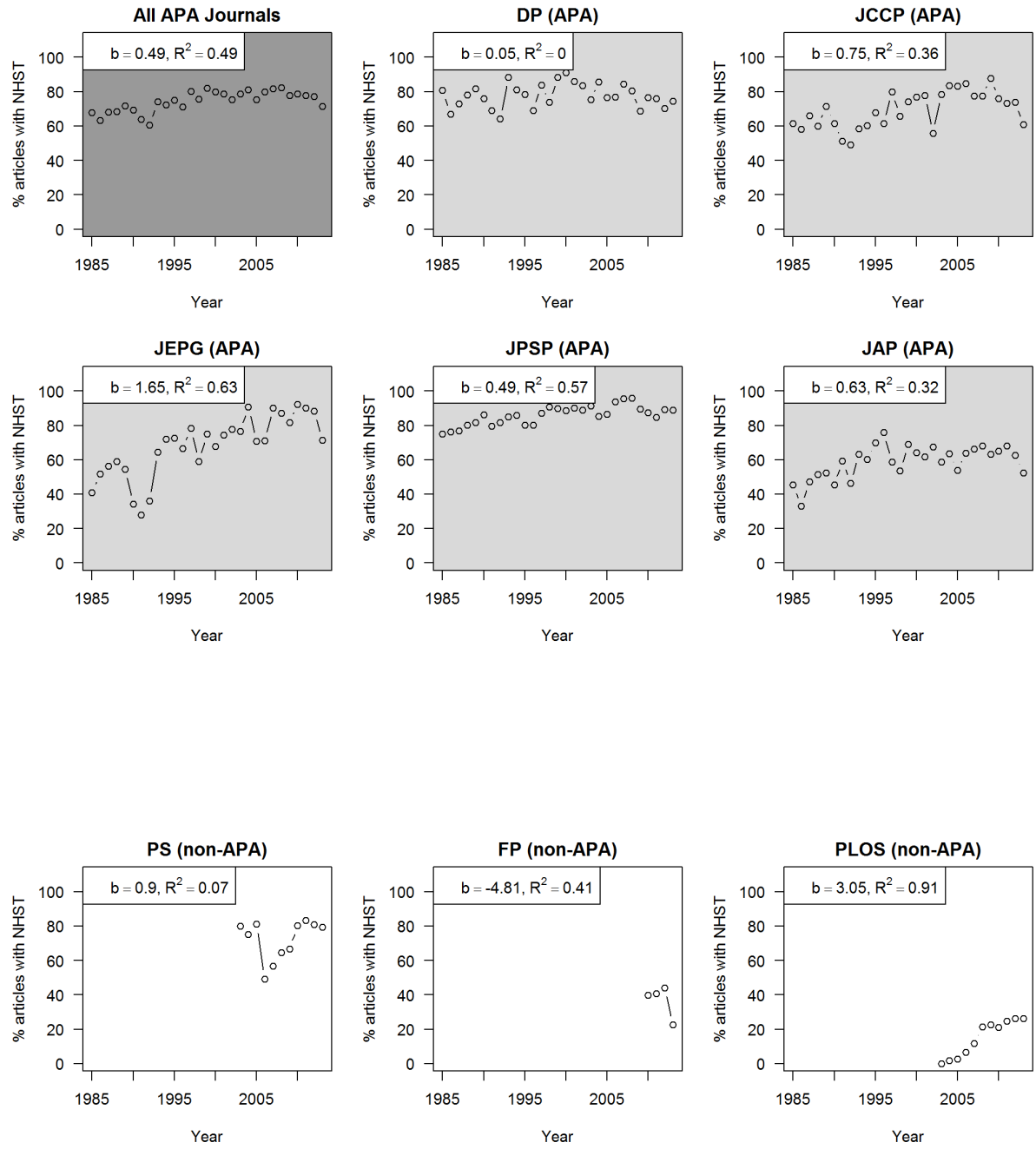


Figure 2.1

The percentage of articles with APA reported NHST results over the years, averaged over all APA journals (DP, JCCP, JEPG, JPSP, and JAP; dark gray panel), and split up per journal (light gray panels for the APA journals and white panels for the non-APA journals). For each trend we report the unstandardized linear regression coefficient (b) and the coefficient of determination (R^2) of the linear trend.

2.2.2 Number of published NHST results over the years

We inspected the development of the average number of APA reported NHST results per article, given that the article contained at least one detectable NHST result (see Figure

2.2). Note that in 1985 the APA manual already required statistics to be reported in the manner that statcheck can read (American Psychological Association, 1983). Hence, any change in retrieved NHST results over time should reflect the actual change in the number of NHST results reported in articles rather than any change in the capability of statcheck to detect results.

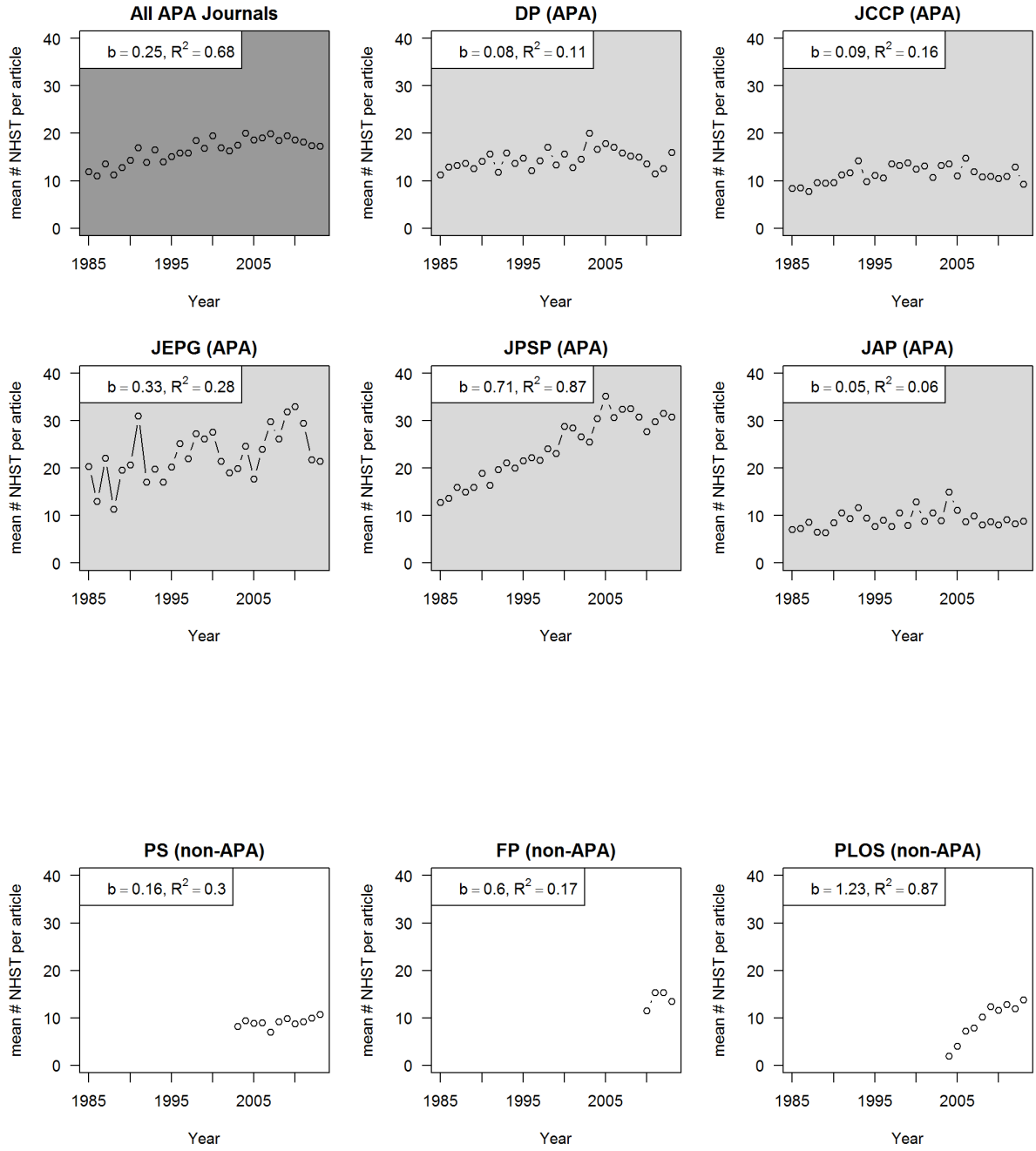


Figure 2.2

The average number of APA reported NHST results per article that contains NHST results over the years, averaged over all APA journals (DP, JCCP, JEPG, JPSP, and JAP; dark gray panel), and split up per journal (light gray panels for the APA journals and white panels for the non-APA journals). For each trend we report the unstandardized linear regression coefficient (b) and the coefficient of determination (R^2) of the linear trend.

Across all APA journals, the number of NHST results per article has increased over the period of 29 years (b = .25, R^2 = .68), with the strongest increases in JEPG and JPSP. These journals went from an average of around 10-15 NHST results per article in 1985 to as much as around 30 results per article on average in 2013. The mean number of NHST results per article remained relatively stable in DP, JCCP, and JAP; over the years, the articles with NHST results in these journals contained on average of ten NHST results. It is hard to say anything definite about trends in PS, FP, and PLOS, since we have only a limited number of years for these journals (the earliest years we have information of are 2003, 2010, and 2004, respectively). Both the increase in the percentage of articles that report NHST results and the increased number of NHST results per article show that NHST is increasingly popular in psychology. It is therefore important that the results of these tests are reported correctly.

2.2.3 General prevalence of inconsistencies

Across all journals and years 49.6% of the articles with NHST results contained at least one inconsistency (8,273 of the 16,695 articles) and 12.9% (2,150) of the articles with NHST results contained at least one gross inconsistency. Furthermore, overall, 9.7% (24,961) of the *p*-values were inconsistent, and 1.4% (3,581) *p*-values were grossly inconsistent. We also calculated the percentage of inconsistencies per article and averaged these percentages over all articles. We call this the "(gross) inconsistency rate". Across journals, the inconsistency rate was 10.6% and the gross inconsistency rate was 1.6%.

2.2.4 Prevalence of inconsistencies per journal

We calculated the prevalence of inconsistencies per journal at two levels. First, we calculated the percentage of articles with NHST results per journal that contained at least one (gross) inconsistency. Second, we calculated the inconsistency rate per journal. The top panel of Figure 2.3 shows the average percentage of articles with at least one (gross) inconsistency, per journal. The journals are ordered from the journal with the highest percentage of articles with an inconsistency to the journal with the least articles with an inconsistency. JPSP showed the highest prevalence of articles with at least one inconsistency (57.6%), followed by JEPG (54.8%). The journals in which the percentage of articles with an inconsistency was lowest are PS and JAP (39.7% and 33.6% respectively). JPSP also had the highest percentage of articles with at least one gross inconsistency (15.8%), this time followed by DP (15.2%). PS had the lowest percentage of articles with gross inconsistencies (6.5%).

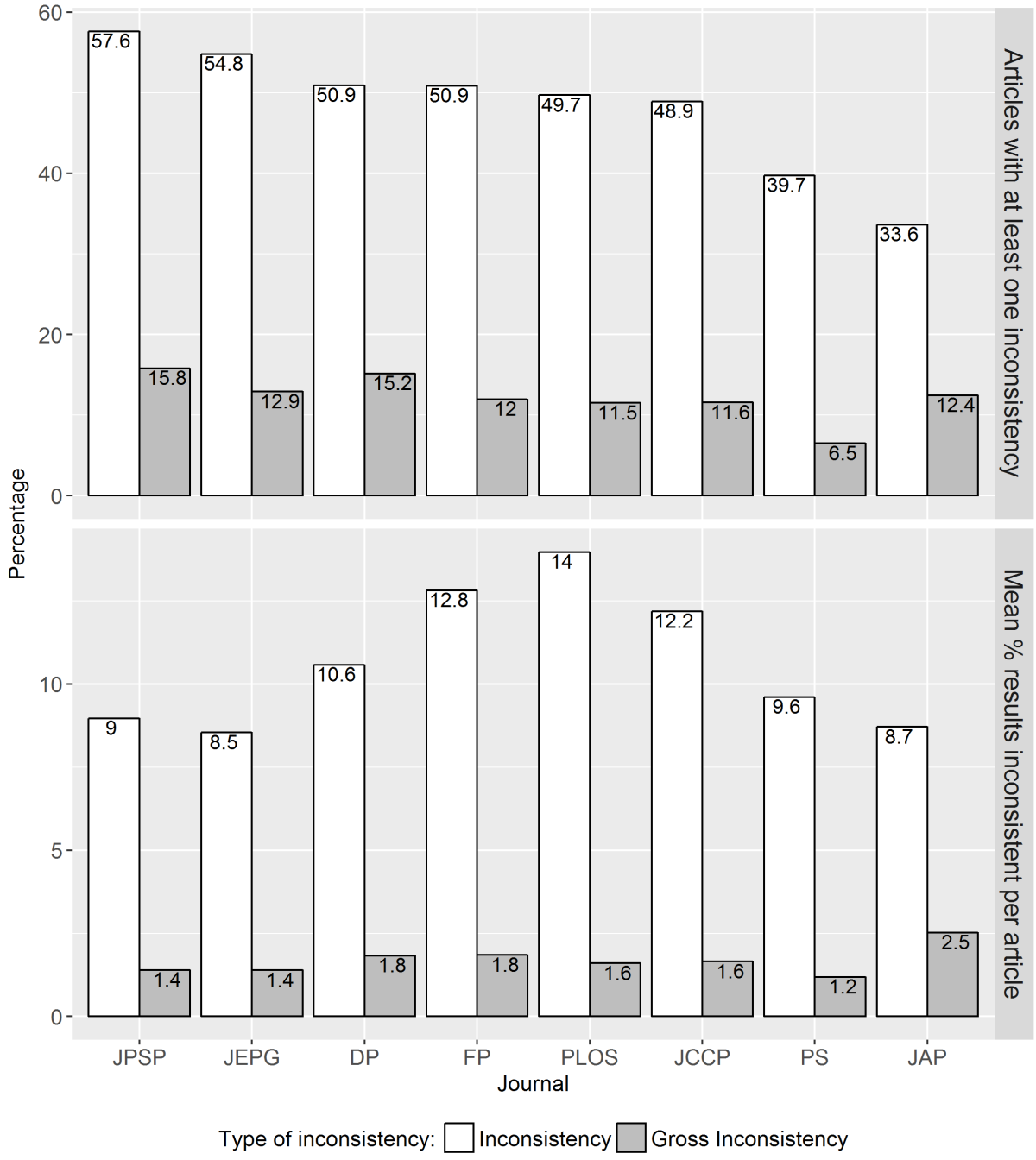


Figure 2.3

The average percentage of articles within a journal with at least one (gross) inconsistency and the average percentage of (grossly) inconsistent p*-values per article, split up by journal. Inconsistencies are depicted in white and gross inconsistencies in grey. For the journals JPSP, JEPG, DP, FP, PLOS, JCCP, PS, and JAP respectively, the number of articles with NHST results is 4346, 821, 2607, 702, 2487, 2413, 1681, 1638, and the average number of NHST results in an article is 23.4, 23.0, 14.4, 14.5, 12.7, 11.4, 9.3, 9.2.*

The inconsistency rate shows a different pattern than the percentage of articles with all inconsistencies. PLOS showed the highest percentage of inconsistent *p*-values per article overall, followed by FP (14.0% and 12.8%, respectively). Furthermore, whereas JPSP was the journal with the highest percentage of articles with inconsistencies, it had one of the lowest probabilities that a *p*-value in an article was inconsistent (9.0%). This discrepancy is caused by a difference between journals in the number of *p*-values per article: the articles in JPSP contain many *p*-values (see Table 2.1, right column). Hence, notwithstanding a low probability of a single *p*-value in an article being inconsistent, the probability that an article contained at least one inconsistent *p*-value was relatively high. The gross inconsistency rate was quite similar over all journals except JAP, in which the gross inconsistency rate was relatively high (2.5%).

2.2.5 Prevalence of inconsistencies over the years

If gross inconsistencies are indicative of QRPs and QRPs have increased over the years, we would expect an increase of gross inconsistencies over the years (see also Leggett et al., 2013). To study this, we inspected the gross inconsistency rate in journals over time. The results are shown in Figure 2.4.

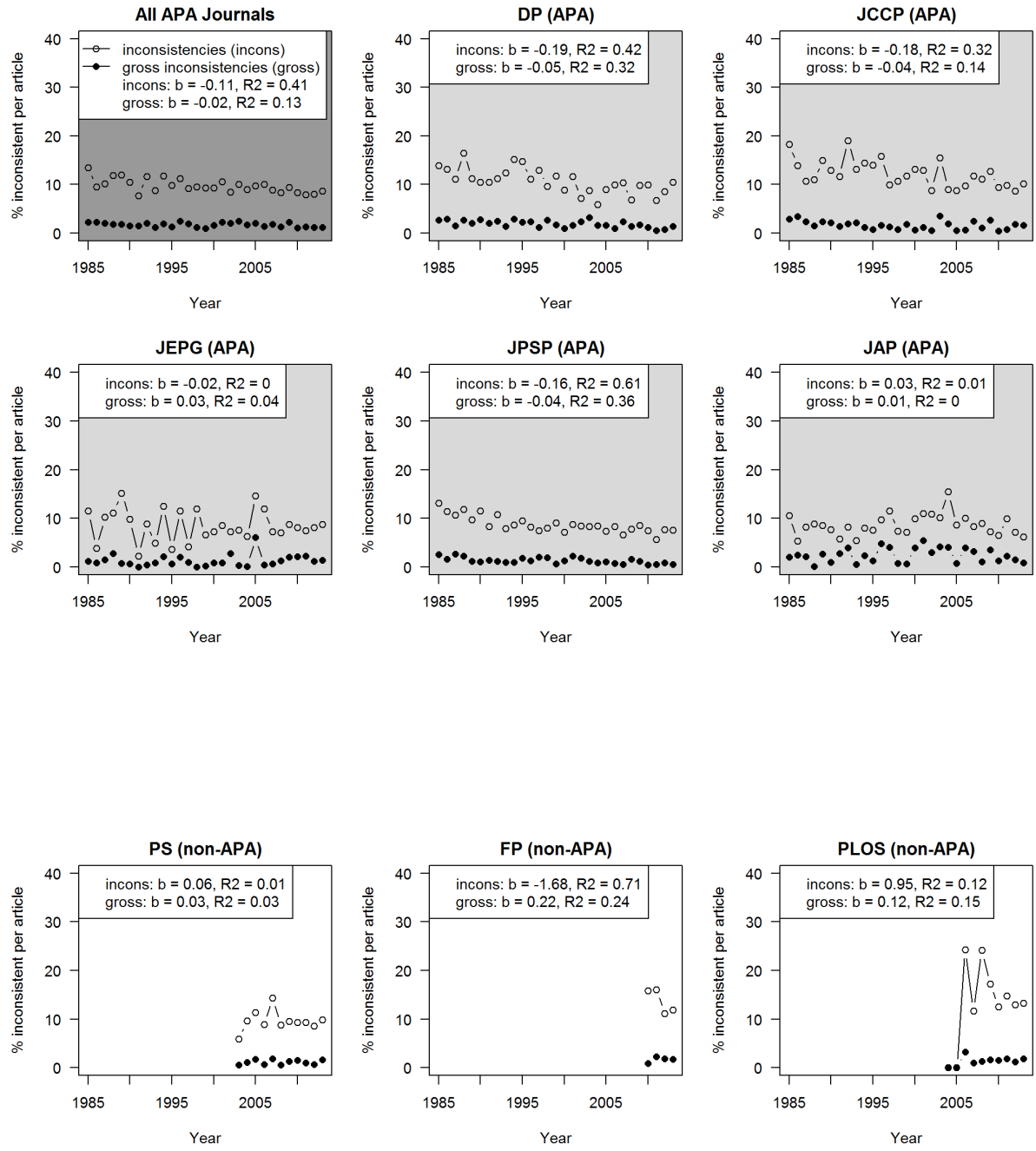


Figure 2.4

Average percentage of inconsistencies (open circles) and gross inconsistencies (solid circles) in an article over the years averaged over all APA journals (DP, JCCP, JEPG, JPSP, and JAP; dark gray panel) and split up per journal (light gray panels for the APA journals and white panels for non-APA journals). The unstandardized regression coefficient b and the coefficient of determination R^2 of the linear trend are shown per journal for both inconsistencies (incons) and gross inconsistencies (gross) over the years.

The number of (gross) inconstencies have decreased or remained stable over the years across the APA journals. In DP, JCCP, JPEG, and JPSP the percentage of all inconsistencies in

an article has decreased over the years. For JAP there is a positive (but very small) regression coefficient for year, indicating an increasing error rate, but the R^2 is close to zero. The same pattern held for the prevalence of gross inconsistencies over the years. DP, JCCP, and JPSP have shown a decrease in gross inconsistencies, in JEPG and JAP the R^2 is very small, and the prevalence seems to have remained practically stable. The trends for PS, FP, and PLOS are hard to interpret given the limited number of years of covarage. Overall, it seems that, contrary to the evidence suggesting that the use of QRPs could be on the rise (Fanelli, 2012; Leggett et al., 2013), neither the inconsistencies nor the gross inconsistencies have shown an increase over time. If anything, the current results reflect a decrease of reporting error prevalences over the years.

We also looked at the development of inconsistencies at the article level. More specifically, we looked at the percentage of articles with at least one inconsistency over the years, averaged over all APA journals (DP, JCCP, JEPG, JPSP, and JAP; dark gray panel in Figure 2.5) and split up per journal (light gray panels for the APA journals and white panels for the non-APA journals in Figure 2.5). Results show that there has been an increase in JEPG and JPSP for the percentage of articles with NHST results that have at least one inconsistency, which is again associated with the increase in the number of NHST results per article in these journals (see Figure 2.2). In DP and JCCP, there was a decrease in articles with an inconsistency. For JAP there is no clear trend; the R^2 is close to zero. A more general trend is evident in the prevalence of articles with gross inconsistencies: in all journals, except PS and PLOS, the percentage of articles with NHST that contain at least one gross inconsistency has been decreasing. Note that the trends for PS, FP, and PLOS are unstable due to the limited number of years we have data for. Overall, it seems that, even though the prevalence of articles with inconsistencies has increased in some journals, the prevalence of articles with gross inconsistencies has shown a decline over the studied period.

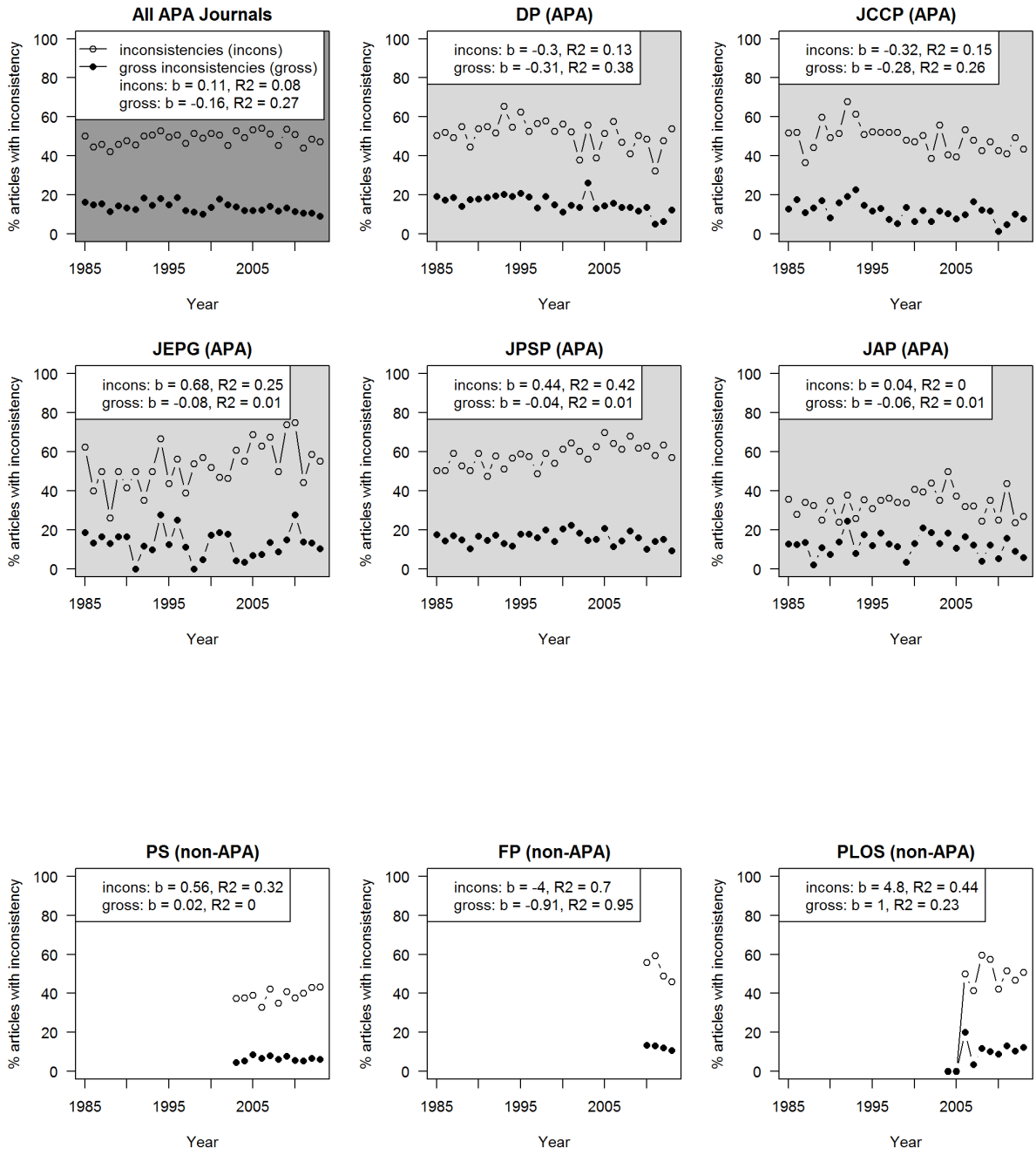


Figure 2.5

Percentage of articles with at least one inconsistency (open circles) or at least one gross inconsistency (solid circles), split up by journal. The unstandardized regression coefficient b and the coefficient of determination R^2 of the linear trend are shown per journal for both inconsistencies (incons) as gross inconsistencies (gross) over the years.

2.2.6 Prevalence of gross inconsistencies in results reported as significant and nonsignificant

We inspected the gross inconsistencies in more detail by comparing the percentage of gross inconsistencies in *p*-values reported as significant and *p*-values reported as nonsignificant. Of all *p*-values reported as significant 1.56% was grossly inconsistent, whereas only .97% of all *p*-values reported as nonsignificant was grossly inconsistent, indicating it is more likely for a *p*-value reported as significant to be a gross inconsistency, than for a *p*-value reported as nonsignificant. We also inspected the prevalence of gross inconsistencies in significant and nonsignificant *p*-values per journal (see Figure 2.6). In all journals, the prevalence of gross inconsistencies is higher in significant *p*-values than in nonsignificant *p*-values (except for FP, in which the prevalence is equal in the two types of *p*-values). This difference in prevalence is highest in JCCP (1.03 percentage point), JAP (.97 percentage point), and JPSP (.83 percentage point) respectively, followed by JEPG (.51 percentage point) and DP (.26 percentage point), and smallest in PLOS (.19 percentage point) and FP (.00 percentage point).

It is hard to interpret the percentages of inconsistencies in significant and nonsignficant *p*-values substantively, since they depend on several factors, such as the specific *p*-value: it seems more likely that a *p*-value of .06 is reported as smaller than .05, than a *p*-value of .78. That is, because journals may differ in the distribution of specific *p*-values we should also be careful in comparing gross inconsistencies in *p*-values reported as significant across journals. Furthermore, without the raw data it is impossible to determine whether it is the *p*-value that is erroneous, or the test statistic or degrees of freedom. As an example of the latter case, a simple typo such as "$F(2,56) = 1.203$, $p < .001$" instead of "$F(2,56) = 12.03$, $p < .001$" produces a gross inconsistency, without the *p*-value being incorrect. Although we cannot interpret the absolute percentages and their differences, the finding that gross inconsistencies are more likely in *p*-values presented as significant than in *p*-values presented as nonsignificant could indicate a systematic bias and is reason for concern.

Figure 2.7 shows the prevalence of gross inconsistencies in significant (solid line) and nonsignificant (dotted line) *p*-values over time, averaged over all journals. The size of the circles represents the total number of significant (open circle) and nonsignificant (solid circle) *p*-values in that particular year. Note that we only have information of PS, FP, and PLOS since 2003, 2010, and 2004, respectively. The prevalence of gross inconsistencies in significant *p*-values seems to decline slightly over the years ($b = -.04$, $R^2 = .65$). The prevalence of the gross inconsistencies in nonsignificant *p*-values does not show any change ($b = .00$, $R^2 = .00$). In short, the potential systematic bias leading to more gross inconsistencies in significant results seems to be present in all journals except for FP, but there is no evidence that this bias is increasing over the years.

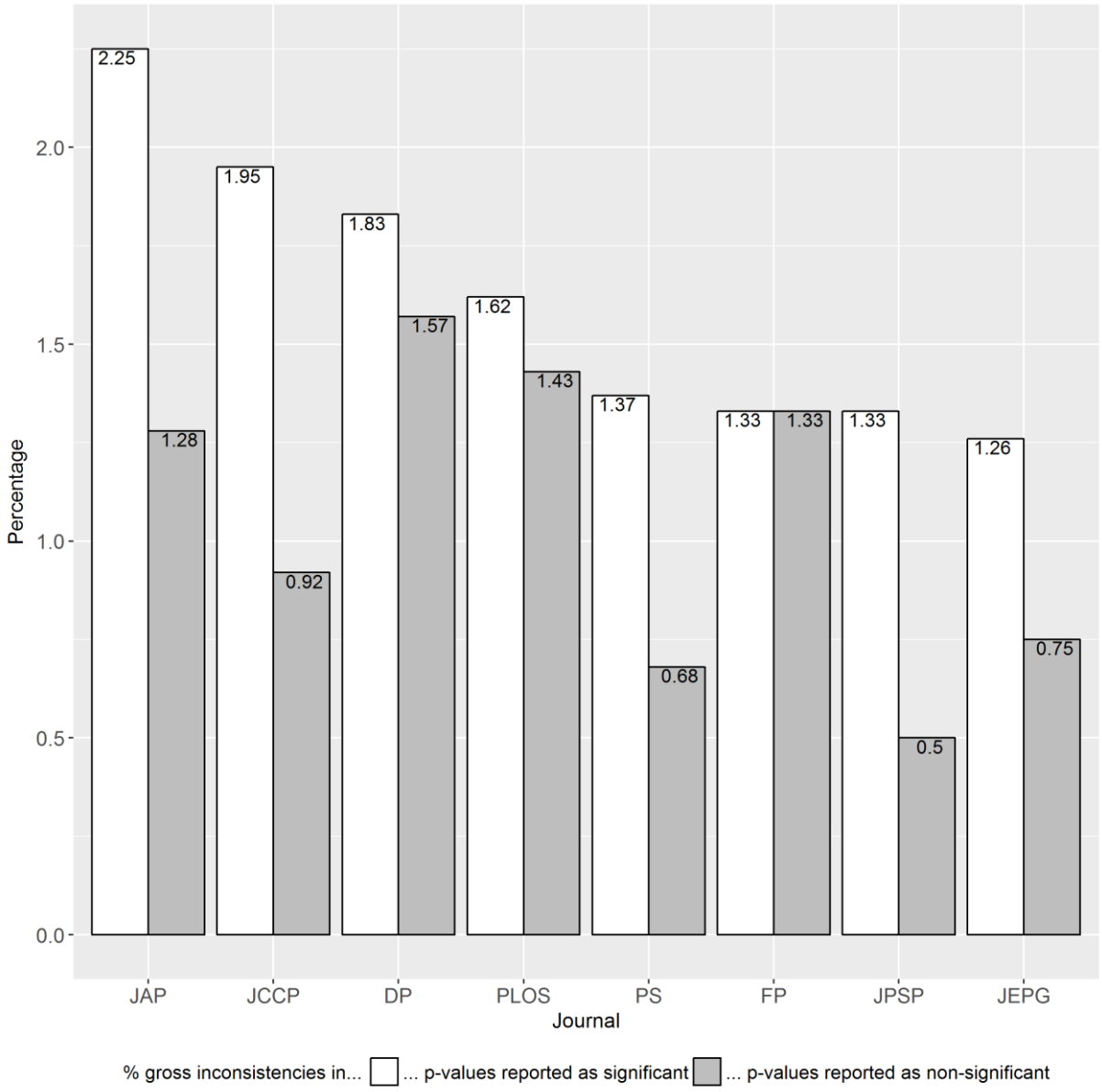


Figure 2.6

The percentage of gross inconsistencies in p*-values reported as significant (white bars) and nonsignificant (gray bars), split up by journal. For the journals JAP, JCCP, DP, PLOS, PS, FP, JPSP, and JEPG respectively, the total number of significant* p*-values was 11654, 21120, 29962, 22071, 12482, 7377, 78889, and 14084, and the total number of nonsignificant* p*-values was 3119, 5558, 6698, 9134, 2936, 2712, 17868, and 4407.*

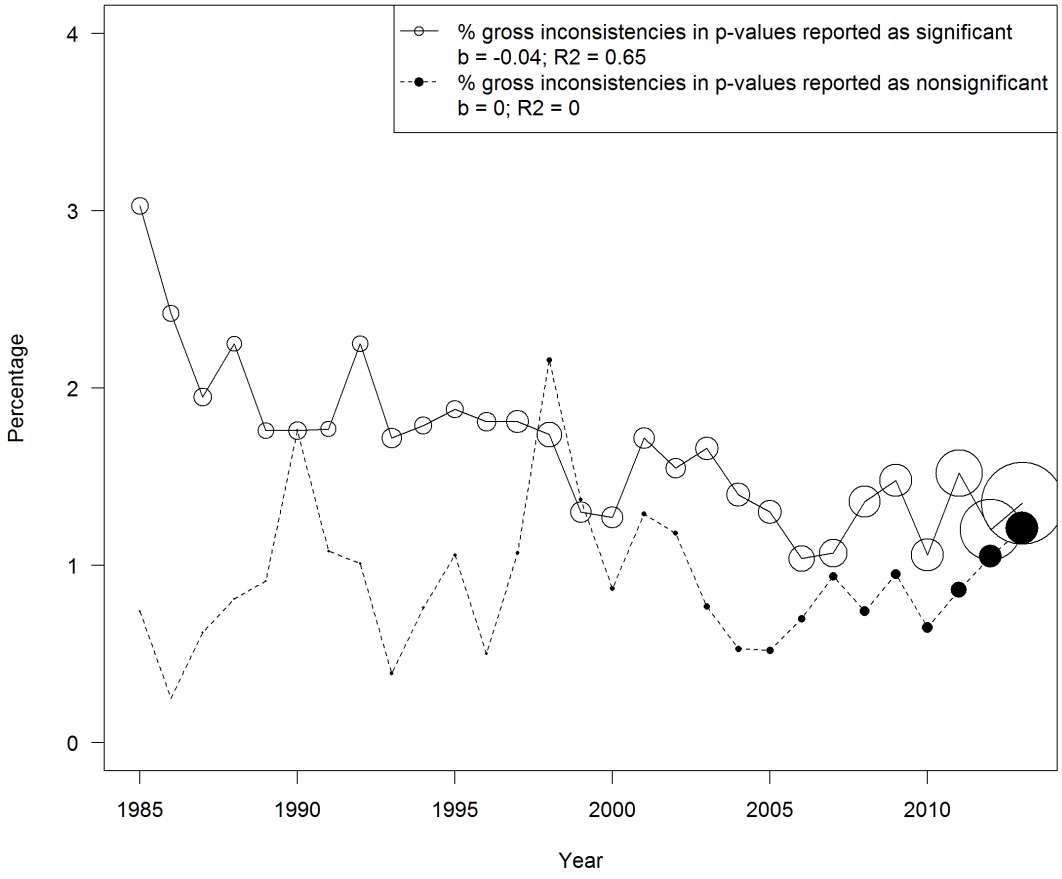


Figure 2.7

The percentage of gross inconsistencies in p*-values reported as significant (solid line) and nonsignificant (dotted line), over the years, averaged over journals. The size of the open and solid circlesrepresents the number of significant and nonsignificant* p*-values in that year, respectively.*

To investigate the consequence of these gross inconsistencies, we compared the percentage of significant results in the reported *p*-values with the percentage of significant results in the computed *p*-values. Averaged over all journals and years, 76.6% of all reported *p*-values were significant. However, only 74.4% of all computed *p*-values were significant, which means that the percentage of significant findings in the investigated literature is overestimated by 2.2 percentage points due to gross inconsistencies.

2.2.7 Prevalence of inconsistencies as found by other studies

Our study can be considered a large replication of several previous studies (Bakker & Wicherts, 2011; Bakker & Wicherts, 2014; Berle & Starcevic, 2007; Caperos & Pardo, 2013; Garcia-Berthou & Alcaraz, 2004; Veldkamp et al., 2014; Wicherts et al., 2011). Table 2.2 shows the prevalence of inconsistent *p*-values as determined by our study and previous studies.

Table 2.2

Prevalence of inconsistencies in the current study and in earlier studies.

Study	Field	# Articles	# Results	% Inconsis-tencies	% Gross inconsis-tencies	% Articles with at least one inconsistency	% Articles with at least one gross inconsistency
Current study	Psychology	30,717	258,105	9.7	1.4	49.6[2]	12.9[2]
Garcia-Berthou and Alcaraz (2004)	Medical	44	244[4]	11.5	0.4	31.5	-
Berle and Starcevic (2007)	Psychiatry	345	5,464	14.3	-	10.1	2.6
Wicherts et al. (2011)	Psychology	49	1,148[1]	4.3	0.9	53.1	14.3
Bakker and Wicherts (2011)	Psychology	333	4,248[3]	11.9	1.3	45.4	12.4
Caperos and Pardo (2013)	Psychology	186	1,212[3]	12.2	2.3	48.0[2]	17.6[2]
Bakker and Wicherts (2014)	Psychology	153[5]	2,667	6.7	1.1	45.1	15.0
Veldkamp et al. (2014)	Psychology	697	8,105	10.6	0.8	63.0	20.5

[1] Only *t*, *F*, and χ^2 values with a $p < .05$.

[2] Number of articles with at least one (gross) inconsistency / number of articles with NHST results.

[3] Only included *t*, *F*, and χ^2 values.

[4] Only exactly reported *p*-values.

[5] Only articles with at least one completely reported *t* or *F* test with a reported *p*-value < .05.

Table 2.2 shows that the estimated percentage of inconsistent results can vary considerably between studies, ranging from 4.3% of the results (Wicherts et al., 2011) to 14.3% of the results (Berle & Starcevic, 2007). The median rate of inconsistent results is 11.1% (1.4 percentage points higher than the 9.7% in the current study). The percentage of gross inconsistencies ranged from .4% (Garcia-Berthou & Alcaraz, 2004) to 2.3% (Caperos & Pardo, 2013), with a median of 1.1% (.3 percentage points lower than the 1.4% found in the current study). The percentage of articles with at least one inconsistency ranged from as low as 10.1% (Berle & Starcevic, 2007) to as high as 63.0% (Veldkamp et al., 2014), with a median of 46.7% (2.9 percentage points lower than the estimated 49.6% in the current study). Finally, the lowest percentage of articles with at least one gross inconsistency is 2.6% (Berle & Starcevic, 2007) and the highest is 20.5% (Veldkamp et al., 2014), with a median of 14.3% (1.4 percentage points higher than the 12.9% found in the current study).

Some of the differences in prevalences could be caused by differences in inclusion criteria. For instance, Bakker and Wicherts (2011) included only *t*, *F*, and χ^2 values; Wicherts et al. (2011) included only *t*, *F*, and χ^2 values of which the reported *p*-value was smaller than .05; Berle and Starcevic (2007) included only exactly reported *p*-values; Bakker and Wicherts (2014) only included completely reported *t* and *F* values. Furthermore, two studies evaluated *p*-values in the medical field (Garcia-Berthou & Alcaraz, 2004) and in psychiatry (Berle & Starcevic, 2007) instead of in psychology. Finally, there can be differences in which *p*-values are counted as inconsistent. For instance, the current study counts $p = .000$ as incorrect, whereas this was not the case in for example Wicherts et al. (2011; see also Appendix A).

Based on Table 2.2 we conclude that our study corroborates earlier findings. The prevalence of reporting inconsistencies is high: almost all studies find that roughly one in ten results is erroneously reported. Even though the percentage of results that is grossly inconsistent is lower, the studies show that a substantial percentage of published articles contain at least one gross inconsistency, which is reason for concern.

2.3 Discussion

In this chapter we investigated the prevalence of reporting errors in eight major journals in psychology using the automated R package statcheck (Epskamp & Nuijten, 2015). Over half of the articles in the six flagship journals reported NHST results that statcheck was able to retrieve. Notwithstanding the many debates on the downsides of NHST (see, e.g., Fidler & Cumming, 2005; Wagenmakers, 2007), the use of NHST in psychology appears to have increased from 1985-2013 (see Figure 2.1 and 2.2), although this increase can also reflect an increase in adherance to APA reporting standards. Our findings show that in general the prevalence of reporting inconsistencies in six flag ship psychology journals is substantial. Roughly half of all articles with NHST results contained at least one inconsistency and about 13% contained a gross inconsistency that may have affected the statistical conclusion. At the

level of individual *p*-values we found that on average 10.6% of the *p*-values in an article were inconsistent, whereas 1.6% of the *p*-values were grossly inconsistent.

Contrary to what one would expect based on the suggestion that QRPs have been on the rise (Leggett et al., 2013), we found no general increase in the prevalence of inconsistent *p*-values in the studied journals from 1985 to 2013. When focusing on inconsistencies at the article level, we only found an increase in the percentage of articles with NHST results that showed at least one inconsistency for JEPG and JPSP. Note this was associated with clear increases in the number of reported NHST results per article in these journals. Furthermore, we did not find an increase in gross inconsistencies in any of the journals. If anything, we saw that the prevalence of articles with gross inconsistencies has been decreasing since 1985, albeit only slightly. We also found no increase in the prevalence of gross inconsistencies in *p*-values that were reported as significant as compared to gross inconsistencies in *p*-values reported as nonsignificant. This is at odds with the notion that QRPs in general and reporting errors in particular have been increasing in the last decades. On the other hand, the stability or decrease in reporting errors is in line with research showing no trend in the proportion of published errata, which implies that there is also no trend in the proportion of articles with (reporting) errors (Fanelli, 2013).

Furthermore, we found no evidence that inconsistencies are more prevalent in JPSP than in other journals. The (gross) inconsistency rate was not the highest in JPSP. The prevalence of (gross) inconsistencies has been declining in JPSP, as it did in other journals. We did find that JPSP showed a higher prevalence of articles with at least one inconsistency than other journals, but this was associated with the higher number of NSHT results per article in JPSP. Hence our findings are not in line with the previous findings that JPSP shows a higher (increase in) inconsistency rate (Leggett et al., 2013). Since statcheck cannot distinguish between *p*-values pertaining to core hypotheses and *p*-values pertaining to, for example, manipulation checks, it is hard to interpret the differences in inconsistencies between fields and the implications of these differences. To warrant such a conclusion the inconsistencies would have to be manually analyzed within the context of the papers containing the inconsistencies.

We also found that gross inconsistencies are more prevalent in *p*-values reported as significant than in *p*-values reported as nonsignificant. This could suggest a systematic bias favoring significant results, potentially leading to an excess of false positives in the literature. The higher prevalence of gross inconsistencies in significant *p*-values versus nonsignificant *p*-values was highest in JCCP, JAP, and JPSP, and lowest in PLOS and FP. Note again that we do not know the hypotheses underlying these *p*-values. It is possible that in some cases a nonsignificant *p*-value would be in line with a hypothesis and thus in line with the researcher's predictions. Our data do not speak to the causes of this overrepresentation of significant results. Perhaps these *p*-values are intentionally rounded down (a practice that 20% of the

surveyed psychological researchers admitted to; John et al., 2012) to convince the reviewers and other readers of an effect. Or perhaps researchers fail to double check significantly reported *p*-values, because they are in line with their expectations, hence leaving such reporting errors more likely to remain undetected. It is also possible that the cause of the overrepresentation of falsely significant results lies with publication bias: perhaps researchers report significant *p*-values as nonsignificant just as often as vice versa, but in the process of publication, only the (accidentally) significant *p*-values get published.

There are two main limitations in our study. First, by using the automated procedure statcheck to detect reporting inconsistencies, our sample did not include NHST results that were not reported exactly according to APA format or results reported in tables. However, based on the validity study and on earlier results (Bakker & Wicherts, 2011), we conclude that there does not seem to be a difference in the prevalence of reporting inconsistencies between results in APA format and results that are not exactly in APA format (see Appendix A). The validity study did suggest, however, that statcheck might slightly overestimate the number of inconsistencies. One reason could be that statcheck cannot correctly evaluate *p*-values that were adjusted for multiple testing. However, we found that these adjustments are rarely used. Notably, the term "Bonferroni" was mentioned in a meager 0.3% of the 30,717 papers.[3] This finding is interesting in itself; with a median number of 11 NHST results per paper, most papers report multiple *p*-values. Without any correction for multiple testing, this suggests that overall Type I error rates in the eight psychology journals are already higher than the nominal level of .05. Nevertheless, the effect of adjustments of *p*-values on the error estimates from statcheck is expected to be small. We therefore conclude that, as long as the results are interpreted with care, statcheck provides a good method to analyze vast amounts of literature to locate reporting inconsistencies. Future developments of statcheck could focus on taking into account corrections for multiple testing and results reported in tables or with effect sizes reported between the test statistic and *p*-value.

The second limitation of our study is that we chose to limit our sample to only a selection of flagship journals from several sub disciplines of psychology. It is possible that the prevalence of inconsistencies in these journals is not representative for the psychological literature. For instance, it has been found that journals with lower impact factors have a higher prevalence of reporting inconsistencies than high impact journals (Bakker & Wicherts, 2011). In this study we avoid conclusions about psychology in general, but treat the APA reported NHST results in the full text of the articles from journals we selected as the population of interest (which made statistical inference superfluous). All conclusions in this paper therefore hold for the APA reported NHST results in the eight selected journals. Nevertheless, the relatively high impact factors of these journals attest to the relevance of the current study.

[3] But see also Chapter 3

There are several possible solutions to the problem of reporting inconsistencies. First, researchers can check their own papers before submitting, either by hand or with the R package statcheck.[4] Editors and reviewers could also make use of statcheck to quickly flag possible reporting inconsistencies in a submission, after which the flagged results can be checked by hand. This should reduce erroneous conclusions caused by gross inconsistencies. Checking articles with statcheck can also prevent such inconsistencies from distorting meta-analyses or analyses of *p*-value distributions (Simonsohn et al., 2014; van Assen et al., 2015). This solution would be in line with the notion of Analytic Review (Sakaluk, Williams, & Biernat, 2014), in which a reviewer receives the data file and syntax of a manuscript to check if the reported analyses were actually conducted and reported correctly. One of the main concerns about Analytic Review is that it would take reviewers a lot of additional work. The use of statcheck in Analytic Review could reduce this workload substantially.

Second, the prevalence of inconsistencies might decrease if co-authors check each other's work, a so-called "co-pilot model" (Wicherts, 2011). In recent research (Veldkamp et al., 2014) this idea has been investigated by relating the probability that a *p*-value was inconsistent to six different co-piloting activities (e.g., multiple authors conducting the statistical analyses). Veldkamp et al. did not find direct evidence for a relation between co-piloting and reduced prevalence of reporting errors. However, the investigated co-pilot activities did not explicitly include the actual checking of each other's *p*-values, hence we do not rule out the possibility that reporting errors would decrease if co-authors double checked *p*-values.

Third, it has been found that reporting errors are related to reluctance to share data (Wicherts et al., 2011; but see Deriemaecker et al., in preparation). Although any causal relation cannot be established, a solution might be to require open data by default, allowing exceptions only when explicit reasons are available for not sharing. Subsequently, researchers know their data could be checked and may feel inclined to double check the result section before publishing the paper. Besides a possible reduction in reporting errors, sharing data has many other advantages. Sharing data for instance facilitates aggregating data for better effect size estimates, enable reanalyzing published articles, and increase credibility of scientific findings (see also Nosek, Spies, & Motyl, 2012; Sakaluk et al., 2014; Wicherts, 2013; Wicherts & Bakker, 2012). The APA already requires data to be available for verification purposes (American Psychological Association, 2010, p. 240), many journals explicitly encourage data sharing in their policies, and the journal Psychological Science has started to award badges to papers of which the data are publicly available. Despite these policies and encouragements, raw data are still rarely available (Alsheikh-Ali, Qureshi, Al-Mallah, & Ioannidis, 2011). One objection that has been raised is that due to privacy concerns data cannot be made publicly

[4] And see the web application at http://statcheck.io

available (see e.g, Finkel, Eastwick, & Reis, 2015). Even though this can be a legitimate concern for some studies with particularly sensitive data, these are exceptions; the data of most psychology studies could be published without risks (Nosek et al., 2012).

To find a successful solution to the substantial prevalence of reporting errors, more research is needed on how reporting errors arise. It is important to know whether reporting inconsistencies are mere sloppiness or whether they are intentional. We found that the large majority of inconsistencies were not gross inconsistencies around $p = .05$, but inconsistencies that did not directly influence any statistical conclusion. Rounding down a *p*-value of, say, .38 down to .37 does not seem to be in the direct interest of the researcher, suggesting that the majority of inconsistencies is accidental. On the other hand, we did find that the large majority of grossly inconsistent *p*-values were nonsignificant *p*-values that were presented as significant, instead of vice versa. This seems to indicate a systematic bias that causes an overrepresentation of significant results in the literature. Whatever the cause of this overrepresentation might be, there seems to be too much focus on getting "perfect", significant results (see also Giner-Sorolla, 2012). Considering that the ubiquitous significance level of .05 is arbitrary, and that there is a vast amount of critique on NHST in general (see, e.g., Cohen, 1994; Fidler & Cumming, 2005; Krueger, 2001; Rozeboom, 1960; Wagenmakers, 2007), it should be clear that it is more important that *p*-values are accurately reported than that they are below .05.

There are many more interesting aspects of the collected 258,105 *p*-values that could be investigated, but this is beyond the scope of this chapter. In another paper, the nonsignificant test results from this dataset are investigated for false negatives (Hartgerink, van Assen, & Wicherts, 2017). Here a method is used to detect false negatives and the results indicate 2 out of 3 papers with nonsignificant test results might contain false negative results. This is only one out of the many possibilities and we publicly share the anonymized data on our Open Science Framework page (https://osf.io/gdr4q/) to encourage further research.

Our study illustrates that science is done by humans, and humans easily make mistakes. However, the prevalence of inconsistent *p*-values in eight major journals in psychology has generally been stable over the years, or even declining. Hopefully, statcheck can contribute to further reducing the prevalence of reporting inconsistencies in psychology.

2.4 Appendix A: Results Validity Check Statcheck

Here we investigate the validity of the R program 'statcheck' (Epskamp & Nuijten, 2015) by comparing the results of statcheck with the results of a study in which all statistics were manually retrieved, recalculated, and verified (Wicherts et al., 2011).

2.4.1 Method

2.4.1.1 *Sample*

We used statcheck to scan the same 49 articles from the Journal of Experimental Psychology: Learning, Memory, and Cognition (JEP:LMC) and the Journal of Personality and Social Psychology (JPSP) that have been manually checked for reporting errors in Wicherts et al., who also double checked each reported error after it had been uncovered. The inclusion criteria for the statistical results to check for inconsistencies differed slightly between the study of Wicherts et al. and statcheck (see Table 2.3).

Table 1.3

Inclusion criteria for the statistical results to check for inconsistencies in Wicherts et al. and statcheck.

Wicherts et al.	**Statcheck**
$p < .05$	$p < .05$
t, *F*, χ^2	*t*, *F*, χ^2
complete (test statistic, DF, p)	APA (test statistic, DF, p)
main text or table in result section	-
NHST	-

Both in Wicherts et al. and in this validity study only *p*-values smaller than .05 and only results from *t*, *F*, or χ^2 tests were included. Wicherts et al. required the result to be reported completely. Statcheck had the equivalent, but stricter criterion that the results had to be reported exactly according to APA guidelines (in general: test statistic (degrees of freedom) =/</> …, *p* =/</>…). Furthermore, Wicherts et al. included all results reported in the text or a table in the results section of an article. Statcheck did not distinguish between different sections in a paper, but included all complete results in APA style. This, in practice, often excludes results reported in a table. Finally, Wicherts et al. stated that they only evaluated

results of NHST. Statcheck did not explicitly have this criterion, but implicitly APA reported results of a *t*, *F*, or χ^2 test will always be a NHST result.

2.4.1.2 *Procedure*

We ran statcheck on the 49 articles twice: once in default mode, and once with an automatic one-tailed test detection. The one-tailed test detection works as follows: if the words “one-tailed”, “one-sided”, or “directional” (with various spacing or punctuation) are mentioned in the article *and* a result is not an inconsistency if it is a one-tailed test, the result is counted as correct. From the complete statcheck results, we selected the cases in which the test statistic was *t*, *F*, or χ^2, and in where $p < .05$.

2.4.2 Results

2.4.2.1 *Descriptives*

Table 2.4 below shows the number of extracted statistics and the number of identified errors for both Wicherts et al., statcheck in default mode, and statcheck with the automatic one-tailed test detection.

Table 2.4

The number of extracted statistics and the number of identified errors for both Wicherts et al. and statcheck (with automatic one-tailed test detection).

	Wicherts et al.	**statcheck**	**statcheck with one-tailed test detection**
# articles	49	43	43
# results	1148	775 (67.5%)	775 (67.5%)
# inconsistencies	49 (4.3%)	70 (9.0%)	56 (7.2%)
# papers with at least one inconsistency	23 (46.9%)	23 (53.5%)[1]	21 (48.8%)[1]
# gross inconsistencies	10 (0.9%)	17 (2.3%)	8 (1.0%)
# papers with at least one gross inconsistency	7 (14.3%)	10 (23.3%)[1]	5 (11.6%)[1]

[1] Number of articles with at least one (gross) inconsistency / number of articles with NHST results

Wicherts et al. extracted 1,148 results from the 49 articles, whereas statcheck extracted 775 results (67.5%). Even though statcheck found fewer results, it found relatively more reporting errors (4.3% of all results in Wicherts et al. versus 9.0% or 7.2% of all results in statcheck, without or with one-tailed detection respectively). In the next sections we will identify possible causes for these differences.

2.4.2.2 *Explanations for discrepancies in the number of extracted statistics*

We found that in 13 articles statcheck reported the exact same amount of statistics as Wicherts et al. In 23 articles Wicherts et al. found more statistics than statcheck, and in 13 articles statcheck found more results than Wicherts et al. Table 2.5 shows the explanations for these discrepancies.

Table 2.5

Explanation of the discrepancies between the number of results that Wicherts et al. and statcheck extracted.

	Type of discrepancy	# Articles	# Results	Example
More results extracted by Wicherts et al.	Value between test statistic and *p*-value	11	201	F1(1, 31) = 4.50, MSE = 22.013, p <.05
	Table (incomplete result)	8	150	
	Result in sentence	3	8	F(1, 15) = 19.9 and 5.16, p <.001 and p <.05, respectively
	Non-APA	5	49	F(1. 47) = 45.98, p <.01; F[1, 95] = 18.11, p <.001; F(l, 76) = 23.95, p <.001; no p value reported
	Article retracted	1	28	
More results extracted by statcheck	G^2 statistic included as χ^2 statistic	1	2	$\Delta G^2(1) = 6.53$, p =.011
	Footnote	12	31	
	Error Wicherts et al.: overlooked result	2	2	
	Inexact test statistic	1	1	
	Not in result section	9	27	Result in materials, procedure, discussion etc.
Total # extracted results Wicherts et al.		**49**	**1148**	
Total # extracted results statcheck		**43**	**775**	

Most of the results that statcheck missed were results that were not reported completely (e.g., results in tables) or not exactly according to the APA format (e.g., an effect size reported in between the test statistic and the *p*-value, or the results being reported in a sentence). Furthermore, one article in the sample of Wicherts et al. has been retracted since 2011, and we could not download it anymore; its 28 *p*-values were not included in the statcheck validity study.

Most of the results that were only included by statcheck but not by Wicherts et al. were results that were that were not reported in the result section but in footnotes, in the method section, or in the discussion. Wicherts et al. did not take these results into account; their explicit inclusion criterion was that the result had to be in the text or in a table in the results section of a paper. Statcheck could not make this distinction and included results independent from their location. Furthermore, Wicherts et al. did not include the two G^2 statistics that statcheck counted as χ^2 statistics. Statcheck also included an inexactly reported *F*-statistic that Wicherts et al. excluded, because it referred to multiple tests. Finally, we found two results that fitted their inclusion criteria, but were inadvertently not included by Wicherts et al. sample.

2.4.2.3 *Explanations for discrepancies in the number of identified inconsistencies*

There were discrepancies in the number of (gross) inconsistencies that Wicherts et al. and statcheck found. Table 2.6 explains these inconsistencies in detail. In 13 cases Wicherts et al. found more errors than statcheck (with default options). However, all these cases were results that statcheck did not scan for one of the reasons mentioned above. There are no other cases in which Wicherts et al. found more errors. The use of default statcheck did not highlight any false negatives.

Table 2.6

Explanation of the discrepancies between the number of inconsistencies found by Wicherts et al. and statcheck (with automatic one-tailed test detection).

		Statcheck		Statcheck with one-tailed test detection	
	Category Inconsistency	**# Articles**	**# Results**	**# Articles**	**# Results**
More inconsistencies found by Wicherts et al.	Not scanned by statcheck	8	13	8	13
	Wrongly marked as one-tailed	0	0	3	6
More inconsistencies found by statcheck	*p* = .000 counted as incorrect	1	7	1	7
	One-tailed	4	9	1	1
	Not checked by Wicherts et al.	5	7	5	7
	Huyn-Feldt correction	2	11	2	11
Total # inconsistencies Wicherts et al.			**49**		**49**
Total # inconsistencies statcheck			**70**		**56**

The default statcheck did, however, find 34 false positives (i.e., it marked results as inconsistent whereas Wicherts et al. did not). Closer inspection of these cases highlighted four main causes. First, seven cases were not included in the sample of Wicherts et al. Second, seven of the results that statcheck classified as an error, but Wicherts et al. did not, were results in which the *p*-value was reported to be zero (*p* = .000). Wicherts et al. counted this as correct, in cases where rounding would indeed render *p* = .000. However, statcheck counts this as inconsistent, because a *p*-value this small should be reported as p < .001, but not as *p* = .000 (American Psychological Association, 2010, p. 114). Third, there were eleven cases (in two articles) in which the *p*-value was inconsistent due to a Huyn-Feldt correction, which statcheck cannot take into account. Fourthly, there were nine cases in which the reported *p*-value was one-tailed and therefore twice as low as statcheck computed.

The discrepancies in the gross inconsistencies between the default statcheck and Wicherts et al. were due to seven one-tailed tests (see Table 2.7). Because of these one-tailed tests, statcheck gives an exaggerated image of how many inconsistencies there are in the literature. Therefore, we also inspect the results of statcheck with the one-tailed test detection.

When statcheck uses the one-tailed test detection all but one one-tailed tests previously marked as inconsistent, are now categorized as correct (see Table 2.6 and Table 2.7)[5]. The one-tailed test detection does result in six more false negatives, in which an inconsistent two-tailed test is counted as correct (see Table 2.6). Overall, statcheck now detected 56 inconsistencies in 775 *p*-values (7.2%) and 8 gross inconsistencies (1.0%), which is closer to the inconsistency prevalence found by Wicherts et al. (4.3% and .9%, respectively) than without the one-tailed test detection. In sum, statcheck performs better with the one-tailed test detection.

2.4.2.4 *Inter-rater reliability manual vs. statcheck*

We also calculated the inter-rater reliability between the manual coding of inconsistencies and gross inconsistencies in Wicherts et al. and the automatic coding in statcheck. We distinguished between three different scenarios: in the first statcheck ran in default mode (without one-tailed test detection), in the second the automatic one-tailed test detection in statcheck was switched on, and in the last we ran statcheck with the automatic one-tailed test detection and we excluded cases in which *p* was reported as *p* = .000, since this was not counted as an inconsistency in Wicherts et al., but statcheck is intentionally programmed to see this as an inconsistency (since *p* cannot be zero and it should have been

[5] The only one-tailed test that is still counted by statcheck as inconsistent, is a result that is reported as one-tailed and has a rounded test statistic: t(14) = 2.0, p <.03. The correct rounding of test statistics is not incorporated in the automatic one-tailed test detection, but this will be incorporated in the next version. For now, this will not bias the results that much, since these are rare cases.

reported as $p < .001$). In all three scenarios we only included *p*-values that were rated both by Wicherts et al. and statcheck.

Table 2.8 shows the inter-rater reliabilities for the inconsistencies and gross inconsistencies in the three scenarios. If statcheck is ran without one-tailed test detection, Cohen's kappa for the inconsistencies is .71 and for the gross inconsistencies .74. If we turn on the automatic one-tailed test detection, Cohen's kappa for the gross inconsistencies increases to .89, but it slightly decreases for the inconsistencies to .69. Note, however, there are fewer *p*-values that statcheck wrongly marked as inconsistent with the one-tailed test detection (see Table 2.5). When both the one-tailed detection is switched on and we exclude cases in which *p* is reported as $p = .000$, Cohen's kappa for the inconsistencies increases to .76, and remains at .89 for the gross inconsistencies.

Table 2.7

Explanation of the discrepancies between the number of gross inconsistencies found by Wicherts et al. and statcheck (with automatic one-tailed test detection).

		Statcheck		Statcheck with one-tailed test detection	
	Category gross inconsistency	**# Articles**	**# Results**	**# Articles**	**# Results**
More gross inconsistencies found by Wicherts et al.	Wrongly marked as one-tailed	0	0	2	2
More gross inconsistencies found by statcheck	One-tailed	4	7	0	0
Total # gross inconsistencies Wicherts et al.			**10**		**10**
Total # gross inconsistencies statcheck			**17**		**8**

Table 2.8

The inter-rater reliability expressed in Cohen's kappa between the manual coding in Wicherts et al. (2011) and the automatic coding in statcheck without or with automatic one-tailed detection, and with and without exclusion of p *= .000.*

	Inconsistencies	**Gross Inconsistencies**
No automatic one-tailed test detection	.71	.74
Automatic one-tailed test detection	.69	.89
Automatic one-tailed test detection & exclude *p* = .000	.76	.89

2.4.3 Discussion

In this validity check we compared the results of Wicherts et al. (2011) with the results of the default version of statcheck and statcheck with automatic one-tailed test detection. The results show that statcheck extracted 67.5% of the manually retrieved results. The main reason for this is that statcheck could not read results that were not reported completely or not in APA style. Even though statcheck included fewer results than Wicherts et al., it found more inconsistencies. These inconsistencies were mainly one-tailed tests that were counted as inconsistent. Specifically, Wicherts et al. found 49 of the 1148 results (4.3%) to be inconsistent and 10 to be grossly inconsistent (.9%), whereas statcheck found 70 of the 775 results (9.0%) to be inconsistent and 17 (2.2%) to be grossly inconsistent. In other words, statcheck found an inconsistency rate that was 4.7 percentage point higher than the one found in a manual search and a gross inconsistency rate that is 1.3 percentage point higher. The inter-rater reliability for inconsistencies was .71 and for gross inconsistencies .74.

When statcheck was run with automatic one-tailed test detection, it still found more errors than did Wicherts et al. but the difference was smaller. Now statcheck found that 56 of 775 results (7.2%) to be inconsistent and 8 results (1.0%) to be grossly inconsistent. That means that with automatic one-tailed test detection statcheck found an inconsistency rate of only 2.9 percentage point higher than the one found in a manual search and a gross inconsistency rate of .1 percentage point higher. The inter-rater reliability for gross inconsistencies was as high as .89, but decreased slightly for inconsistencies to .69. However, since there are fewer *p*-values wrongly marked as inconsistent with the automatic one-tailed test detection, we advise users to use this option when searching for reporting inconsistencies.

The main limitation of statcheck is that it seems to give an overestimation of the number of inconsistencies in a sample. A large part of these false positives was due to the

conscious choice to count $p = .000$ as inconsistent. If we exclude these cases, the inter-rater reliability for inconsistencies goes up to .76, and remains .89 for gross inconsistencies (with automatic one-tailed test detection). Furthermore, the false positives caused by one-tailed tests are mostly solved by statcheck's one-tailed test detection. That leaves only the false positives due to *p*-values adjusted for multiple testing, eventually resulting in only a slight overestimation of the inconsistencies. Herein lies a possibility for future improvement of the program.

In conclusion, since statcheck slightly overestimated the prevalence of inconsistencies in our study, its results should be interpreted with care. We also advise against using statcheck blindly to point out mistakes in a single article. The main two usages of statcheck are 1) to give an overall indication of the prevalence of inconsistencies in a large amount of literature, and 2) to give a first indication of inconsistent *p*-values in a single article, after which the results should be checked by hand. The final verdict on whether a result is erroneous should be based on careful consideration by an expert.

2.5 Appendix B: Additional Analyses

2.5.1 Number of articles with NHST results

Figure 2.8 shows the number of articles that contain NHST results over the years averaged over all APA journals (DP, JCCP, JEPG, JPSP, and JAP; dark gray panel) and split up per journal (light gray panels for the APA journals and white panels for the non-APA journals). The number of articles with NHST results seems to remain relatively stable over the years in JCCP and JAP. JPSP has published fewer articles with NHST results over the years. In DP and JEPG the number of articles with NHST results increased over the years. The newer journals PS, FP, and especially PLOS show a steep increase in articles with NHST results in the last few years.

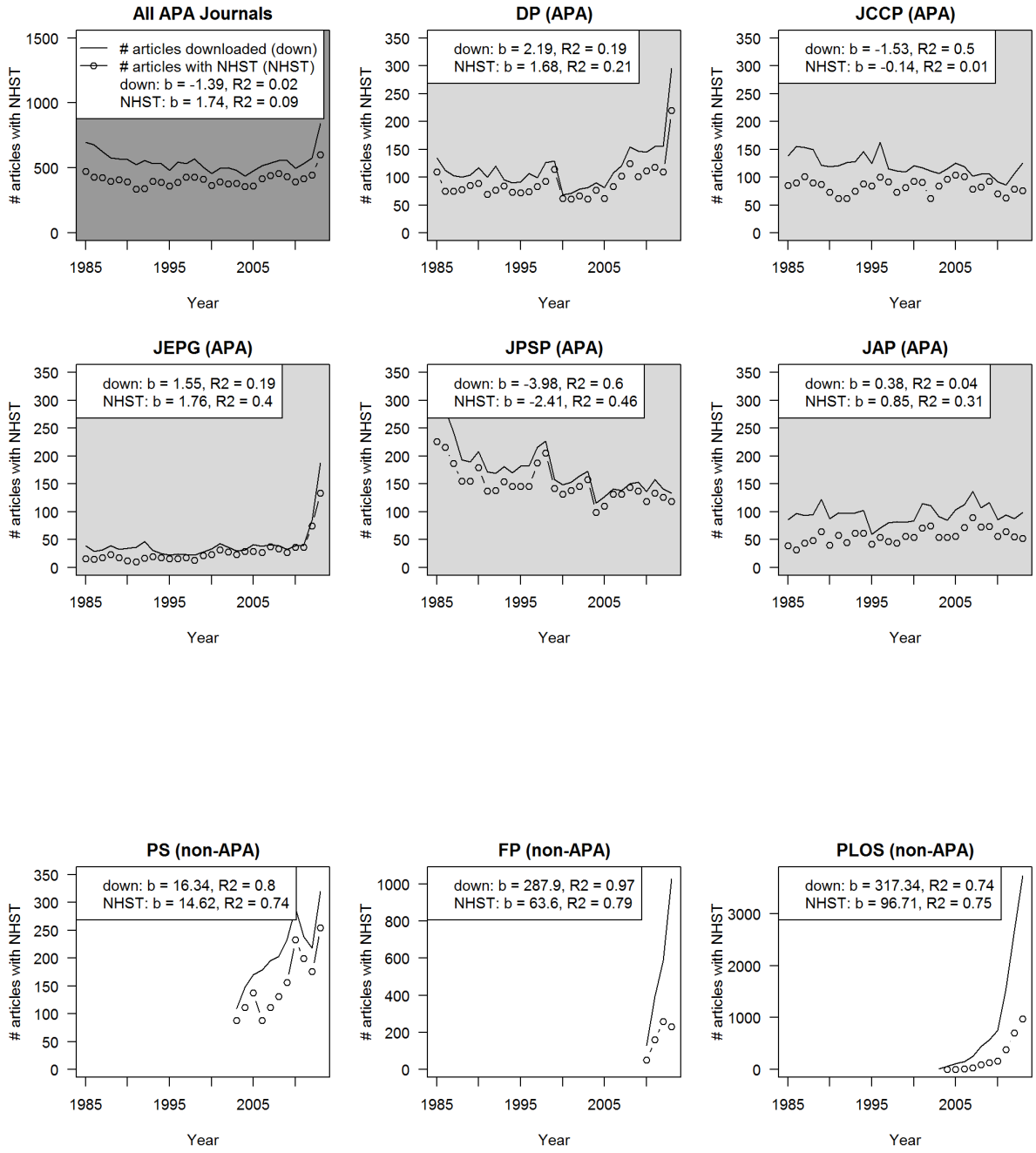


Figure 2.8

The total number of downloaded articles and the number of published articles that contain NHST results over the years, averaged over all APA journals (DP, JCCP, JEPG, JPSP, and JAP; dark gray panel), and split up per journal (light gray panels for the APA journals and white panels for the non-APA journals). Note that the y-axes in the plot for "All APA Journals", FP, and PLOS are different from the others and continue until 1,000, 1,050, and 3,750, respectively. The unstandardized regression coefficient 'b' and the coefficient of determination 'R^2' of the linear trend are shown per journal for both the downloaded articles (down) as articles with NHST results (NHST) over the years.

2.5.2 Number of exactly and inexactly reported *p*-values over the years

Besides the general prevalence of NHST results over the years, we were also interested in the prevalence of exactly reported *p*-values (*p* = …) and inexactly reported *p*-values (*p* </> …, or "ns", which could be interpreted the same as *p* > .05).[6] From the fourth edition of the APA Publication Manual onwards (1994), researchers have been encouraged to report *p*-values exactly, so we expected to find an increase of exactly reported *p*-values.

We inspected the prevalence of exact and inexact *p*-values over time averaged over all APA journals (DP, JCCP, JEPG, JPSP, and JAP; dark gray panel in Figure 2.9), and split up per journal (light gray panels for the APA journals and white panels for the non-APA journals in Figure 2.9). The average number of exact *p*-values per article with NHST results increases for all journals. For all journals except JAP and PS the number of inexact *p*-values per article with NHST results increased, although the increase is less steep than for exact *p*-values.

[6] Note that the APA advises any *p*-value smaller than .001 to be reported as *p* < .001. These cases could be considered as exactly reported. Our analysis does not take this into account. Furthermore, statements like "all tests >.05" are not included in our analysis.

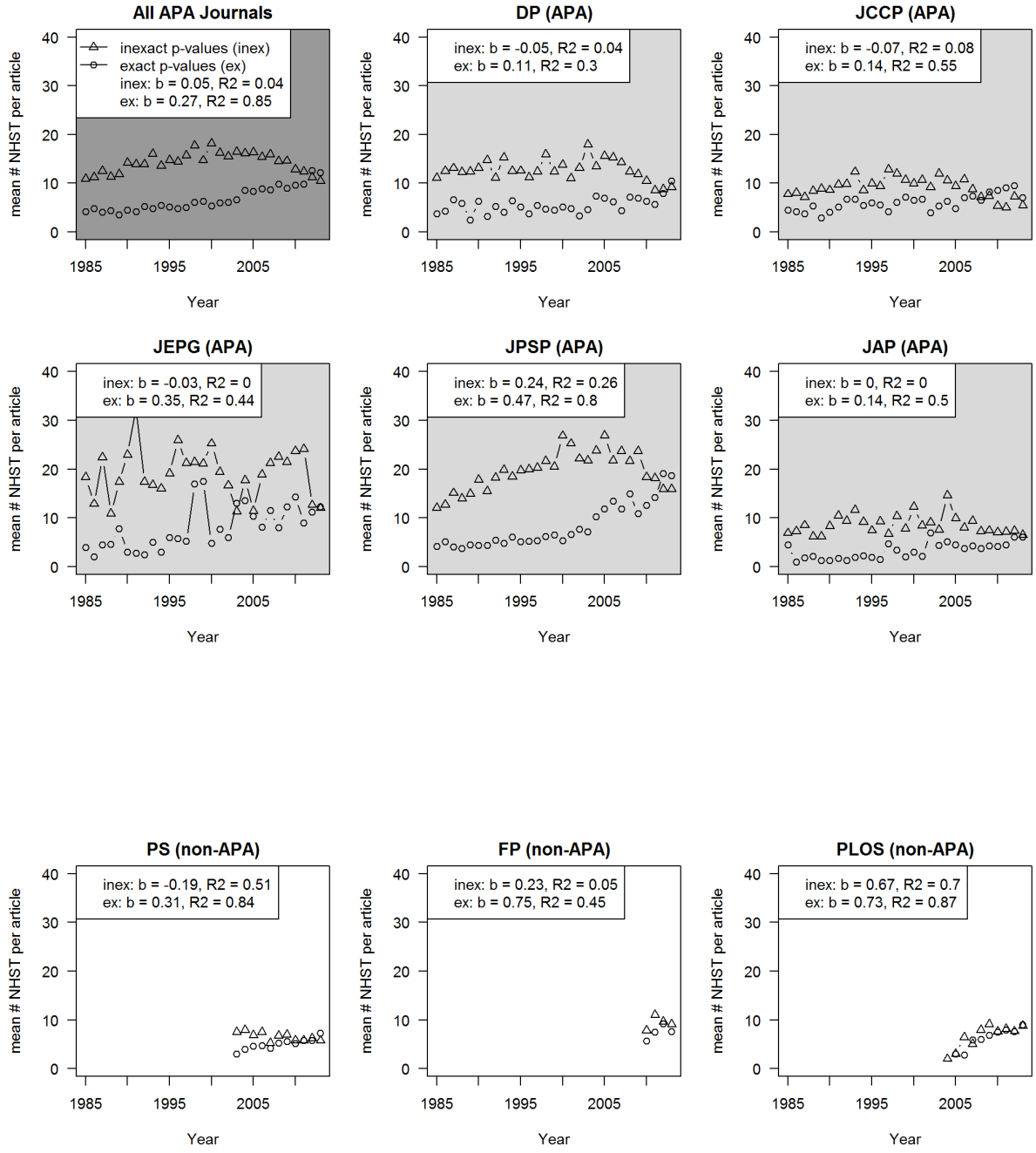


Figure 2.9

The average number of exact and inexact NHST results per article over the years, averaged over all journals (grey panel), and split up by journal (white panels). The unstandardized regression coefficient 'b' and the coefficient of determination 'R^2' of the linear trend are shown per journal for both exact (ex) as inexact (inex) p*-values over the years.*

Chapter 3

The Validity of the Tool "statcheck" in Discovering Statistical Reporting Inconsistencies

This chapter is submitted as Nuijten, M. B., Van Assen, M. A. L. M., Hartgerink, C. H. J., Epskamp, S., & Wicherts, J. M. (2017). The validity of the tool "statcheck" in discovering statistical reporting inconsistencies. *Preprint available from https://psyarxiv.com/tcxaj/*.

Abstract

The R package “statcheck” (Epskamp & Nuijten, 2016) is a tool to extract statistical results from articles and check whether the reported *p*-value matches the accompanying test statistic and degrees of freedom. A previous study showed high interrater reliabilities (between .76 and .89) between statcheck and manual coding of inconsistencies (.76 - .89; Nuijten, Hartgerink, Van Assen, Epskamp, & Wicherts, 2016; Chapter 2). Here we present an additional, detailed study of the validity of statcheck. In Study 1, we calculated its sensitivity and specificity. We found that statcheck’s sensitivity (true positive rate) and specificity (true negative rate) were high: between 85.3% and 100%, and between 96.0% and 100%, respectively, depending on the assumptions and settings. The overall accuracy of statcheck ranged from 96.2% to 99.9%. In Study 2, we investigated statcheck’s ability to deal with statistical corrections for multiple testing or violations of assumptions in articles. We found that the prevalence of corrections for multiple testing or violations of assumptions in psychology was higher than we initially estimated in Chapter 2. Although we found numerous reporting inconsistencies in results corrected for violations of the sphericity assumption, we demonstrate that inconsistencies associated with statistical corrections are not what is causing the high estimates of the prevalence of statistical reporting inconsistencies in psychology.

In psychological research, most conclusions are based on Null Hypothesis Significance Testing (NHST; see, e.g., Cumming et al., 2007; Sterling, 1959; Sterling et al., 1995). Unfortunately, there is increasing evidence that reported NHST results are often inconsistent. In psychology, roughly half of all articles published in reputable journals contain at least one inconsistent result in which the reported *p*-value does not correspond with the accompanying test statistic and degrees of freedom. In roughly one in eight articles there is at least one grossly inconsistent result in which the reported *p*-value is statistically significant (i.e., $p < .05$) but the recomputed *p*-value based on the test statistic and degrees of freedom is not, or vice versa (Chapter 2; Bakker & Wicherts, 2011; Bakker & Wicherts, 2014; Caperos & Pardo, 2013; Veldkamp et al., 2014). These inconsistencies can lead to erroneous substantive conclusions and affect the reliability of meta-analyses, so it is important that these inconsistencies can be easily spotted, corrected, and hopefully prevented.

To facilitate the process of detecting and correcting statistical reporting inconsistencies, we developed the R package "statcheck" (Epskamp & Nuijten, 2016; http://statcheck.io). Statcheck is a free and open-source algorithm that extracts NHST results reported in APA style from articles and recalculates *p*-values based on the reported test statistic and degrees of freedom. If the reported *p*-value does not match the computed *p*-value, the result is flagged as an inconsistency. If the reported *p*-value is significant ($\alpha = .05$) and the computed *p*-value is not, or vice versa, the result is flagged as a gross inconsistency.

To ensure that statcheck is a valid tool for detecting statistical inconsistencies, we included a detailed validity study in Chapter 2, in which we ran statcheck on a set of articles that had previously been manually coded for statistical reporting inconsistencies by Wicherts et al. (2011). Using the manually coded results as the standard we found that the interrater reliability of statcheck was .76 for flagging inconsistencies and .89 for gross inconsistencies. We reported in detail where any discrepancies came from and concluded that the validity of statcheck was sufficiently high to recommend its use for self-checks, peer review, and research on the prevalence of (gross) inconsistencies in large bodies of literature (for details, see Appendix A in Chapter 2).

Since the publication of the study in which statcheck was introduced and validated, statcheck has begun to be used in large-scale assessments (Baker, 2015, 2016b; Hartgerink, 2016) and in the peer-review process of the journals *Psychological Science* and the *Journal of Experimental Social Psychology*. Additionally, we have received additional questions about different aspects of statcheck's validity. First, we chose to express statcheck's validity in interrater reliability coefficients, but both in personal communications and in an anonymous review, researchers asked for more information about statcheck's false positive and false negative rate. Therefore, in Study 1 in this chapter we present an analysis of statcheck's accuracy by calculating its sensitivity (true positive rate) and specificity (true negative rate;

Altman & Bland, 1994). Second, Schmidt (2016) published a critique online in which he questioned the validity of statcheck in the presence of NHST results that are adjusted to correct for multiple testing or possible violations of assumptions. In Study 2 in this chapter we estimate the prevalence of such statistical corrections in psychology in the large sample of articles used in Chapter 2, and investigate whether the presence of such corrections is associated with statistical reporting inconsistencies and could have caused the high prevalence of these inconsistencies.

3.1 Study 1: Sensitivity & Specificity

In Chapter 2 we determined statcheck's validity by means of the interrater reliability between manual coding and statcheck's results. Another common way to determine an instrument's accuracy is to calculate its sensitivity and specificity (Altman & Bland, 1994). In calculating sensitivity and specificity we use the following terminology (Baratloo, Hosseini, Negida, & El Ashal, 2015):

True Positive (TP): the number of results correctly flagged as a (gross) inconsistency
False Positive (FP): the number of results incorrectly flagged as a (gross) inconsistency
True Negative (TN): the number of results correctly *not* flagged as a (gross) inconsistency
False Negative (FN): the number of results incorrectly *not* flagged as a (gross) inconsistency

Sensitivity refers to the "true positive rate": the proportion of "true" (gross) inconsistencies that were also flagged by statcheck as such:

$$sensitivity = \frac{TP}{TP+FN}.$$

Equation 3.1

Specificity refers to the "true negative rate": the proportion of results that are "truly" not (grossly) inconsistent, and statcheck correctly did not flag them as (gross) inconsistencies:

$$specificity = \frac{TN}{TN+FP}.$$

Equation 3.2

Together, sensitivity and specificity say something about statcheck's accuracy: the ability to correctly differentiate between consistent and (grossly) inconsistent results, or more mathematically:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}.$$

Equation 3.3

Ideally, accuracy should be 100%, which would mean there are no false positives or false negatives, but for an automated algorithm such as statcheck this is not tenable. Statcheck will probably be less accurate than a manual check, but we do want to minimize false positives and false negatives in flagging (gross) inconsistencies. To calculate statcheck's accuracy, sensitivity, and specificity, we used the same reference data set that we used to

calculate the interrater reliabilities in Chapter 2, namely data from the manual checks of Wicherts et al. (2011).

3.1.1 Reference Sample

As a reference set, we used the same sample as Wicherts et al. (2011), who manually coded the internal consistency of NHST results from 49 articles from the *Journal of Experimental Psychology: Learning, Memory, and Cognition (JEP:LMC)* and the *Journal of Personality and Social Psychology (JPSP)*. These authors included results from *t*, *F*, or χ^2-tests that were reported completely (test statistic, degrees of freedom, and *p*-value) in the Results section of an article. From this set, they only selected results with *p*-values smaller than .05. This resulted in a total set of 1,148 NHST results.

3.1.2 Procedure

We ran different versions of statcheck over the articles included in Wicherts et al. (2011). One article was excluded from the set, because it was retracted due to misconduct. Our final sample consisted of 48 articles and 1,120 NHST results. In all runs, we ran statcheck both with and without automated one-tailed test detection (OneTailedTxt = TRUE). With this option, statcheck considers a result consistent if (1) the reported *p*-value would be consistent if it belonged to a one-tailed test (specifically: if the reported *p* times two equals the computed *p*), and (2) if anywhere in the full text of the article statcheck found the word "one-tailed", "one-sided", or "directional". All other statcheck options were set to their default settings (see section 3.5 in the statcheck manual at http://rpubs.com/michelenuijten/statcheckmanual).

We compared the sensitivity and specificity of the three different versions of statcheck that are published on CRAN: statcheck 1.0.0 (Epskamp & Nuijten, 2014), statcheck 1.0.1 (Epskamp & Nuijten, 2015), and statcheck 1.2.2 (Epskamp & Nuijten, 2016). There were no major changes in the core code of statcheck 1.0.1 as compared to statcheck 1.0.0, but there were some relevant changes in version 1.2.2. In statcheck 1.2.2 there was a bug fix to ensure that statcheck does not misread *t*, *F*, or *r* statistics with a subscript as chi-square tests. We also adapted the code so that statcheck would still recognize a degree of freedom reported as the lower case letter L ("l") as the number "1". Furthermore, if statcheck detects a correlation that is reported as > 1, it neither calculates a *p*-value nor determines whether the result is inconsistent. The main reason is that when statcheck found a correlation larger than one the risk was too high that it had mistakenly identified a different test as a correlation, leading to a falsely flagged inconsistency. Finally, in version 1.2.2 we fixed a bug in the way statcheck flagged inconsistencies in inexactly reported test statistics (e.g., $t(38) < 1.00$, p = …). The full history and specific code of all changes to statcheck can be found on GitHub at https://github.com/michelenuijten/statcheck.

From the statistical results that statcheck detected, we selected results from *t*, *F*, or χ^2-tests that had a *p*-value smaller than .05 to match the inclusion criteria of Wicherts et al. (2011). Wicherts et al. only included results reported in the Results section, but statcheck cannot distinguish in which section of an article a result was reported and hence also extracted results from different sections. Conversely, Wicherts et al. included results from tables, but statcheck only extracts complete NHST results reported in APA style, which are usually not reported in tables.

Next, we compared the results from Wicherts et al. (2011) with the results from statcheck. We first checked what percentage of manually extracted results were also detected by statcheck. We then selected the NHST results that were extracted both by Wicherts et al. (2011) and statcheck, and continued to investigate if the manual classifications of inconsistencies and gross inconsistencies matched those of statcheck.

The reference data from Wicherts et al. (2011) and the full R scripts to clean and select the data and calculate sensitivity and specificity are available from https://osf.io/753qd/. The articles on which the data are based and on which statcheck was run are published on the private web page https://osf.io/ske8z/, and can be shared upon request.

3.1.3 Results

The accuracy, sensitivity, and specificity were almost exactly the same for the three versions of statcheck. The reason for this is that all updates to the code were made to solve specific problems that only occurred in a single instance the set of reference articles: one result was reported as "F(l, 76) = 23.95, p <.001", and only statcheck version 1.2.2 recognized the first degree of freedom as a 1. This means that the versions 1.0.0 and 1.0.1 of statcheck detected 684 NHST results of the 1,120 results (61.1%) that were included in Wicherts et al. (2011), and version 1.2.2 detected 685 NHST results (61.2%).[7] To calculate statcheck's sensitivity and specificity in detecting (gross) inconsistencies, we only focused on the NHST results that were detected both by Wicherts et al. (2011) and statcheck. The results of the sensitivity and specificity analysis for all three statcheck versions are displayed in Table 3.1.

[7] We excluded one article with 28 NHST results reported in APA style because it was retracted due to misconduct; this article was included by Wicherts et al. (2011). Note that in the validity study in Chapter 2 we reported that statcheck detected 775 NHST results, as that study also included results that were detected by statcheck but not included in the manual check (mainly results reported in a section other than the Results section).

Table 3.1

Results of the sensitivity and specificity analysis of statcheck 1.0.0 (Epskamp & Nuijten, 2014), statcheck 1.0.1 (Epskamp & Nuijten, 2015), and statcheck 1.2.2 (Epskamp & Nuijten, 2016), with and without one-tailed test detection. The statcheck version was indicated (e.g., "v. 1.0.0") if results differed between versions. The reference set consisted of manually coded data by Wicherts et al. (2011). TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative.

	statcheck (default)				**statcheck (with automated one-tailed test detection)**			
	TP	**FP**	**TN**	**FN**	**TP**	**FP**	**TN**	**FN**
Inconsistencies	34	26	624 v. 1.0.0 - 1.0.1 625 v. 1.2.2	0	29	19	631 v. 1.0.0 - 1.0.1 632 v. 1.2.2	5
Sensitivity				100%				85.3%
Specificity				96.0%				97.1%
Accuracy				96.2%				96.5%
Inconsistencies (strict)*	52	8	624 v. 1.0.0 - 1.0.1 625 v. 1.2.2	0	47	5	631 v. 1.0.0 - 1.0.1 632 v. 1.2.2	5
Sensitivity				100%				90.4%
Specificity				98.7%				99.8%
Accuracy				98.8%				99.1%
Gross Inconsistencies	8	6	670 v. 1.0.0 - 1.0.1 671 v. 1.2.2	0	7	0	676 v. 1.0.0 - 1.0.1 677 v. 1.2.2	1
Sensitivity				100%				87.5%
Specificity				99.1%				100%
Accuracy				99.1%				99.9%

* Here we consider the 7 results reported as "*p* = .000" and the 11 cases in which a Huynh-Feldt correction was applied, but the uncorrected degrees of freedom were reported, as true inconsistencies.

3.1.3.1 *True inconsistencies*

The diagnostic accuracy of statcheck depended on whether statcheck was run with or without its automated one-tailed test detection. In default mode (without one-tailed test detection) statcheck's sensitivity was 100%; all 34 true inconsistencies in the selected set of NHST result were correctly flagged by statcheck as such. In other words, in default mode there were no false negatives when flagging inconsistencies. When statcheck was run with one-

tailed test detection, however, sensitivity decreased to 85.3%. In this case, statcheck failed to flag 5 of the 34 true inconsistencies as such. In these cases, statcheck was too lenient in counting results as one-tailed tests: if the words “one-tailed”, “one-sided”, or “directional” were mentioned anywhere in the article, *all* results that would be consistent if they were one-tailed were counted as correct. To be able to decide for individual results whether or not it is one-tailed, we would need an algorithm that can interpret text substantively, which does not currently fit within the scope of the statcheck project.

The specificity of statcheck also depended on whether the one-tailed test detection was used, but here we see the opposite pattern: specificity was higher with one-tailed test detection. When statcheck was run with default options (without one-tailed test detection) its specificity was 96.0%: depending on the version, either 624 of the 650 truly consistent results (version 1.0.0 and 1.0.1) or 625 of the 651 truly consistent results (version 1.2.2) were correctly flagged as consistent by statcheck. This means that 26 results were “false positives”: results that statcheck flagged as an inconsistency, whereas they were counted as correct in the manual check. Eight of these false positives were one-tailed tests that statcheck did not recognize. Indeed, when we ran statcheck with one-tailed test detection, the specificity increased to 97.1%: now statcheck correctly flagged 631 of the 650 consistent results in version 1.0.0 and 1.0.1, and 632 of the 651 results in version 1.2.2. Note that one one-tailed result was still wrongly flagged as an inconsistency. In this case, the one-tailed *p*-value was probably based on an unrounded test statistic, whereas the (correctly) rounded test statistic was reported. In general, statcheck takes into account correct rounding of the test statistic, but when the one-tailed test detection is used, statcheck uses the exact reported test statistic to calculate whether a one-tailed *p*-value would be consistent. In future versions of statcheck, this feature will be adapted to take into account one-tailed *p*-values based on correctly rounded test statistics.

The sensitivity and specificity of statcheck can be combined to reflect its accuracy: the ability to correctly differentiate between consistent and (grossly) inconsistent results (see Equation 3.3). In default mode, without one-tailed test detection, the accuracy was 96.2%. If one-tailed test detection was switched on, statcheck’s overall performance slightly increased to an accuracy of 96.5%.

Eight of the 26 false positives could be explained by the use of one-tailed tests. The remaining 18 false positives were caused by two main things. First, statcheck counted results reported as $p = .000$ as inconsistent, because this is not in line with APA reporting practices. Cases like this should, according to the APA, be reported as $p < .001$. Wicherts et al. (2011) did not automatically count this as incorrect. There were seven such cases in this data set. Second, there were eleven results in which a Huynh-Feldt correction was applied to correct for a violation of the assumption of sphericity. A Huynh-Feldt correction adjusts the result by multiplying the degrees of freedom by a factor “ε”. However, in all eleven cases that we

detected, the unadjusted degrees of freedom were reported along with the adjusted *p*-value, which rendered the result internally inconsistent. In the manual check, this was still counted as consistent, because it was traceable how the correction had been applied. We consider and discuss such corrections in more detail below.

3.1.3.2 *True inconsistencies (strict)*

Our results showed that most of the "false positives" (i.e., results marked by statcheck as inconsistent, but counted as consistent in the manual check) resulted from conscious choices in programming statcheck. We deliberately chose to consider $p = .000$ as inconsistent, because a *p*-value can never be exactly zero. The APA prescribes that such results should be reported as $p < .001$ (American Psychological Association, 2010). The same line of reasoning applies to the Huynh-Feldt corrections. If these corrections were correctly reported (i.e., the adjusted degrees of freedom together with the adjusted *p*-value), they would have been consistent. If we retain both of these stricter criteria for flagging inconsistencies, statcheck's accuracy in detecting inconsistencies increases (see Table 3.1). With these stricter criteria, statcheck's sensitivity remains at 100% when it is run in default mode, or increases from 85.3% to 90.4% when one-tailed test detection is used. Similarly, the specificity increases from 96.0% to 98.7% (default) and from 97.1% to 99.8% (one-tailed test detection). These increases in sensitivity and specificity are also reflected in the overall accuracy, which increases from 96.2% to 98.8% (default) and from 96.5% to 99.1% (one-tailed test detection). Retaining these stricter criteria for flagging inconsistencies had no bearing on the sensitivity and specificity in detecting gross inconsistencies.

3.1.3.3 *True gross inconsistencies*

Similar to its performance when flagging inconsistencies, statcheck's sensitivity in detecting true gross inconsistencies depended on whether one-tailed test detection was used. In default mode, without one-tailed test detection, statcheck's sensitivity was 100%: all 8 true gross inconsistencies were correctly flagged as such. However, when one-tailed test detection was used, the sensitivity dropped to 87.5%: statcheck correctly identified 7 out of the 8 gross inconsistencies. The one missed gross inconsistency was due to the automatic one-tailed test detection being too lenient. The article that contained this specific gross inconsistency mentioned "directional" in the full text, which caused statcheck to count the result as a one-tailed test, but the manual check revealed that "directional" did not refer to the statistical analyses.

The specificity of statcheck in detecting results that were truly not grossly inconsistent also depended on the one-tailed test detection. In default mode, without one-tailed test detection, statcheck's specificity was 99.1%: 670 of the 676 results that were truly not gross inconsistencies were correctly identified as such. There were 6 results that statcheck wrongly

flagged as a gross inconsistency, because statcheck did not recognize that these were one-tailed tests. Indeed, when we ran statcheck with its one-tailed test detection the specificity increased to 100%. In other words, with one-tailed test detection, there were no false positives in detecting gross inconsistencies.

The sensitivity and specificity in detecting gross inconsistencies combined led to an accuracy of 99.1% if statcheck was run in default mode without one-tailed test detection, and increased to 99.9% when one-tailed test detection was used.

3.1.4 Conclusion

The analysis of statcheck's diagnostic accuracy showed low false positive and false negative rates in flagging inconsistencies and gross inconsistencies. The sensitivity (flagging true [gross] inconsistencies) ranged from 85.3% to 100%, and the specificity (flagging results that are truly not [grossly] inconsistent) ranged from 96.0% to 100%. Combined, these results indicate that the accuracy ranged from 96.2% to 99.9%. We considered these results evidence that the validity of statcheck is high.

The exact sensitivity and specificity depended on several conditions. First, we found that statcheck's sensitivity and specificity depended on whether its automated one-tailed test detection was used. The sensitivity of statcheck was highest without one-tailed test detection, whereas the specificity was highest when statcheck was run *with* one-tailed test detection. Users can take this into account when they decide whether or not to use the one-tailed test detection; if they find it most important to avoid false positives and not falsely flag correct results as inconsistent, they should use one-tailed test detection. Conversely, when they find it most important to avoid false negatives and they want to flag every result that could potentially be inconsistent, they should use statcheck without one-tailed test detection. This is a standard trade-off with any diagnostic instrument.

3.1.5 Generalizability Sensitivity & Specificity

A clear limitation of this additional validity study is that we used a single manually coded sample as a reference set. However, we do not believe that statcheck's diagnostic accuracy varies considerably across articles, journals, or disciplines. This belief is strengthened by the fact that we keep finding very similar inconsistency rates across different samples (see, e.g., the different estimates in the three studies in Chapter 4, or the summary of different studies about the prevalence of inconsistencies in Table 2.2 in Chapter 2).

We also see no reason to expect large differences in sensitivity and specificity between the different versions of statcheck. The adaptations to the code across different versions were mainly aimed at improving statcheck's detection rate of NHST results in peculiar and rather infrequent cases. Hence, statcheck's sensitivity and specificity will remain the same or likely only slightly increase over versions. As an extra check, we ran statcheck 1.0.0, 1.0.1, and 1.2.2

on a different sample of articles to see if there were any changes in the detected prevalence of (gross) inconsistencies. For this check we used a set of 137 psychology meta-analyses that we had collected for a different study (see Chapter 9). We found that even though the detection rate of statcheck increased in version 1.2.2 (226 NHST results as opposed to 215 results in the previous two versions), the numbers of detected inconsistencies (25) and gross inconsistencies (4) were the same. Since the total number of detected results increased, the percentage of results that were inconsistent or grossly inconsistent decreased slightly (from 11.6% to 11.1%, and from 1.9% to 1.8%, respectively). In short, even though the sensitivity and specificity were calculated based on only one reference set of articles, we argue that these results can be generalized to different versions of statcheck and different sets of articles.

3.2 Study 2: Accounting for Corrections for Multiple Testing, Post Hoc Testing, or Possible Violations of Assumptions

A possible cause for a detected inconsistent result is the use of statistical corrections for multiple testing, post hoc testing, or violations of assumptions. Take for example the Bonferroni correction for multiple testing. This correction is used to control the Type I error rate by dividing the level of significance (α) by the number of hypotheses tested. However, we often see articles in which instead of dividing α, researchers multiply the *p*-values by the number of tests. This then results in an internally inconsistent statistical result: the original test statistic and degrees of freedom no longer correspond to the reported (multiplied) *p*-value. Such cases will be flagged by statcheck as an inconsistency.

In Chapter 2 we intended to give a rough estimate of the prevalence of corrected *p*-values to illustrate that these were an unlikely cause of the many inconsistent *p*-values we found. We used the "Search" function in Windows Explorer to search the entire folder of downloaded articles for "Bonferroni" and "Huynh-Feldt", and reported the following results: *"[...] when we automatically searched our sample of 30,717 articles, we found that only 96 articles reported the string "Bonferroni" (0.3 %) and nine articles reported the string "Huynh-Feldt" or "Huynh Feldt" (0.03 %). We conclude from this that corrections for multiple testing are rarely used and will not significantly distort conclusions in our study."* (Chapter 2; or see Nuijten et al., 2016, p. 1207)

On the post-publication peer review forum “PubPeer” and the e-print service “arXiv”, Schmidt (2016) expressed his concern that we underestimated the prevalence of corrections for two reasons. First, we only searched for "Bonferroni" and "Huynh-Feldt", but did not include several other types of corrections. Second, estimates based on Schmidt’s own library and our validation sample (Wicherts et al., 2011) resulted in a higher prevalence of corrections than the one we had found in our full text search. Therefore, we decided to re-estimate the prevalence of corrected p-values in our sample of the literature. We also examined whether corrections were associated with inconsistently reporting statistical results.

3.2.1 Re-estimating Prevalence of Correction-Related Strings

We re-estimated the prevalence of articles that might contain statistical results that were adjusted by one of the corrections mentioned by Schmidt: Bonferroni, Scheffé, Tukey, Greenhouse-Geisser, and Huynh-Feldt. To this end, we ran a shell script with full text searches on the full database of 30,717 articles used in Chapter 2. The full script is available at https://osf.io/v9msf/. The script counts all articles that contain the string "Bonferroni", "Tukey", "Scheff", "Greenhouse", or "Huynh". Using this method, the results showed a much higher prevalence of strings of text that could point to corrected statistical results than we had found in our original estimate, confirming Schmidt's (2016) suspicion that we had initially missed many of these corrected results (see Table 3.2). We speculate that something went wrong in the Windows Explorer Search function in Chapter 2, but we were not able to determine this with certainty. However, more important than finding the cause of this discrepancy is investigating whether these results of statistical corrections are associated with reporting inconsistencies and hence might influence our original conclusion that statistical corrections are an unlikely cause for the high prevalence of statistical reporting inconsistencies in the psychological literature.

Table 3.2

New estimates of the number and percentage of articles that mentioned any of the listed types of corrections compared to the estimates we mentioned in Chapter 2, based on the total number of 30,717 downloaded articles.

Correction Type	# Articles found with shell script	% Articles estimated with shell script	% Originally estimated with Windows Explorer
Bonferroni	2,744	8.93	0.30
Tukey	1,691	5.51	NA*
Scheffé	667	2.71	NA
Greenhouse-Geisser	898	2.92	NA
Huynh-Feldt	234	0.76	0.03
Articles with one or multiple corrections	5,513	17.9	NA

* NA = Not Available, i.e., these were not examined in Chapter 2.

3.2.2 Percentage of Inconsistencies in Articles without Adjusted Statistics

One way to determine whether the presence of statistical corrections and adjustments has influenced our estimate of the prevalence of (gross) inconsistencies in the psychological literature is to remove all articles from the analysis that show any sign of containing such a correction or adjustment. If a large part of the detected inconsistencies in Chapter 2 were due to statistical adjustments, one would expect that a subset of articles without any corrections

would show a lower prevalence of inconsistencies and gross inconsistencies (all else being equal).

For this analysis, we first needed to identify which of the 16,695 articles that we analyzed in Chapter 2 contained evidence for adjusted statistics. We used the same shell script as above to determine which articles mentioned any of the keywords "Bonferroni", "Tukey", "Scheff", “Greenhouse”, or “Huynh” (the full script can be found at https://osf.io/v9msf/). This resulted in a list of 6,234 article titles, of which 5,513 were unique (some of the articles contained multiple keywords and appeared in the list two or more times). We found that 2,396 of the 16,695 (14.4%) articles in which statcheck detected APA reported NHST results contained at least one of the keywords that possibly indicated the presence of statistical corrections.[8] To extract the NHST results and detect inconsistencies, statcheck version 1.0.1 was used with automated one-tailed test detection (Chapter 2).

We removed all articles with any evidence for the presence of statistical adjustments from the sample and re-estimated the general prevalence of inconsistencies and gross inconsistencies (see Table 3.3). The results showed that removing articles with possible corrections led to a slightly *higher* prevalence of inconsistencies and gross inconsistencies than was found in Chapter 2. This suggests that the original estimates of the high prevalence of inconsistencies in the psychological literature are not driven by the presence of tests corrected for multiple testing, post hoc testing, or violations of assumptions. A possible explanation for this unexpected increase in the detection rate of inconsistencies here is that researchers who apply statistical corrections might be more diligent when it comes to their statistics, which might also decrease the probability that they report one or more results inconsistently.

The full R code of this analysis can be found at https://osf.io/t7b6m/. The raw data from Chapter 2 with article identifiers, from which we selected a sample of articles to manually code, is published on a private page at OSF (https://osf.io/sa87e/); due to ethical restrictions these data are only available upon request. The articles that were scanned in Chapter 2 are also published on a private page (https://github.com/MicheleNuijten/sampleStatcheck) and are available upon request.

[8] For 20 articles in the original sample we were unable to automatically check if these titles also occurred in the list of articles with possible corrections, because problems in automatically reading in the file names due to special symbols in the article title. Because we were not sure if these articles contained evidence for statistical corrections, we ran our analyses with and without these articles. Removing these articles on top of the articles that did contain evidence for corrections did not change the estimates of inconsistency prevalence.

Table 3.3

Estimates of the prevalence of (gross) inconsistencies in the full sample from Chapter 2 compared to the sample without articles that showed evidence for containing one or more statistical corrections or adjustments.

	All articles (data from Chapter 2)	Articles without evidence for corrections
% articles with at least one inconsistency	49.6%	49.8%
% articles with at least one gross inconsistency	12.9%	14.4%
average % of p-values that are inconsistent per article	10.6%	11.1%
average % of p-values that are grossly inconsistent per article	1.6%	1.9%

3.2.3 Percentage of Inconsistencies that are Associated with a Statistical Correction

Schmidt (2016) argued that it is misleading if statistics are flagged as inconsistencies if they were affected by statistical corrections, because the use of statistical corrections is usually good practice and statcheck would "punish" that practice by flagging its results as inconsistencies. However, we disagree that flagging inconsistently reported corrected statistics as such is misleading. Each of the five examples of statistical corrections that Schmidt mentioned can (and should be reported) in an internally consistent way.

First, a Bonferroni correction for multiple testing consists of dividing the level of significance, α, by the number of tests to adjust the level of significance. For instance, if you run 6 different tests, and you want to retain an overall α of .05, the Bonferroni corrected α for each of the tests is $\alpha = .05/6 = .00833$. However, instead of correcting α, researchers often adjust the *p*-values themselves by multiplying each *p*-value by the number of tests. In fact, this is also how SPSS carries out the Bonferroni post hoc test with the POSTHOC Bonferroni command in UNIANOVA. However, if one reports the original test statistic and degrees of freedom, but a multiplied *p*-value, the result is no longer consistent. An additional reason not to multiply the *p*-value is that with this procedure it is possible to obtain *p*-values larger than one, which are meaningless by definition.

Furthermore, Schmidt (2016) mentions the Greenhouse-Geisser correction and Huynh-Feldt correction to adjust for violations of the assumption of sphericity. In both procedures, the degrees of freedom of an *F*-test are multiplied by a factor ε that lies between 0 and 1, which increases the *p*-value of the observed *F*-statistic. Sometimes, as Schmidt also illustrates, researchers report the original, uncorrected degrees of freedom with the corrected test statistic and *p*-value. In these cases, the original degrees of freedom may be reported so that others may deduce the sample size on which the test was based, but making the statistical

results inconsistent. It is recommended to report the corrected statistical result, as well as the value of ε (see, e.g., Field, 2009, p. 481).

We initially did not expect problems with statistical results of the Tukey and Scheffé post hoc tests that Schmidt mentioned. The Tukey test has its own statistic and *p*-value, and as such is not a correction of another statistical result. The Scheffé test compares the original *F*-statistic with a recalculated critical value of the *F*-test (i.e., $(K\text{-}1)\times F_{CV(K\text{-}1,N\text{-}K)}$, with *K* and *N*-*K* denoting the degrees of freedom of the *F*-test, and F_{CV} denoting the critical value of the test). Since the Scheffé test does not yield an exact *p*-value but a comparison with a significance level, e.g. $p < .05$, statcheck will not detect these results because they are not reported in line with APA guidelines. However, as we did not know how researchers report these results, we also examined reported results of these statistical tests.

In short, we contend that all of the five corrections Schmidt mentioned (or any other, to our knowledge) can be reported in an internally consistent and informative way. However, we agree with Schmidt that some of these corrections might be reported incorrectly in research articles. To see how statistical corrections are usually reported and how often a statistical correction led to a flagged inconsistency, we manually coded a subsample of the articles investigated in Chapter 2.

3.2.3.1 *Method*

We selected all articles in which we found a keyword that could indicate the use of statistical corrections (see the shell script at https://osf.io/v9msf/). For each of the five corrections (Bonferroni, Scheffé, Tukey, Greenhouse-Geisser, and Huynh-Feldt), we randomly selected 100 articles that contained the specific keyword and had at least one NHST result that statcheck was able to verify. There were only 39 articles that contained both the keyword "Huynh" and had statcheck output, so we included all of those. This procedure led to a sample of 439 articles. For the current analysis, we were only interested in cases where a statistical correction could have led to an inconsistency, so from the 439 articles we then only selected those articles in which statcheck flagged at least one inconsistency. This resulted in a final sample of 229 articles (see Table 3.4).

Table 3.4

The number of randomly selected articles that contained both a keyword indicating a statistical correction and statcheck output, and the number of these articles that also contained a flagged inconsistency.

Correction type	# Selected articles selected with statcheck output	# Selected articles with statcheck output that contained at least one inconsistency
Bonferroni	100	50
Tukey	100	51
Scheffé	100	50
Greenhouse-Geisser	100	65
Huynh-Feldt	39	25
Total	439	229

We then manually coded the inconsistencies in the selected articles, using three coders (MN, MvA, and a research assistant). As we could not make a specific protocol beforehand that anticipated all possible errors and contingencies, we decided to code results by discussion. That is, while coding different results independently in the same office, we discussed those instances where the coder was unsure about how to code the result. Each type of correction required a slightly different approach in coding, but we retained the following general approach for each article:

1. Use the Search function to find sentences that mentioned the correction that was the selection criterion for the article, to determine whether any results might be associated with this correction.
   Example: "Note that all ANOVAs reported in this article use the Greenhouse-Geisser correction for violations of sphericity."
2. Use the Search function to find all results that statcheck flagged as an inconsistency, to determine whether these results are associated with the correction that was the selection criterion for the article.
   Example: "Greenhouse–Geisser F(1.77, 1098.32) = 2.34, p = 0.06."
3. If, based on the text, no inconsistency is associated with the correction, classify this as 0.
   Example: "F(3, 357) = 5.44, p<.001 (all ps still reliable when the Geisser–Greenhouse adjustment was applied)."
4. If, based on the text, an inconsistency is associated with the correction, and the cause of the inconsistency seems to be the use of the correction, classify this as 1.
   Example: "A main effect of rank [F(5, 155) = 3.57, p = 0.006, epsilon = 0.89] was observed"
   Additional signs that a result is associated with any of the corrections are:

   a. The reported *p*-value is higher than the *p*-value computed by statcheck
   b. Scheffé, Greenhouse-Geisser, and Huynh-Feldt corrections only apply to *F*-tests
   c. Tukey tests/corrections only apply to *t*-tests or *F*-tests where $df_1 = 1$
5. If, based on the text, an inconsistency is associated with the correction, but the result is still inconsistent when this correction is taken into account, classify this as 2.
   Example: "F(6,114) = 2.67, p = 0.057, ε = 0.45." If we multiply the degrees of freedom (6 and 114) by epsilon (.45), the result would be F(2.7, 51.3) = 2.67, and the accompanying p-value should be .063, which does not correspond to the reported p-value of .057.
   Example: "Greenhouse–Geisser F(1.77, 1098.32) = 2.34, p = 0.06." This result is reported with the corrected degrees of freedom and should be internally consistent. However, based on the reported degrees of freedom and test statistic, the p-value should be .103.

Note that following this approach, we only looked at the type of correction that was the selection criterion for the article. If an article belonged in the category "Greenhouse-Geisser" but a result was affected by a Bonferroni correction, this was coded as a 0. Furthermore, it was sometimes hard to distinguish between category 0 and 2; there were cases in which it was unclear whether a result was not associated with a correction, or whether the result was associated with a correction but simply wrongly reported. For instance, there were cases in which the text stated that "all *p*-values were corrected for multiple testing with a Bonferroni procedure", but then some reported *p*-values were *lower* than the recomputed ones, which is impossible if the original *p*-value was multiplied to correct for multiple testing. In this case, it is unclear whether this result was in fact not Bonferroni corrected at all (category 0), or whether the impossibly low *p*-value was the result of a typo (category 2). Therefore, in our analyses, we only focused on category 1: results in which the inconsistency was clearly associated with the statistical correction. The coded data and R scripts to analyze them are available at https://osf.io/gf9zx/.

3.2.3.2 ***Results***

An overview of the results is shown in Table 3.5. In total, the 229 selected articles contained 5,606 APA reported NHST results that were extracted by statcheck, of which 798 results were inconsistently reported (14.2%). From these 798 inconsistencies, 97 (12.2%) were associated with one of the five investigated statistical corrections.

If we zoom in on the specific types of corrections, it turns out that Tukey or Scheffé corrections never led to a reporting inconsistency being flagged by statcheck. As stated earlier, Scheffé's method does not yield an exact *p*-value, which means that any test result based on Scheffé's method will not be reported in the APA style that statcheck can detect. Typically,

results of Scheffé tests are reported along these lines: “Scheffé multiple-comparisons tests revealed significant differences (*p*s < .05)”. It is therefore not surprising that none of the inconsistencies flagged by statcheck were associated with Scheffé’s test. Similarly, we also did not expect to find many cases in which Tukey’s test seemed to have affected the consistency of a result. Tukey’s test has its own test statistic (*q*), which statcheck cannot detect. Furthermore, it turned out that the results of Tukey’s test were often reported in-text, e.g., “Tukey's HSD test was used to specify the nature of the differences between conditions (*p* < .05, for all differences reported)”, or in tables in which the significance of the results was indicated by stars. In neither of these cases could statcheck have detected the results, let alone flag an inconsistency.

In the articles that contained the keyword “Bonferroni”, 17 of the 184 inconsistencies (9.2%) were caused by researchers multiplying the *p*-values instead of dividing α. The percentage of inconsistencies associated with a correction was higher in articles that showed evidence for a correction for the violation of the assumption of sphericity; in all articles that mentioned Huynh-Feldt, 14 of the 73 inconsistencies were caused by reporting the uncorrected degrees of freedom (19.2%), and in the articles that mentioned Greenhouse-Geisser, 66 of 198 inconsistencies (33.3%) were caused by reporting the uncorrected degrees of freedom.

Table 3.5

The total number of APA reported NHST results extracted by statcheck, the total number of those NHST results that were inconsistently reported, and the number of those inconsistencies that were caused by the use of a statistical correction. The results are split up per type of correction.

Correction type	**# APA reported NHST results in selected articles**	**# inconsistent results**	**# inconsistencies associated with correction**
Bonferroni	1,108	184	17 (9.2%)
Tukey	1,185	208	0 (0.0%)
Scheffé	898	135	0 (0.0%)
Greenhouse-Geisser	1,646	198	66 (33.3%)
Huynh-Feldt	769	73	14 (19.2%)
Total	5,606	798	97 (12.2%)

The results of this analysis of articles using statistical corrections showed that the vast majority of inconsistencies were not associated with these corrections. Test results based on Tukey’s test or Scheffé’s test were never reported in such a way that statcheck could detect them, which meant that these corrections never led to a reporting inconsistency being flagged by statcheck. When a Bonferroni correction was used, less than one in ten inconsistencies was

actually caused by a multiplied *p*-value. Corrections for violations of sphericity, in contrast, led to more inconsistencies. Here, the uncorrected degrees of freedom were often reported alongside the corrected *p*-value (e.g., "Hereafter, when violations of sphericity occurred, we report Huynh-Feldt corrected *p*-values; for clarity, unadjusted degrees of freedom are reported.").

3.3 General Discussion

In this chapter we investigated statcheck's diagnostic accuracy by calculating its sensitivity and specificity, and examined whether statistical corrections could have caused the high prevalence of statistical reporting inconsistencies as found in Chapter 2. The results of Study 1 showed that all current versions of statcheck have high sensitivity and specificity. The majority of "false positives" in flagged inconsistencies were caused by the deliberate choice to always count "$p = .000$" as incorrect (not applied in the manual checking by Wicherts et al., 2011), and by results that had been subject to a statistical correction and therefore inconsistently reported.

The fact that statistical corrections can lead to inconsistently reported results has been presented as an argument against the use of statcheck (Schmidt, 2016).[9] However, we argue that there is no reason to report the result of a corrected test in a manner that creates an inconsistency between the test statistic, degrees of freedom, and the *p*-value. In Study 2, we found no reporting inconsistencies associated with Scheffé and Tukey tests, and only some inconsistencies associated with the Bonferroni correction (9.2% of inconsistencies). We did find numerous inconsistencies associated with corrections for violations of the sphericity assumption (Greenhouse-Geisser and Huynh-Feldt; 33.3% and 19.2% of the inconsistencies, respectively). We therefore conclude that Schmidt (2016) was correct to raise the issue that some statistical corrections may be detected as reporting inconsistencies, as some of these corrections may not be consistently reported. As the APA manual does not discuss reporting of statistical corrections, we sent a message via APA's feedback form recommending that a future edition of the APA Publication Manual should incorporate specific examples of how to report these corrections in articles (e.g., "Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(5) = 11.41$, $p = .044$), therefore degrees of freedom were corrected using Huynh-Feldt estimates of sphericity ($\varepsilon = 0.67$). The results show that X was

[9] Schmidt's (2016) critique even led to an official statement from the DGPs (the German Psychological Society) in which they argue against the use of statcheck. In our reply we maintained our position that even though statcheck is not 100% accurate, its validity is high enough to recommend its use. Their letter, our reply, and a summary of the discussion can be found on the following Retraction Watch post: http://retractionwatch.com/2016/10/25/psychological-society-wants-end-to-posting-error-finding-algorithm-results-publicly/.

significantly affected by Y, $F(2, 13.98) = 3.79$, $p = .048$, $\omega^2 = .24$."; adapted from Field, 2009; p. 482).

Even the reporting inconsistencies associated with these tests and corrections could not explain the high prevalence of reporting inconsistencies in psychology as reported in Chapter 2, for two reasons. First, we found that these corrections are infrequently used. Second, the subset of articles that showed no evidence for any of these corrections had a higher prevalence of (gross) inconsistencies than the full set of articles. Furthermore, the estimates of the prevalence of (gross) inconsistencies in Chapter 2 are very similar to the estimates reported in other studies in which manual procedures were used (see, e.g., Bakker & Wicherts, 2011; Bakker & Wicherts, 2014; Caperos & Pardo, 2013).

The results of this validity study, the results of the earlier validity study in Chapter 2, and the convergence of estimates from the present study with studies that were based on manual checking (e.g., Bakker & Wicherts, 2011) highlight statcheck's high level of diagnostic accuracy. Therefore, we recommend the use of statcheck for checking one's own work, for use in peer review, and as a tool to estimate the general prevalence of reporting inconsistencies across a large sample of articles. We stress that statcheck is an algorithm that will, like any automated procedure exposed to real-world data, sometimes lead to false positives or false negatives. These limitations should be taken into account, preferably by manually double-checking inconsistencies detected by statcheck.

Chapter 4

Journal Data Sharing Policies and Statistical Reporting Inconsistencies in Psychology

This chapter is published as Nuijten, M. B., Borghuis, J., Veldkamp, C. L. S., Dominguez-Alvarez, L., Van Assen, M. A. L. M., & Wicherts, J. M. (2017). Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra: Psychology, 3(1)*, 1-22.

Abstract

In this chapter, we present three retrospective observational studies that investigate the relation between data sharing and statistical reporting inconsistencies. Previous research found that reluctance to share data was related to a higher prevalence of statistical errors, often in the direction of statistical significance (Wicherts et al., 2011). We therefore hypothesized that journal policies about data sharing and data sharing itself would reduce these inconsistencies. In Study 1, we compared the prevalence of reporting inconsistencies in two similar journals on decision making with different data sharing policies. In Study 2, we compared reporting inconsistencies in psychology articles published in PLOS journals (with a data sharing policy) and Frontiers in Psychology (without a stipulated data sharing policy). In Study 3, we looked at papers published in the journal Psychological Science to check whether papers with or without an Open Practice Badge differed in the prevalence of reporting errors. Overall, we found no relationship between data sharing and reporting inconsistencies. We did find that journal policies on data sharing are extremely effective in promoting data sharing. We argue that open data is essential in improving the quality of psychological science, and we discuss ways to detect and reduce reporting inconsistencies in the literature.

Most psychological researchers use Null Hypothesis Significance Testing (NHST) to evaluate their hypotheses (Cumming et al., 2007; Hubbard & Ryan, 2000; Sterling, 1959; Sterling et al., 1995). The results of NHST underlie substantive conclusions and serve as the input in meta-analyses, which makes it important that they are reported correctly. However, NHST results are often misreported. Several large-scale studies estimated that roughly half of psychology articles using NHST contain at least one *p*-value that is inconsistent with the reported test statistic and degrees of freedom, while around one in eight such articles contain a gross inconsistency, in which the reported *p*-value was significant and the computed *p*-value was not, or vice versa (Chapter 2; Bakker & Wicherts, 2011; Caperos & Pardo, 2013; Veldkamp et al., 2014). In the medical sciences roughly one in three articles contains an inconsistent *p*-value (Garcia-Berthou & Alcaraz, 2004), and in psychiatry about one in ten articles (Berle & Starcevic, 2007).

There is evidence that inconsistent *p*-values are associated with reluctance to share data, especially when the inconsistencies concern statistical significance (Wicherts et al., 2011). Wicherts et al. speculated that it is possible that authors are reluctant to share data because they fear that other research teams will arrive at different conclusions, or that errors in their work will be exposed (see also Ceci, 1988; Hedrick, 1985; Sterling & Weinkam, 1990). Along these lines, one may expect that if authors intend to make their data available from the start, they will double-check their results before writing them up, which would result in fewer inconsistencies in the final paper. Wicherts et al. also offered the alternative explanation that the relation between data sharing and misreporting is caused by differences in the rigor with which data are managed; researchers who work more diligently in their handling and archiving of data are probably less likely to commit a reporting error.

In psychology, the availability of research data in general is already strikingly low (Vanpaemel, Vermorgen, Deriemaecker, & Storms, 2015; Wicherts, Borsboom, Kats, & Molenaar, 2006), although this problem is not limited to psychology (see, e.g., Alsheikh-Ali et al., 2011). This is a worrying finding in itself, since the availability of original research data is essential to reproduce or verify analyses. However, this problem becomes worse if data are even less likely to be shared if the research article contained statistical inconsistencies, because in these cases verification of the analyses is even more important. Over the past few years there has been increasing awareness that the availability of research data is essential for scientific progress (Anagnostou et al., 2015; Nosek et al., 2015; Wicherts, 2011; Wilkinson et al., 2016), and several journals have started to request authors to share their data when they submit an article (e.g., in PLOS and Psychological Science; see Bloom, Ganley, & Winker, 2014; Lindsay, 2017; respectively). We theorized that such journal policies on data sharing could help decrease the prevalence of statistical reporting inconsistencies, and that articles with open data (regardless of journal policy) contained fewer inconsistencies.

In this chapter, we present three retrospective observational studies that investigate the relation between data sharing and reporting inconsistencies. Our two main hypotheses were that 1) journals that encourage data sharing will show a (larger) decrease in inconsistencies and gross inconsistencies compared to similar journals that do not encourage data sharing (an open policy effect), and 2) articles that are accompanied with open data have fewer inconsistencies and fewer gross inconsistencies than articles without open data (an open data effect). We compared inconsistency rates between two similar journals on decision making with different data sharing policies (Study 1), between psychology articles from journals from the open access publisher PLOS that requires open data and Frontiers that has less strict data sharing policies (Study 2), and between papers in the journal Psychological Science with and without Open Practice Badges (Study 3). Studies 2 and 3 are pre-registered and the relevant registrations can be found at https://osf.io/538bc/. Exploratory findings across the three studies are reported in a final results section.

4.1 Study 1

In Study 1 we documented the prevalence of reporting inconsistencies in two similar journals on decision making that have different data sharing policies: the Journal of Behavioral Decision Making (JBDM; no data sharing policy) and Judgment and Decision Making (JDM; recommended data sharing). Furthermore, we compared the number of reporting inconsistencies in articles that actually did or did not include shared data, regardless of the journal they were published in. We hypothesized that JDM would show a (larger) decrease in inconsistencies and gross inconsistencies compared to JBDM after the introduction of the data sharing policy in JDM (open policy effect), and that articles that are accompanied with open data contain fewer inconsistencies and gross inconsistencies than articles that are not accompanied with open data (open data effect).

4.1.1 Method

4.1.1.1 *Sample*

We examined the relation between open data journal policy on statistical reporting inconsistencies in two similar psychological journals: JBDM (ISI impact factor in 2015: 2.768) and JDM (ISI impact factor in 2015: 1.856). Both journals focus on human decision processes and accept empirical research as well as theoretical papers. Furthermore, there is considerable overlap between their editorial boards: in 2015, seventeen researchers sat in the editorial boards of both JDM (51 members in total) and JBDM (125 members in total). A difference between the journals is that JDM is completely open access, whereas in JBDM the authors can pay a fee to make their article open access. The main difference of concern here, however, is that since 2011 JDM editors have started encouraging authors to submit their raw

data at the time of review (Baron, 2011).[10] When the articles are accepted, these data are subsequently published on the web site along with the articles. Before 2011, there was no explicit data policy in JDM. JBDM did not adopt a similar data sharing policy in the relevant years.[11]

We downloaded the articles of JDM in the periods before and after their policy change, and we included articles from JBDM in the corresponding time periods. The first issue in JDM was published in 2006, and from April 2011 (Issue 3, 2011; corresponding to Issue 2 2011 of JBDM) onwards JDM started to implement the new data policy. We collected data in 2015, so we included papers up until the end of 2014 to include the most recent full year. Our final sample contained papers published in the years 2006 to February 2011 (T1), and in April 2011 to 2014 (T2). See Table 4.1 for the number of articles collected per journal and time period. We included all research articles and special issue papers from these periods in both journals, but no book reviews and editorials. All articles of JDM were HTML files, whereas all articles of JBDM were PDF files because no HTML files were available in T1.

Table 4.1

Number of articles (N) downloaded per journal and time period: 2006 to February 2011 (T1; published before open data policy of JDM), and from April 2011 to 2014 (T2; published after open data policy of JDM).

	N in T1	N in T2	Total N
JBDM	157	149	306
JDM	236	222	458
Total	393	371	764

[10] The data sharing recommendation of JDM states: *"We encourage the submission of raw data at the time of review, and we include the data of accepted articles with the articles (unless this is for some reason difficult). We will also include stimuli, questionnaires, and code, when these are necessary to understand exactly what was done (again, unless this is difficult for some reason)."*, (http://journal.sjdm.org/).

[11] At the time of writing, JBDM actually did implement a data sharing policy: "*Journal of Behavioral Decision Making* encourages authors to share the data and other artefacts supporting the results in the paper by archiving it in an appropriate public repository. Authors should include a data accessibility statement, including a link to the repository they have used, in order that this statement can be published alongside their paper." (retrieved from http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1099-0771/homepage/ForAuthors.html, October 2017). We emailed JBDM's editorial office to ask when they changed their data policy and if it had stayed the same from 2006 to 2014, but unfortunately they did not reply. Based on information from web archives, we can see that in July 2017 this data policy was not yet part of the author guidelines and therefore does not affect our conclusions (information retrieved from https://web.archive.org/web/20170713015402/http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1099-0771/homepage/ForAuthors.html, October 2017).

4.1.1.2 *Procedure*

For each article, we coded in which journal and time period it was published and whether the (raw) data were published alongside the articles. Published data files in matrix format with subjects in the rows (so no correlation matrices) as well as simulation codes and model codes were considered open data. The data had to be published either in the paper, an appendix, the journal's website, or a website with a reference to that website in the paper. Remarks such as "data are available upon request" were not considered open data (as such promises are often hollow; Krawczyk & Reuben, 2012). Note that we did not assess whether any published data were also relevant, usable, and/or complete, which is by no means guaranteed (Kidwell et al., 2016).

We assessed the consistency of the reported statistical results through an automated procedure: an adapted version[12] of the R package "statcheck" (version 1.0.0; Epskamp & Nuijten, 2014). Statcheck extracts NHST results and recomputes *p*-values in the following steps. First, statcheck converts PDF and HTML files into plain text files and extracts statistical results based on *t*-tests, *F*-tests, correlations, *z*-tests, and χ^2-tests that are reported completely (i.e., test statistic, degrees of freedom, and *p*-value) and according to the guidelines in the APA Publication Manual (American Psychological Association, 2010). Next, the extracted *p*-values are recomputed based on the reported test statistic and degrees of freedom. Finally, statcheck compares the reported and recomputed *p*-value, and indicates whether they are congruent. Incongruent *p*-values are marked as an inconsistency, and incongruent *p*-values that possibly change the statistical conclusion from significant to nonsignificant (and vice versa) are marked as a gross inconsistency.

The program statcheck contains an automated one-tailed test detection: if the words "one-tailed", "one-sided", or "directional" are mentioned somewhere in the article *and* a *p*-value would have been consistent if it was one-sided, it is counted as consistent. Furthermore, statcheck takes rounding of the reported test statistic into account. Take for instance the result t (48) = 1.43, p = .158. Recalculation would give a *p*-value of .159, which seems incongruent with the reported *p*-value. However, the true *t*-value could lie in interval [1.425, 1.435), with *p*-values ranging from .158 to .161, statcheck will count any *p*-value within this range as consistent. We assumed that all studies retained an overall alpha of .05. We also counted results reported as p = .05 as significant, since previous research showed that over 90% of the instances in which p = .05 was reported, the authors interpreted the result as significant (Chapter 2). Finally, note that when erroneously only one of the three components of an NHST result (test statistic, degrees of freedom, or *p*-value) is adjusted to correct for

[12] In the conversion from PDF to plain text, "=" signs were often translated to "¼". We adapted statcheck such that it would also recognize these cases. Furthermore, the downloaded articles contained a non-standardly reported test results that statcheck wrongly recognized as chi-square tests. This we also fixed in this adapted version of statcheck.

multiple testing, post-hoc testing, or violations of assumptions, the result becomes internally inconsistent and statcheck will flag it as such. However, in an extended validity study of statcheck, we found that such statistical corrections do not seem to cause the high estimates of the general prevalence of inconsistencies (for details, see Chapter 3). For a more detailed explanation of statcheck, see Chapter 2, or the statcheck manual at http://rpubs.com/michelenuijten/statcheckmanual.

In Chapter 2 we investigated the validity of statcheck and found that the interrater reliability between manual coding and statcheck was .76 for inconsistencies and .89 for gross inconsistencies. In an additional validity study, we found that statcheck's sensitivity (true positive rate) and specificity (true negative rate) were high: between 85.3% and 100%, and between 96.0% and 100%, respectively, depending on the assumptions and settings. The overall accuracy of statcheck ranged from 96.2% to 99.9%. For details, see Appendix A in Chapter 2 and the additional validity study, see Chapter 3.

Using statcheck, we extracted 6,482 statistical results from 498 of the 764 articles (65.2%) that contained APA reported NHST results. Note that the conversion of articles to plain text files can be different for PDF and HTML files, which can cause statcheck to recognize or miss different statistical results. Since all articles for JBDM were PDF files, and all articles in JDM HTML files, we could not reliably compare overall inconsistency rates between the journals. However, since over time the file types for each journal stayed the same, we could compare change in inconsistencies over time between the journals. All tests in this study are two-tailed unless otherwise specified and we maintained an alpha level of .05.

4.1.2 Results

4.1.2.1 *General Descriptives*

In total, we extracted 6,482 NHST results, which is on average 13.0 NHST results per article. On average, the articles in JBDM contained more NHST results than JDM articles (15.4 and 10.9 results, respectively). We found that on average 9.3% of the reported NHST results within an article was inconsistent and 1.1% grossly inconsistent. These inconsistency rates are similar to what we found in previous research (9.7% and 1.4%, respectively; Chapter 2).

Note that the general prevalence of inconsistencies can be estimated in several ways. A first way is to look at the complete set of NHST results, and calculate which percentage of these are inconsistent or grossly inconsistent. The downside of this method is that it does not take into account that results within one article may be statistically dependent. A second method is to calculate for each article which proportion of reported NHST results are inconsistent, and average this over all articles. The downside of this method is that articles with fewer results get as much weight in the calculations as articles with more results, whereas they contain less (precise) information. The third method is to use multilevel logistic models

that estimate the probability that a single NHST result is inconsistent while including a random effect at the article level. The downsides of this method are that the assumption of normally distributed random effects may be violated and that the conversion of logits to probabilities in the tails of the distribution leads to inaccurate probability estimates. Taking into account the pros and cons of all these methods, we decided to focus on the second method: the average of the average percentage of inconsistencies within an article, which we call the "inconsistency rate". We retained this method throughout this chapter to estimate the general prevalence of inconsistencies. To test relations between inconsistencies and open data or open data policies, we used multilevel models.

4.1.2.2 ***Confirmatory analyses***

Our first hypothesis was that JDM would show a larger decrease in (gross) inconsistencies than JBDM after the introduction of the data sharing policy in JDM. However, the mean prevalence of (gross) inconsistencies actually shows a pattern opposite to what we expected: the inconsistency rate increased in JDM after its open data policy from 9.7% to 11.0%, and the inconsistency rate decreased in JBDM from 9.1% to 7.0% (see Table 4.2). For illustration purposes, we also plotted the inconsistency rates in both journals over time in Figure 4.1. The Figure shows a drop in the inconsistency rate in JDM in 2013 onwards (two years after introduction of the data policy). However, there are only few inconsistencies in absolute sense in 2013 and 2014, which makes it hard to interpret this drop substantively; this decrease is in line with only random fluctuations from year to year. More details about the general trends in (gross) inconsistencies over time can be found in the Supplemental Materials at https://osf.io/5j6tc/.

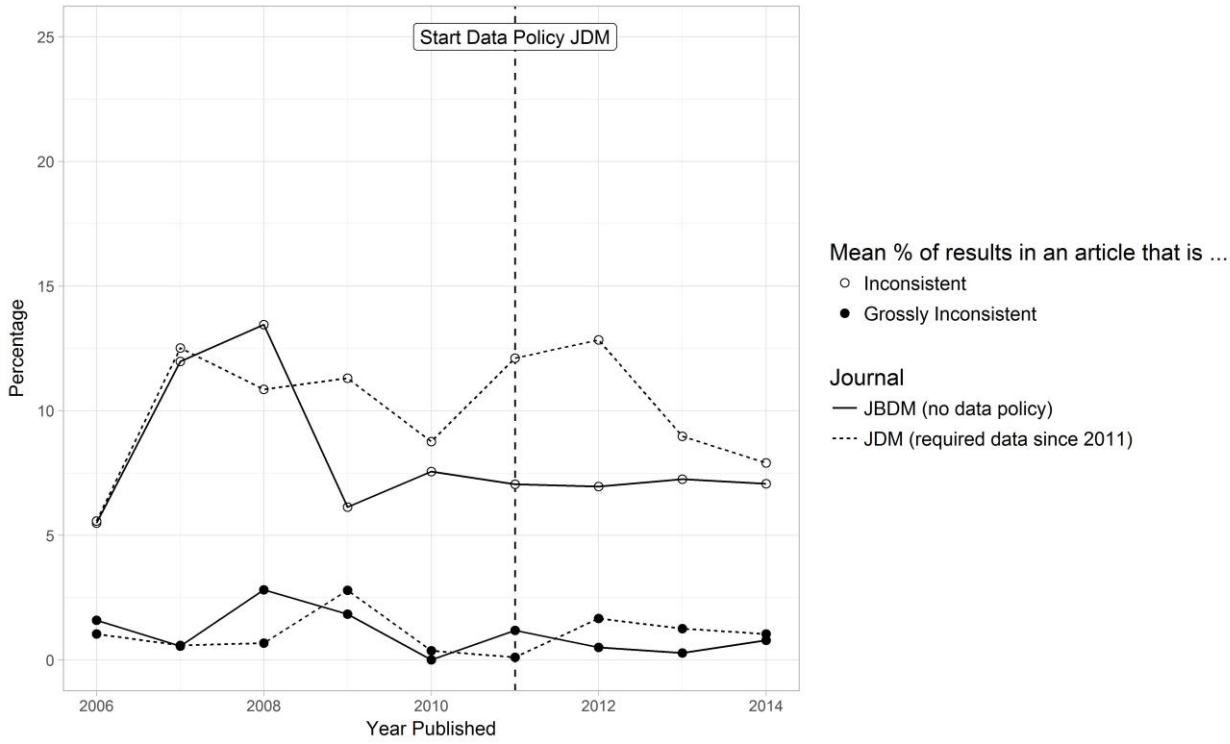


Figure 4.1

Per publication year and journal the average percentage of results within an article that was inconsistent or grossly inconsistent (the "inconsistency rate").

We tested the interaction between journal and the period in which a paper was published with a multilevel logistic regression analysis in which we predicted the probability that a *p*-value was (grossly) inconsistent with Time (0 = before data sharing policy, 1 = after data sharing policy), Journal (0 = JBDM, 1 = JDM), and the interaction Time * Journal:

$$Logit[(gross)\ inconsistency] = b_0 + b_1 Time_i + b_2 Journal_i + b_3 Time_i * Journal_i + \theta_i,$$

Equation 4.1

where subscript *i* indicates article, Time is the period in which an article is published (0 = published before JDM's data sharing policy, 1 = published after JDM's data sharing policy), Journal is the journal in which the article is published (0 = JBDM and JDM = 1), and θ_i is a random effect on the intercept b_0. We included a random intercept because the statistical results are nested within article, which means there can be dependency in the inconsistencies within the same article.

The interaction effect was not significant (b = 0.37, 95% CI = [-0.292; 1.033], *Z* = 1.10, *p* = .273), which means that there is no evidence that changes in inconsistencies over time differed for the journals. Second, we looked at the change in the prevalence of gross inconsistencies, but these showed patterns opposite to those expected as well. The gross

inconsistency rate stayed at 1.1% in JDM after its open data policy, whereas the gross inconsistency rate in JBDM decreased from 1.4% to 0.6%. To test this finding, we performed the same multilevel logistic regression with Time, Journal, and Time * Journal as predictors, but this time we predicted the probability that a *p*-value was a *gross* inconsistency. Again, we included a random effect for article. In this analysis, too, we found that the interaction effect was not significant (b = 0.58, 95% CI = [-1.412; 2.580], $Z = 0.57$, $p = .566$), meaning that there is no evidence that any change in gross inconsistencies over time depends on journal.

Table 4.2

Number of (gross) inconsistencies per journal (JDM = Judgment and Decision Making and JBDM = Journal of Behavioral Decision Making) and time period (T1 = published in 2006-Feb 2011 and T2 = published in April 2011-2014). In April 2011 JDM started encouraging open data.

		# articles	# articles with APA reported NHST results	# articles with APA reported NHST results and open data	# APA reported NHST results	average # APA reported NHST results per article	average % inconsistencies per article	average % gross inconsistencies per article
JBDM	**T1**	157	117 (74.5%)	0	1,543	13.2	9.1%	1.4%
	T2	149	118 (79.2%)	2	2,074	17.6	7.0%	0.6%
JDM	**T1**	236	128 (54.2%)	11	1,313	10.3	9.7%	1.1%
	T2	222	135 (60.8%)	118	1,552	11.5	11.0%	1.1%
Total		764	498 (65.2%)	131	6,482	13.0	9.3%	1.1%

Our second hypothesis was that articles that are accompanied with open data contain fewer (gross) inconsistencies than articles that are not accompanied with open data. Again, we observed the opposite pattern in the prevalence of inconsistencies: on average, in articles without open data 8.8% of the results was inconsistent as opposed to 10.7% in articles with open data (see Table 4.3). To test this pattern, we again fitted a multilevel logistic regression model in which we predicted the probability that a *p*-value was an inconsistency with Open Data (0 = the *p*-value is from an article without open data, 1 = the *p*-value is from an article with open data), and a random effect for article. Open Data did not significantly predict whether a *p*-value was inconsistent (b = 0.30, 95% CI = [-0.069; 0.672], $Z = 1.59$, $p = .111$). Next, we looked at the relation between gross inconsistencies and open data. We found a pattern in the predicted direction: articles with open data had on average a lower rate of gross inconsistencies than articles without open data (1.0% of the results versus 1.1%, respectively). To test this relation, we fitted a multilevel logistic regression model to see if Open Data predicts the probability that a *p*-value is a gross inconsistency, including a random effect for article. Again, Open Data was not a significant predictor (b = 0.001, 95% CI = [-1.150; 1.153], $Z = 0.002$, $p = .998$). A problem with this analysis is that the large majority of papers with open data were published in JDM, which makes this analysis a comparison of the inconsistency rates in both journals. Since we only have HTML files from JDM and only PDF files from JBDM, this comparison could therefore reflect differences in the performance of statcheck instead of an actual difference in inconsistency prevalence.

Table 4.3

Number of (gross) inconsistencies in articles with and without open data.

Open data?	# articles with APA reported NHST results	average % inconsistencies per article	average % gross inconsistencies per article
No	367	8.8%	1.1%
Yes	131	10.7%	1.0%

Based on our analyses we found no evidence for our two hypotheses: JDM did not show a larger decrease in inconsistencies and gross inconsistencies than JBDM after the introduction of the data sharing policy in JDM. We also did not find that articles that are accompanied by open data contained fewer inconsistencies or gross inconsistencies than articles without open data, but this analysis is possibly confounded.

4.1.3 Conclusion

In this study, we investigated whether there is a relationship between recommended data sharing and statistical reporting inconsistencies, by comparing the number of

inconsistencies over time in the journal JDM, which introduced a data sharing policy, and JBDM, that has no such policy. We hypothesized that JDM would show a stronger decrease in (gross) inconsistencies than JBDM (open policy effect), and that *p*-values from articles accompanied by open data were less likely to be inconsistent (open data effect). We found no evidence of an open policy effect or an open data effect.

It is worth noting that even though we found no relation between data sharing policy and reporting inconsistencies, the data sharing policy of JDM did result in the retraction of an article after anomalies in the (open) data were discovered.[13] This emphasizes the potential importance of open data (Simonsohn, 2013; Wicherts, 2011).

The main limitation of this study is its lack of power. Even though we downloaded a considerable number of articles for each cell in the design, statcheck did not retrieve statistics from every paper, and of the retrieved statistics only a small percentage was inconsistent, resulting in potentially underpowered regression analyses. Based on these data alone we cannot draw firm conclusions about the relation between data sharing and reporting inconsistencies. We therefore designed Studies 2 and 3 to obtain more power and thus more reliable results.

4.2 Study 2

In Study 2 we compared the prevalence of inconsistencies and gross inconsistencies in psychological articles the open access journal Frontiers in Psychology (FP) and in journals from the major open access publisher PLOS. From March 1st 2014 onwards PLOS required submissions to be accompanied with open data. Their online policy on data availability states that "The data underlying the findings of research published in PLOS journals must be made publicly available. Rare exceptions may apply and must be agreed to with the Editor." (https://www.plos.org/editorial-publishing-policies; retrieved October 2017). Furthermore, all submissions had to have an official Data Availability Statement explaining how the data were shared or why the data could not be shared. (Bloom et al., 2014). Not sharing data could affect the publication decision. The author guidelines of FP also state that data must be made available, but the guidelines are not as explicit as those of PLOS: FP does not require a standardized data availability statement, and it is not clear if not sharing data could affect the publication decision. [14] We again hypothesized that the inconsistencies and gross

[13] See Uri Simonsohn's post on Data Colada: http://datacolada.org/2013/09/17/just_posting_it_works/ and the analysis of the case by Retraction Watch: http://retractionwatch.com/2013/09/10/real-problems-with-retracted-shame-and-money-paper-revealed/#more-15597.

[14] FP's data policy: "*To comply with best practice in their field of research, authors must also make certain types of data available to readers at time of publication in stable, community-supported repositories such as those listed below, unless in case of serious confidentiality concerns (for example, research involving human subjects). Although not mandatory, authors may also consider the deposition of additional data-types (see below).*" FP's editorial office let us know via email that they supported the TOP guidelines since 2015: "*Frontiers supports the*

inconsistencies in articles from PLOS would show a stronger decrease (or less strong increase) over time than in FP. Furthermore, we again hypothesized that data sharing (regardless of whether it was required) is associated with fewer inconsistencies and gross inconsistencies in an article.

4.2.1 Method

4.2.1.1 *Preregistration*

The hypotheses as well as the design and analysis plan were preregistered and can be found at https://osf.io/a973d/. The hypotheses, procedure, and power analysis were registered in detail, whereas the analysis plan was registered more generally, and consisted of the regression equations we intended to test. We followed our preregistered plan, except for one detail: we did not preregister any data exclusion rules, but we did exclude one article from the analysis because it was unclear when it was received.

4.2.1.2 *Sample*

We downloaded all articles available in HTML from FP and all HTML articles with the topic "Psychology" from PLOS in two time periods to capture change in inconsistencies before and after the introduction of PLOS' requirement to submit raw data along with an article.

4.2.1.2.1 Articles from FP

We already had access to all FP articles published from 2010 to 2013 that were downloaded for the research in Chapter 2. On top of that, in the period of 9 to 15 June 2015 we manually downloaded all FP articles published from January 1st 2014 up until April 30th 2015. In total we had 4,210 articles published from March 8th 2010 to April 30th 2015.

For our sample, we selected only the research articles, clinical (case) studies, and methods articles (excluding editorials, retractions, opinions, etc.). We used systematic text searches in R to automatically select these articles, which resulted in 2,693 articles. Next, we also used systematic text searches in R to extract whether the articles were received before

Transparency and Openness Promotion (TOP) guidelines, which state that materials, data, and code described in published works should be made available, without undue reservation, to any qualified researcher, to expedite work that builds on previous findings and enhance the reproducibility of the scientific record." Both quotes retrieved from http://home.frontiersin.org/about/author-guidelines, Materials and Data Policies, May 17, 2017.

or after PLOS' data sharing policy[15] that came into effect March 1st 2014.[16] 1,819 articles in the sample were received before the policy and 873 after the policy. One article was excluded because it was unclear when it was received. Table 4.4 shows the number of downloaded articles per period and journal.

4.2.1.2.2 Articles from PLOS

PLOS has the option of selecting articles based on the date they were received, which made it straightforward to download articles and categorize them in received before or after PLOS' data sharing policy. Using the R package rplos (Chamberlain et al., 2014) we first automatically downloaded all PLOS articles with the subject "Psychology" that were received before March 1st 2014, which rendered 7,719 articles. Next, we downloaded all "Psychology" articles received after March 1st 2014, rendering 1,883 articles. We restricted this sample to articles that were published in the same time span that the FP articles were published, which means that we excluded all PLOS articles published before March 8th 2010 (4 articles excluded) or after April 30th 2015 (376 articles excluded). Next, using systematic text searches in R we only selected the research articles from this sample,[17] rendering 7,700 articles from before the data sharing policy, and 1,515 articles from after the policy. The final sample size is described in Table 4.4.

[15] We could use systematic text searches because all research articles in FP have a standard header indicating the type of article. We included articles with the header "Original Research ARTICLE", "Clinical Trial ARTICLE", "Methods ARTICLE", and "Clinical Case Study ARTICLE", which resulted in 2,693 articles. We also wanted to extract whether the articles were received before or after PLOS' data sharing policy that came into effect March 1st 2014. In FP, this is also systematically indicated at the bottom of the article (e.g., "Received: 22 October 2010; Paper Pending Published: 10 November 2010; Accepted: 01 December 2010; Published online: 14 December 2010"). Because these dates were always reported in the same place and in the same way, we could use systematic text searches in R again to extract when the articles were received and published.

[16] The exact date at which the open data policy at PLOS was implemented is not entirely clear. In the editorial announcing the policy it was stated the policy was implemented at March 1st 2014 (Bloom et al., 2014), but at the data availability web page, it was stated that the starting date was March 3rd (http://journals.plos.org/plosone/s/data-availability). For our study we retained March 1st.

[17] Similar to articles in FP, PLOS articles also have a standard header indicating the type of article. Again, we used systematic text searches in R to identify the research articles, but for this it was not enough to only search for "Research Article", since this phrase could also just occur in the full text of the manuscript. We therefore also specified the context in which the phrase "Research Article" should occur. We included either the phrase "Open Access Peer-Reviewed Research Article" or "Browse Topics Research Article", rendering 7,700 articles from before the data sharing policy, and 1,515 articles from after the policy.

Table 4.4

Number of research articles downloaded from PLOS and FP before and after PLOS introduced obligatory data sharing. All articles were published between March 8th 2010 and April 30th 2015.

	Before PLOS' data sharing policy: Received before March 1st 2014	After PLOS' data sharing policy: Received after March 1st 2014	
			Total
FP	1,819 articles	873 articles	2,692 articles
PLOS	7,700 articles	1,515 articles	9,215 articles
Total	9,519 articles	2,388 articles	11,907 articles

4.2.1.3 *Power analysis*

Based on the number of downloaded articles and the previous results from Chapter 2, we conducted a power analysis. We retained a baseline probability that a result in FP or PLOS was inconsistent of 6.4%.[18] We concluded that we have a power of .80 if the decrease in inconsistencies in PLOS over time is 2 to 3 percentage points steeper than in FP. The full details of the power analysis including all R code have been included in the preregistration and can be found at https://osf.io/ay6sh/.

4.2.1.4 *Procedure*

We used statcheck version 1.0.2 (Epskamp & Nuijten, 2015) to extract all APA reported NHST results from the PLOS and FP articles. Due to feasibility constraints, we decided not to check all the downloaded articles for open data, but only the ones that statcheck extracted results from (1,108 articles from FP and 2,909 articles from PLOS[19]).

For each downloaded article with detectable NHST results, we coded whether the (raw) data were available. Published data files in matrix format with subjects in the rows (so no correlation matrices) were considered open data. The data had to be published either in the paper, an appendix, or a website (with a reference to that website in the paper). Remarks such as "data are available upon requests" were not considered open data. Again, we did not assess whether any available data were relevant, usable, and/or complete.

[18] Note that this probability is smaller than one would expect based on the general inconsistency prevalence in Chapter 2. This is due to the estimation method in the power analysis, which takes into account the random intercept, resulting in a lower probability of an inconsistency than observed directly in the data.

[19] It is possible that our sample contained articles that contained reporting inconsistencies because those inconsistencies were the topic of investigation (Bakker & Wicherts, 2014; Veldkamp et al., 2014; Wicherts et al., 2011). However, this sample is so small that it is unlikely to affect our general conclusions.

Due to the large number of articles that needed to be coded with respect to data availability, we had seven coders: the six authors and a student assistant. We tested the coding protocol by assessing interrater reliability by coding 120 articles that were randomly selected from the full sample and calculating the intraclass correlation (ICC). In this set-up, every article was coded by two randomly selected coders. Per article three variables were coded. We coded whether the authors stated that the data was available (ICC(2, 2) = .948), whether there actually was a data file available (ICC(2, 2) = .861), and finally whether there was a URL linking to the data available (ICC(2, 2) = .282). The last ICC was quite low. After further inspection of the coding it turned out that there was some confusion among coders whether a link to a data file that was also embedded in the article should be counted as a URL. Since this was not crucial for testing our hypotheses, we adapted the protocol to only code two variables: whether the authors state that the data were available, and whether the data actually were available. The final protocol is available on https://osf.io/yq4mt/.

The total sample was coded for open data with the help of an extra student assistant, resulting in eight coders in total. As a final reliability check, 399 articles (approximately 10% of all articles with APA reported NHST results) were coded twice by randomly assigned coders. The interrater reliability was high: for the data availability statement the ICC(2, 2) was .900[20], and for whether the data was actually available the ICC(2, 2) was .913[21]. Furthermore, the first author blindly recoded all cases in which a coder had added a remark, and solved any discrepancies by discussion. The first author also solved any discrepancies between coders when an article was coded twice. All coders were blind for the statcheck results, but not for the journal and time period in which the article was published.

4.2.2 Results

4.2.2.1 *General Descriptives*

Table 4.5 shows the descriptive results per journal and time. It turned out that statcheck extracted NHST results from more articles than expected based on the data from Chapter 2. On average, 41.2% of the articles in FP and 31.6% of the articles in PLOS contained APA reported NHST results that statcheck could detect. This means that we obtained more power than expected based on our power analysis. Across journal and time, on average 13.0% of NHST results in an article was inconsistent, and 1.6% was grossly inconsistent. The average percentage of inconsistencies within an article in FP increased over time from 13.1% to 16.2%,

[20] Discrepancies in coding data availability statements mainly arose in PLOS articles before they introduced the standardized data availability statements. There were also a few instances in which coders disagreed whether a statement such as "all *relevant* data are available" could be counted as a data availability statement.

[21] Discrepancies in coding data availability mainly arose in cases where the shared data deviated from "standard" experimental data (e.g., in a meta-analysis or in genetic research), or when data about the stimuli were confused with collected data.

whereas the inconsistency rate in PLOS increased from 12.5% to 13.5%. The percentage of gross inconsistencies in FP increased slightly from 1.7% to 2.0%, and the gross inconsistencies in PLOS increased from 1.4% to 1.7%. The steeper increase in inconsistencies in FP as compared to PLOS seems to be in line with our hypothesis that an open data policy influences the inconsistency rates, but we will test this in the next section. For the sake of completeness, we also added a plot that shows the trends over time in the inconsistency rates per journal (see Figure 4.2). Note that this plot shows the average inconsistency rates in the year the articles were published, not the years in which the articles were received. That means that even though some articles were published after PLOS introduced the data policy, they may have been submitted before the policy was implemented. Even so, the figure gives a good indication of the prevalence of (gross) inconsistencies in PLOS and FP over time. More details about the general trends in (gross) inconsistencies over time can be found in the Supplemental Materials at https://osf.io/5j6tc/.

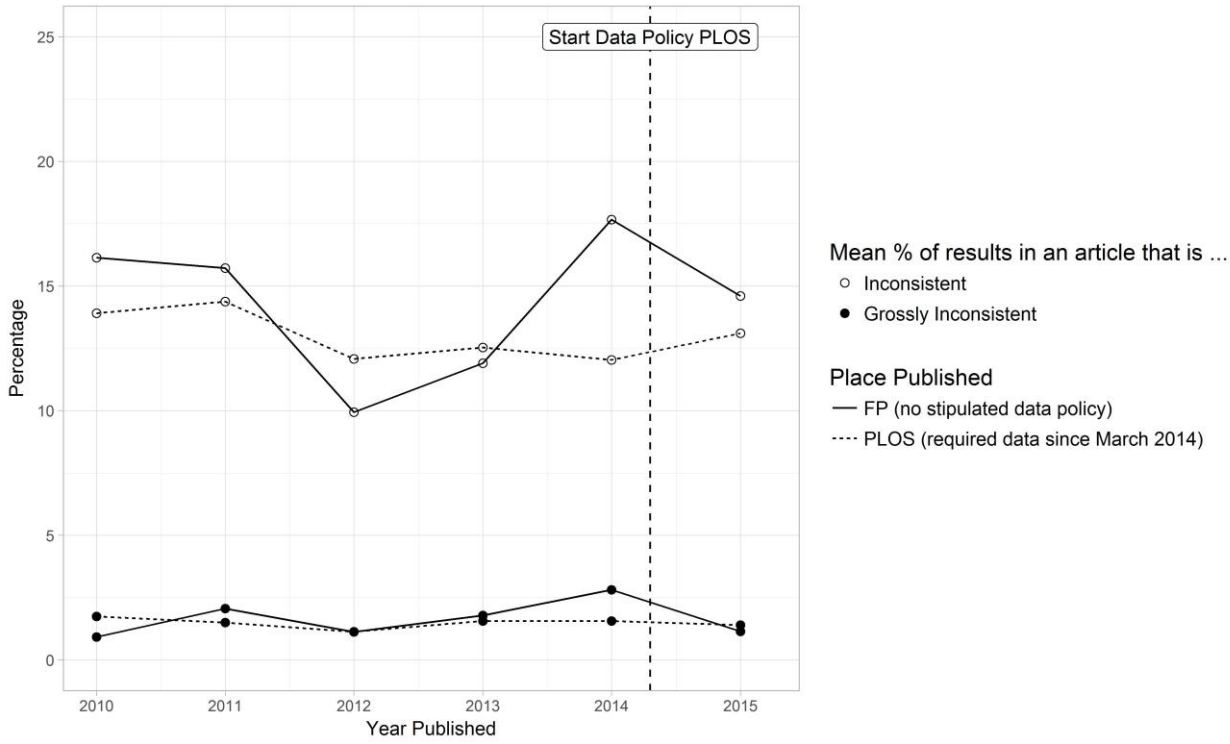


Figure 4.2

Per publication year and place published the average percentage of results within an article that was inconsistent or grossly inconsistent (the "inconsistency rate").

Table 4.5

Number of (gross) inconsistencies per journal (FP and PLOS) and time period (T1 = received before March 1st 2014 and T2 = received after March 1st 2014). PLOS required articles submitted after March 1st 2014 PLOS to be accompanied by open data.

		# articles	# articles with APA reported NHST results	# articles with APA reported NHST results and open data	# APA reported NHST results	average # APA reported NHST results per article	average % inconsistencies per article	average % gross inconsistencies per article
FP	T1	1,819	804 (44.2%)	11	11,079	13.8	13.1%	1.7%
	T2	873	304 (34.8%)	4	2,432	8.0	16.2%	2.0%
PLOS	T1	7,700	2,462 (32.0%)	110	33,064	13.4	12.5%	1.4%
	T2	1,515	447 (29.5%)	247	5,801	13.0	13.5%	1.7%
Total		11,907	4,017 (33.7%)	372	52,376	13.0	13.0%	1.6%

4.2.2.2 *Confirmatory analyses*

For our first set of preregistered hypotheses we hypothesized that the probability that a result is inconsistent decreases more strongly in PLOS after they introduced a data sharing policy than in FP, where there was no data sharing policy (open policy effect). More specifically, we expected that there is a negative interaction effect of Time (0 = received before PLOS' data sharing policy, 1 = received after PLOS' data sharing policy) times Journal[22] (0 = FP, 1 = PLOS) on the probability that a result is inconsistent or grossly inconsistent. The raw probabilities of an inconsistency and gross inconsistency split up per time and journal can be found in Table 4.5. We tested our hypotheses by estimating the following multilevel logistic models:

[22] Technically, we should call this variable "Journal/Publisher", since the results from PLOS did not all come from a single article. However, for the sake of readability and consistency with the preprint, we will call this variable "Journal".

$$Logit[(gross)\ inconsistency] = b_0 + b_1 Time_i + b_2 Journal_i + b_3 Time_i * Journal_i + \theta_i,$$

Equation 4.2

where subscript *i* indicates article, Time is the period in which an article is published (0 = received before PLOS' data sharing policy, 1 = received after PLOS' data sharing policy), Journal is the outlet in which the article is published (0 = FP and PLOS = 1), and θ_i is a random effect on the intercept b_0. We included a random intercept because the statistical results are nested within article, which means there can be dependency in the inconsistencies within the same article. We hypothesized that in both models the coefficient b_3 is negative. We maintained an α of .05. We did not preregister that we would use one-tailed tests, so we tested our hypotheses two-tailed.

When predicting the inconsistencies, we found a significant interaction effect of Time * Journal in the predicted direction, b_3 = -0.43, 95% CI = [-0.77; -0.085], *Z* = -2.45, *p* = .014. This indicates that the prevalence of inconsistencies decreased more steeply (or more accurately: increased *less* steeply) in PLOS than in FP. This finding is in line with the notion that requiring open data as a journal could decrease the prevalence of reporting errors.

When predicting gross inconsistencies, we did not find a significant interaction effect of Time * Journal; b_3 = -0.12, 95% CI = [-1.04; 0.80], *Z* = -0.25, *p* = .804. This means that there is no evidence that any change in gross inconsistencies over time depended on the journal in which the result was published. This finding is not in line with our hypothesis. Since we found no significant interaction effect, we (exploratively) tested the model again without the interaction effect to see if there is a main effect for Time and/or Journal. We found no evidence for a main effect of Time (b_1 = 0.169, 95% CI = [-0.266; 0.605], *Z* = 0.762, *p* = .446) or a main effect of Journal (b_2 = -0.012, 95% CI = [-0.413; 0.438], *Z* = -0.063, *p* = .950). Note that our power analysis was based on the prevalence of inconsistencies, and not gross inconsistencies. The power of our analysis to find an effect of data sharing on the prevalence of gross inconsistencies is much lower since gross inconsistencies are much less prevalent.

For our second set of hypotheses we tested whether results in articles that are accompanied by open data have a lower probability of being inconsistent and grossly inconsistent than results in articles that are not accompanied by open data, regardless of the journal in which they were published (open data effect). We found that the average percentage of inconsistencies in an article was 13.7% when an article had open data, and 12.9% when an article did not have open data. The average percentage of gross inconsistencies in an article was 2.1% and 1.5% for articles with and without open data, respectively. These patterns are the opposite of what we expected. We tested whether there is a relationship between open data and the probability of a (gross) inconsistency by estimating the following two multilevel logistic models:

$$Logit[(gross)\ inconsistency] = b_0 + b_1 Open\ Data_i + \theta_i,$$

Equation 4.3

where subscript *i* indicates article, Open Data indicates whether the data is published along with the article (0 = no open data, 1 = open data), and θ_i is a random effect on the intercept b_0. We hypothesized that in both models the coefficient b_1 is negative.

We found no effect of Open Data on the prevalence of inconsistencies (b_1 = 0.06, 95% CI = [-0.16; 0.27], Z = 0.50, p = .617) or the prevalence of gross inconsistencies (b_1 = 0.23, 95% CI = [-0.33; 0.79], Z = 0.79, p = .429). This finding is not in line with our hypothesis that articles accompanied by open data should have lower inconsistency rates.

4.2.3 Conclusion

In this study, we investigated the relation between required data sharing and statistical reporting inconsistencies using a larger dataset than in Study 1, by comparing the number of statistical reporting inconsistencies over time in open access articles. We compared psychology articles from journals in PLOS, which since March 2014 requires articles to be accompanied by open data, with articles in FP, which does encourage data sharing, but does not require it in the same strong terms as PLOS does. We hypothesized that PLOS would show a stronger decrease in (gross) inconsistencies than FP, and that *p*-values from articles accompanied by open data were less likely to be inconsistent. We found that the prevalence of inconsistencies over time increases less steeply in PLOS than in FP, which is in line with our hypotheses. However, we did not find evidence for our other hypotheses: there is no evidence that any change in gross inconsistency prevalence is different for PLOS and FP, and we also found no relationship between open data and *p*-value inconsistency.

4.3 Study 3

In Study 3, we examined the prevalence of reporting inconsistencies in the journal Psychological Science (PS). Before 2014, the policy of PS concerning data sharing was simply the general policy of the APA, which roughly states that data should be available upon request. From 2014 onwards, however, PS has started to award so-called “Open Practice Badges” in recognition of open scientific practices (Eich, 2014). “Open Practice Badges” is a collective term for three types of badges: authors can earn an Open Data Badge, an Open Materials Badge, and a Preregistration Badge. This simple intervention has proven to be very effective: the frequency of reported data sharing in PS increased almost ten-fold after introduction of the badges, compared to reported data sharing in PS before the badges, and data sharing in four comparable journals (Kidwell et al., 2016). Furthermore, articles in PS with an open data badge had a much higher probability of actually providing the data (93.8%) than articles without a badge that promised data (40.5%; Kidwell et al., 2016).

We again theorized that open practices in general and data sharing in particular would decrease inconsistencies and gross inconsistencies. To test this, we focused on articles published in PS from 2014 onwards, because in this time frame the Open Practice Badges enable a straightforward check for the availability of data and/or engagement in other open practices (sharing materials and preregistration). One of the main advantages of this study as compared to Study 1 and Study 2 in this chapter, is that authors have to meet certain criteria before they are awarded any of the Open Practice Badges. For instance, for an Open Data Badge authors need to publish their data in an open-access repository that is time-stamped, immutable, and permanent[23]. Therefore, in this study we are better able to assess whether an article actually has (high quality) open data than in Study 1 or Study 2.[24] It is possible that articles published before 2014 also engaged in data sharing and other open practices, but due to feasibility constraints we did not attempt to code this. Furthermore, from July 2016 onwards, PS started using statcheck to screen articles for inconsistencies.[25] In our study, we only included PS articles published up until May 2016, because any drop in the prevalence of statistical reporting inconsistencies after May 2016 could have been caused by the use of statcheck in the review process instead of the introduction of the Open Practice Badges.

To investigate the relation between open practices in general and reporting inconsistencies, we tested the following two hypotheses (open practice effects), as stated in the preregistration at https://osf.io/6nujg/:

"Statistical results in articles published in PS from 2014 onwards with one or more Open Practice Badges have a lower probability to be inconsistent (Hypothesis 1) and grossly inconsistent (Hypothesis 2) than statistical results in PS articles published from 2014 onwards without an Open Practice Badge."

These hypotheses concern an effect of open practices in general (including sharing materials and preregistration), but we were also interested in the effect of open data in particular on reporting inconsistencies. To that end, we also focused on the Open Data Badges

[23] See https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ for details.

[24] We note that Kidwell et al. (2016) found that some articles that did share data did not receive an Open Data Badge, but it was unclear why. Conversely, there were also articles with an Open Data Badge that did not have available data. Even though these cases were rare, they indicate that having one of the Open Practice Badges is not necessarily a perfect indicator of open practice.

[25] "Please note: Psychological Science uses statcheck, an R program written by Sacha Epskamp and Michele B. Nuijten that is designed to detect inconsistencies between different components of inferential statistics (e.g., t value, df, and p). Statcheck is not designed to detect fraud, but rather to catch typographical errors (which occur often in psychology; see https://mbnuijten.com/statcheck/). We run statcheck only on manuscripts that are sent out for extended review and not immediately rejected after extended review. Authors are informed if statcheck detects any inconsistencies. Authors are welcome to run statcheck before submitting a manuscript (http://statcheck.io/)."
Retrieved from http://www.psychologicalscience.org/publications/psychological_science/ps-submissions#OPEN, October 2017.

specifically, by testing the following two hypotheses (open data effects) from our preregistration at https://osf.io/6nujg/:

"Statistical results in articles published in PS from 2014 onwards with an Open Data Badge have a lower probability to be inconsistent (Hypothesis 3) and grossly inconsistent (Hypothesis 4) than statistical results in articles published from 2014 onwards without an Open Data Badge."

Finally, we theorized that PS' policy to award open practice with badges has caused the journal to become known as a journal focused on open, solid science. Because of this, we speculated that after the installation of the badge policy in 2014, the articles submitted to PS were of higher quality, regardless of whether they actually received a badge or not. Therefore, we also hypothesized that (open policy effects), as stated in the preregistration at https://osf.io/6nujg/:

"Statistical results in articles published in PS before 2014 have a higher probability to be inconsistent (Hypothesis 5) and grossly inconsistent (Hypothesis 6) than statistical results in articles published in PS from 2014 onwards."

4.3.1 Method

4.3.1.1 *Preregistration*

The hypotheses and analysis plan (including the full R code) of this study were preregistered. The preregistration can be found at https://osf.io/8j56r/. All elements of the preregistration were written up in a high level of detail. We followed our preregistered plan, except for one aspect of the analysis. We preregistered the R code for the intended analyses, but did not take into account convergence problems. Our solutions to deal with these problems were ad hoc.

4.3.1.2 *Sample*

To investigate the prevalence of inconsistencies and gross inconsistencies in PS, we looked at HTML articles published in PS from 2003 to 2016. We already downloaded the articles published from 2003 to 2013 in previous research, which resulted in a sample of 2,307 articles (Chapter 2). In June 2016, a research assistant downloaded all HTML articles except editorials published from January 2014 up until May 2016, which resulted in 574 articles (see Table 4.6 for details).

4.3.1.3 *Power Analysis*

As we did in Study 2, we conducted power analyses for all hypotheses based on the number of downloaded articles and the results from Chapter 2. We concluded that for hypothesis 1 and 3 we have 80% power if the probability of an inconsistency drops with about

50% after introduction of the badges (from .049[26] to .024), and if the probability of an inconsistency drops with about 25% for hypothesis 5 (from .049 to .036; see the preregistration for details). Furthermore, we concluded that we probably do not have sufficient power to detect predictors of a reasonable size of gross inconsistencies (hypotheses 2, 4, and 6). Consequently, we do not trust the test results on gross inconsistencies. However, we still reported the results of the multilevel logistic regression analyses of gross inconsistencies for the sake of completeness. The full details of this power analysis including all R code has been included in the preregistration and can be found at https://osf.io/xnw6u/.

4.3.1.4 ***Procedure***

For the articles published from 2014 onwards a research assistant coded which (if any) badges accompanied the article. A detailed protocol (in Dutch) with instructions for the research assistant on which articles to download and how to code the open practice badges is available on OSF: https://osf.io/kktk5/. For full sample details, see Table 4.6.

Table 4.6

Total number of downloaded research articles published before and after PS introduced the Open Practice Badges, and how many of these articles were accompanied by the different badges.

Year published	Total # articles downloaded	Open Data Badge	Open Material Badge	Preregistration Badge
2003-2013	2,305[27]	0	0	0
2014-2016	574[28]	97	69	4

We used statcheck version 1.2.2 (Epskamp & Nuijten, 2016) to extract all APA reported NHST results from the downloaded PS articles and check them on internal consistency.

[26] Note that this probability is lower than one would expect based on the general inconsistency prevalence of roughly .10 in PS (Chapter 2). This is due to the estimation of the regression coefficients, which takes into account the random intercept, resulting in a lower probability of an inconsistency than observed directly in the data.
[27] In the preregistration we stated that we had 2,307 articles in total, but this seems to have been a mistake.
[28] In the preregistration we stated that we had 576 articles in total, but this seems to have been a mistake.

4.3.2 Results

4.3.2.1 *General Descriptives*

Of the 2,879 downloaded articles, 2,106 (73.2%) contained APA reported NHST results. In total, we extracted 20,926 NHST results, which is on average 9.9 NHST results per article. Per article we found that on average 9.3% of the reported NHST results was inconsistent and 1.1% grossly inconsistent. These inconsistency rates are similar to what we found in Studies 1 and 2 in this chapter, and to the results in Chapter 2.

4.3.2.2 *Hypotheses 1 & 2: Open Practice Badges*

Hypothesis 1 and 2 focused on whether the probability that a result is a (gross) inconsistency was lower if the article acquired one or more Open Practice Badges. We found that 574 articles were published in the period from 2014 onwards when PS started to award badges. In our sample, the probability that a result was inconsistent is slightly higher for articles with a badge (11.8%) than articles without a badge (9.7%), but the probability that a result was a gross inconsistency is equal in the two groups (see Table 4.7 for details).

Table 4.7

Number of (gross) inconsistencies for articles published in PS after 2014 with at least one Open Practice Badge and without any badges.

	# Articles downloaded	**# Articles with APA reported NHST results (%)**	**# APA reported NHST results**	**Average # APA reported NHST results per article**	**Average % inconsistencies per article**	**Average % gross inconsistencies per article**
No Badges	469	351 (74.8%)	4240	9.7	9.7%	1.5%
Open Practice Badge(s)	105	75 (71.4%)	1039	10.3	11.8%	1.5%
Total	574	426 (74.2%)	5279	9.8	10.0%	1.5%

We tested hypothesis 1 and 2 with the following logistic multilevel models:

$$Logit[(gross)inconsistency] = b_0 + b_1 OpenPracticeBadge_i + \theta_i,$$

Equation 4.4

where subscript *i* indicates article, OpenPracticeBadge indicates whether an article had one or more of the three available Open Practice Badges (1) or not (0), and θ_i is a random effect on the intercept b_0. We hypothesized that in both models the coefficient b_1 is negative. We tested these hypotheses maintaining an α of .05, and we tested one-sided ($b_1 < 0$).

Consistent with our preregistered analysis plan, we took into account the possibility that the year in which the paper was published could cause a spurious relation between having a badge and the prevalence of (gross) inconsistencies: it is imaginable that a gradual change in research culture caused both the prevalence of open practice badges to increase and the prevalence of (gross) inconsistencies to decrease (although Figure 4.3 does not seem to show such a trend in inconsistencies, see the next sections for more details). We therefore first intended to test whether there was an interaction effect between OpenPracticeBadge and Year on the prevalence of (gross) inconsistencies. Due to convergence problems, we re-estimated this model by altering the number of nodes in the Gauss-Hermite quadrature formula to 0 and 0.9. The results of these analyses revealed no effect of the year in which an article was published. Therefore, we proceeded with fitting the originally hypothesized models. Based on our analyses, we found no evidence for an effect of OpenPracticeBadge on the probability that a result is inconsistent (b_1 = -0.349, 95% CI = [-0.867; 0.169], z = -1.320, p = .093, one-tailed) or grossly inconsistent (b_1 = -0.894, 95% CI = [-3.499; 1.711], z = -0.673, p = .250, one-tailed).

4.3.2.3 *Hypotheses 3 & 4: Open Data Badges*

In Hypotheses 3 and 4, we looked at the relation between whether articles had an Open Data Badge or not and the probability that a result in that article was inconsistent. Of the 574 articles published in PS from 2014 onwards, 97 had an Open Data Badge and 477 did not. The average percentage of both inconsistencies and gross inconsistencies per article in this sample was higher in articles with an Open Data Badge than in articles without one (see Table 4.8 for details).

Table 4.8

Number of (gross) inconsistencies for articles published in PS after 2014 with and without an Open Data Badge.

	# Articles downloaded	**# Articles with APA reported NHST results (%)**	**# APA reported NHST results**	**Average # APA reported NHST results per article**	**Average % inconsistencies per article**	**Average % gross inconsistencies per article**
No Open Data Badges	477	354 (74.2%)	4,259	9.8	9.6%	1.5%
Open Data Badge	97	72 (74.2%)	1,020	9.8	12.0%	1.6%
Total	574	426 (74.2%)	5,279	9.8	10.0%	1.5%

We estimated the following logistic multilevel models to test Hypothesis 3 and 4:

$$Logit[(gross)inconsistency] = b_0 + b_1 OpenDataBadge_i + \theta_i,$$

Equation 4.5

where OpenDataBadge indicates whether an article had an Open Data Badges (1) or not (0). We hypothesized that in both models the coefficient b_1 is negative. We tested these hypotheses maintaining an α of .05, and we tested one-sided ($b_1 < 0$).

Similar to Hypotheses 1 and 2 and following the preregistration, we first tested the models including two extra control variables: in which year the article was published and whether the article had a badge other than an Open Data Badge. We included the latter control because we wanted to distinguish between effects of open practice in general and open data in particular. We first intended to test a three-way interaction between Open Data Badge, other badges, and year published, because if there would be a three-way interaction, any two-way interactions or main effects could not be interpreted. However, these models were too complex to fit and failed to converge. We therefore continued to fit the models with three two-way interactions. Similar to hypotheses 1 and 2, we fitted the models with the node-parameter set to 0 and 0.9. Based on these analyses, we continued to estimate the simple effects of the following model:

$$Logit[inconsistency] = b_0 + b_1 OpenDataBadge_i + b_2 OtherBadge_i + b_3 Year_i + b_4 OpenDataBadge_i * Year_i + \theta_i,$$

Equation 4.6

where we looked at the coefficients of the model when Year was centered on 2014, 2015, and 2016. The results show that the negative relation between whether an article had an Open Data Badge and the probability that a result was inconsistent was stronger for articles published in 2014 than in 2015 or 2016 (see Table 4.9 for details). This finding would be in line with a scenario in which open data (badges) led to a lower prevalence of reporting inconsistencies in 2014, but that this effect decreased over time.

Table 4.9

Results of the simple effects analysis to predict the probability that a result is inconsistent when Year is centered on 2014, 2015, and 2016. The Table shows the regression coefficients and their standard errors. The main predictor of interest, Open Data Badge, is printed in bold.

	b (SE)		
Year centered on	2014	2015	2016
Intercept	-2.96 (.15)***	-3.18 (.15)***	-3.40 (.27)***
Open Data Badge	**-1.60 (.78)***	**-0.65 (.48)**	**0.29 (.56)**
Other Badge	0.22 (.50)	0.22 (.50)	0.22 (.50)
Year	-0.22 (.16)	-0.22 (.16)	-.22 (.16)
Year * Open Data Badge	0.94 (.48)*	0.94 (.48)*	.94 (.48)*

* $p < .05$; ** $p < .01$; *** $p < .001$.

Then, to predict the probability that a result was grossly inconsistent, we fitted a model including the two-way interactions to compare it with a model with only the main effects. However, the model with the two-way interactions was too complex to fit, and failed to converge. We therefore continued with the model with only main effects, which we again fitted with the node-parameter set to 0 and 0.9. We compared these models with a model with only Open Data Badge as a predictor and found that adding control variables did not significantly improve the model ($\chi^2(2) = .531$, $p = .767$). Based on the final model including only Open Data Badge as a predictor, we found that there was no significant relation between the probability that a result was grossly inconsistent and whether the article had an Open Data Badge or not (b = -.869, 95% CI = [-3.481; 1.743], $z = -0.652$, $p = .257$, one-tailed).

4.3.2.4 *Hypotheses 5 & 6: Time Period*

For Hypothesis 5 and 6 we were interested if there was a change in the probability that a result was (grossly) inconsistent when PS started to award badges, so we looked at articles published in PS before and after 2014 when the badge system was introduced. In our sample, we had 2,305 downloaded articles from before 2014, and 574 articles from 2014 onwards. The prevalence of inconsistencies and gross inconsistencies was slightly higher in the second period (see Table 4.10 for details). We were interested in the difference in inconsistency rates before and after the introduction of the badges, but to sketch a more complete picture we also plotted the inconsistency rates per year (see Figure 4.3). This figure shows that there is a steep drop in the inconsistency rate in articles that were published after the Open Practice Badges were introduced, but that this trend is inconsistent. More details about the general

trends in (gross) inconsistencies over time can be found in the Supplemental Materials at https://osf.io/5j6tc/.

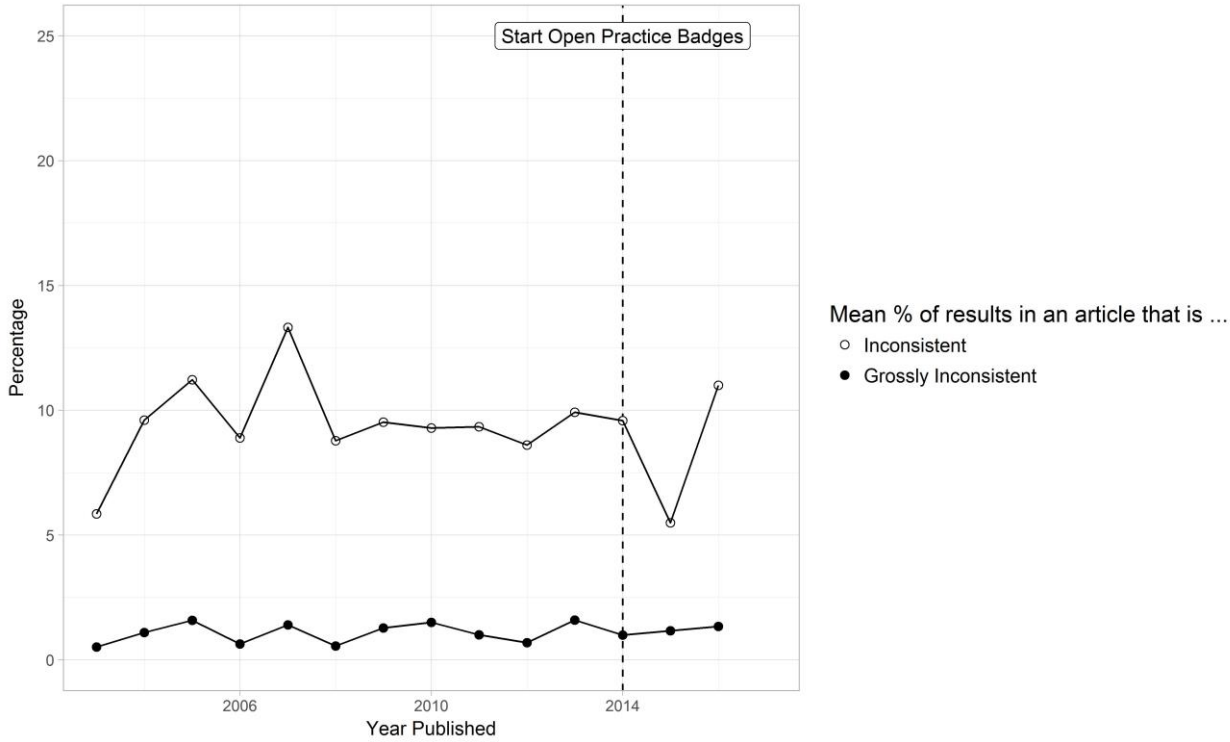


Figure 4.3

The average percentage of results within an article that was inconsistent or grossly inconsistent (the "inconsistency rate") per publication year.

Table 4.10

Number of (gross) inconsistencies for articles published in PS before 2014 (Period 1) and from 2014 onwards (Period 2). From 2014 onwards PS started to award Open Practice Badges.

	# Articles downloaded	**# Articles with APA reported NHST results (%)**	**# APA reported NHST results**	**Average # APA reported NHST results per article**	**Average % inconsistencies per article**	**Average % gross inconsistencies per article**
Period 1	2,305	1,680 (72.9%)	15,647	10.0	9.1%	1.1%
Period 2	574	426 (74.2%)	5,279	9.8	10.0%	1.5%
Total	2,879	2,106 (73.2%)	20,926	9.9	9.3%	1.1%

We tested our hypotheses using the following multilevel logistic models:

$$Logit[(gross)\ inconsistency] = b_0 + b_1 Period_i + \theta_i,$$

Equation 4.7

where Period indicates the time period in which the article was published (0 = T1, published before 2014 and the badge policy; 1 = T2, published from 2014 onwards when the badge policy was installed). Again, we included a random intercept to account for dependencies of results within articles. We hypothesized that in both models the coefficient b_1 is negative. We tested this hypothesis maintaining an α of .05 using a one-sided ($b_1 < 0$) test.

Following the strategy from the previous hypotheses, we first intended to test the models controlling for possible effects of whether an article had any of the badges, and the specific year in which the article was published. Again, we first intended fit the models including a three-way interaction between Period, Badges, and Year, and in case there was no significant three-way interaction continue with a model with all two-way interactions, as we preregistered. However, we later realized that testing an interaction between Period and Badges does not make sense because badges were always awarded in T2. Similarly, any interaction between Year and Period also does not make sense, because all years up to 2014

were per definition T1 and from 2014 onwards T2.[29] We therefore ran models with a main effect for Period and only one two-way interaction between Badges and Year. Including this two-way interaction did not improve the models, so we continued to fit the models including all main effects and compared them to the models with only Period as predictor. The models that included all main effects did not significantly improve in fit as compared to the models with only Period as predictor when predicting inconsistencies ($\chi^2(2) = 2.244$, $p = .326$) or gross inconsistencies ($\chi^2(2) = 0.263$, $p = .877$). We therefore proceeded with fitting the originally hypothesized models.

In line with our hypothesis, we found evidence that a result has a lower probability of being inconsistent when it was published from 2014 onwards (b_1 = -0.204, 95% CI = [-0.424; 0.015], z = -1.823, p = .034, one-tailed). Note that this conclusion differs from the descriptives in Table 4.10 that show that the average percentage of inconsistencies actually increased from Period 1 to 2 (from 9.1% to 10.0%). These differences in results arise because these analyses reflect different ways to estimate the prevalence of inconsistencies, each with its own advantages and disadvantages (see the section General Descriptives in Study 1 for details). However, despite these seemingly discrepant results for both methods, the effect of open data policy was invariably very small at best. When we looked at gross inconsistencies, we found no evidence for an effect of Period on the probability that a result is grossly inconsistent (b_1 = -0.186, 95% CI = [-1.140; 0.768], z = -0.382, p = .351, one-tailed).

The full details on the analyses of hypotheses 1 through 6 and the ad-hoc solutions to the convergence problems can be found in the Supplemental Information at https://osf.io/4gx53/ and in the R code at https://osf.io/8e3gr/.

4.3.3 Conclusion

In Study 3, we documented the prevalence of reporting inconsistencies in the journal *Psychological Science*. We hypothesized that articles with any of the Open Practice Badges had a lower prevalence of inconsistencies and gross inconsistencies than articles without any badges, but we found no evidence to support this. Furthermore, we hypothesized that articles with an Open Data Badge in particular had a lower prevalence of inconsistencies and gross inconsistencies than articles without an Open Data Badge. We found that for articles published in 2014 there was a lower probability that a result was inconsistent if an article had an open data badge, but this pattern did not hold for other years or for gross inconsistencies. Finally, we hypothesized that the prevalence of inconsistencies and gross inconsistencies was lower from 2014 onwards, when PS installed the badge policy. We found evidence that the prevalence of inconsistencies was indeed lower from 2014 onwards than before 2014, but this only held when we looked at the multilevel logistic models and were not in line with the

[29] We thank Julia Rohrer for pointing this out to us in her review.

descriptives in Table 4.10. Furthermore, we did not find a similar pattern for gross inconsistencies. Our results indicate that if there is any effect of the introduction of the policy on reporting inconsistencies, it is very small at best.

4.4 Exploratory Findings across Studies 1, 2, and 3

We distinguish between confirmatory and exploratory analyses. Confirmatory analyses are intended to test a priori formulated hypotheses, as opposed to exploratory analyses, which are more data-driven. Although confirmatory findings are more reliable than exploratory findings, exploratory findings can be important in formulating new hypotheses. As long as the distinction is made clear, both confirmatory and exploratory findings have their own merits (see also Wagenmakers et al., 2012).

The results of Studies 2 and 3 in the sections above can be considered purely confirmatory, since we preregistered the hypotheses, procedure, and analysis plans. This also means that the results of Study 1 cannot be considered purely confirmatory, because this study was not preregistered. Beside confirmatory analyses, we also performed several additional, more explorative analyses. We looked at cases in which data were promised but not delivered, the effectiveness of journal policies on data sharing, and whether articles with different types of gross inconsistencies also differ in how often they have open data. Finally, we also looked at the prevalence of inconsistencies over time, but since we did not find clear trends (similar to the findings in Chapter 2), we only included these results in the Supplemental Information at https://osf.io/5j6tc/. We did not test any of the exploratory findings for statistical significance, because *p*-values are only interpretable in confirmatory tests (see Wagenmakers et al., 2012).

4.4.1 Data missing when promised

A large part of this study focuses on the availability of research data. Ideally, open data should follow the FAIR Guiding Principles (Wilkinson et al., 2016), which state that data should be Findable, Accessible, Interoperable, and Reusable. Here, we only focused on the first and least stringent of these principles: findability. However, in Study 2 (PLOS vs. FP) we noticed that in many cases articles stated that all data were available, whereas in fact this was not the case. We analyzed these cases in detail below.

We recorded 134 cases in articles from PLOS journals where data were promised but not available. This is as much as 29.0% of all PLOS articles that promised data. This is in line with the findings of Chambers (2017, p. 86), who found that 25% of a random sample of 50 PLOS papers employing brain-imaging methods stated their data was available, whereas in fact it was not. In FP, we found a similar percentage: of the twelve articles that promised data, three articles (25.0%) did not have available data. In Table 4.11 we categorized all articles from

Study 2 on whether data were promised and whether data were actually available, split up by journal.

Table 4.11

Number of articles in which data were promised or not and data were actually available or not, split up per journal. The cases in which data were promised but not available are printed in bold.

Journal	Data Available	Data Promised		
PLOS		**Yes**	**No**	**Total**
	Yes	328	32	360
	No	**134**	2,415	2,549
	Total	462	2,447	2,909
FP		**Yes**	**No**	**Total**
	Yes	9	6	15
	No	**3**	1,090	1,093
	Total	12	1,096	1,108

We examined papers that promised but did not deliver data according to the type of "missing" data. In a minority of the cases (N = 11), the data were hard or impossible to find due to broken URLs, links to Chinese websites, or directions to general data websites (e.g., http://osf.io). The large majority of cases (N = 126, all in PLOS) were articles that only reported summary data, such as tables with means and standard deviations or bar plots, instead of actual raw data files. All but two of these cases were published after PLOS started requiring open data and every published article contained an explicit data availability statement. These data availability statements roughly fell in two categories: "Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All data are included within the manuscript" (N = 9) and "Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All *relevant* data are within the paper" (italics added; N = 115).

Based on our findings, we speculate that there are two likely causes for the high rate of "missing" promised data in PLOS. First, it is possible that the definition of "data" is unclear to the authors, PLOS editorial staff, or both. Perhaps summary data are considered enough information to comply with PLOS' open data regulations. Second, a lot of flexibility is

introduced by allowing the data statement to promise all "relevant" data to be available. The word "relevant" is open to interpretation and might lead to underreporting of actual raw data files. We note that this high rate of missing promised open data is by no means unique for PLOS. A recent study found that as much as 40.5% of articles published in the journals *Clinical Psychological Science, Developmental Psychology, Journal of Experimental Psychology: Learning, Memory, and Cognition, and Journal of Personality and Social Psychology* that promised open data did not deliver (Kidwell et al., 2016). Whatever the cause may be, we are concerned about the high percentage of papers with missing open data.

4.4.2 Effectiveness open data policy

We noted that journal policy on sharing data seems highly effective. Figure 4.4 shows that the percentage of articles with open data increased dramatically right after JDM, PLOS, and PS introduced a data sharing policy (in 2011, 2014, and 2014, respectively), whereas JBDM and FP without a data policy did not show such an increase. Specifically, in Study 1 we saw that the percentage of articles in JDM with open data increased dramatically from 8.6% to 87.4% after the introduction of their data policy (see Table 4.12). Moreover, in 2013 and 2014, 100% of the articles in JDM contained open data (see Figure 4.4). In the similar journal JBDM that did not introduce a data policy, none of the articles had open data in period 1, and only 1.7% of the articles had open data in period 2 (see Table 4.12). We found a similar pattern in Study 2. There, the articles in PLOS that were accompanied by open data increased from 4.5% to 55.9% after PLOS introduced a data sharing policy. In the comparable open access journal FP without such a stringent policy, we see no such increase (1.4% to 1.3%; see Table 4.12). Note that these percentages reflect whether data are actually available or not, so despite the worrying finding that roughly a third of the articles in Study 2 that promised data did not deliver (see the previous section), we still see a steep increase in the prevalence of open data in PLOS. In Study 3, we found that after the introduction of Open Practice Badges in PS, 16.9% of the articles earned an Open Data Badge. Previous research investigating the effectiveness of the badges in more detail found that after the introduction of the badges, data was more often available, correct, usable, and complete (Kidwell et al., 2016). These results are in line with the finding that journal submission guidelines in general can inspire desirable change in authors' behavior (Giofrè, Cumming, Fresc, Boedker, & Tressoldi, 2017; but see also Morris & Fritz, 2017).

Note, however, that our design is observational, which does not allow us to draw a causal conclusion. It is imaginable that there is an alternative explanation for the increase in data availability after data policies were introduced. For instance, it is possible that the introduction of data policies changed the image of these journals, which inspired "open-science-minded" researchers who always share their data to submit to these journals instead of elsewhere. In that case, it would not be the policy per se that increased data availability,

but the way these journals present themselves. We would need an experimental design to be able to investigate whether data policies actually lead to higher data availability. For instance, one way to investigate this would be to have one or multiple journals randomly assign submissions to a "required data sharing condition" and a control condition in which no explicit requests concerning data sharing are made. This way, any systematic difference in the prevalence of statistical reporting inconsistencies between conditions is likely to be due to the presence or absence of a data sharing request.

Table 4.12

Percentage of articles that was accompanied by open data, split up per journal and period. The periods were decided per study based on the dates that one of the journals implemented their open data policy.

		% Articles with open data			
		Before implementation		**After implementation**	
Study 1		**Up to April 2011**		**From April 2011**	
	JBDM (no data policy)	0%	N = 0/117	1.7%	N = 2/118
	JDM (data policy)	8.6%	N = 11/128	87.4%	N = 118/135
Study 2		**Up to March 2014**		**From March 2014**	
	FP (no stipulated data policy)	1.4%	N = 11/804	1.3%	N = 4/304
	PLOS (data policy)	4.5%	N = 110/2462	55.9%	N = 250/447
Study 3		**Up to 2014**		**From 2014**	
	PS (data policy)	Not coded		16.9%	N = 72/426

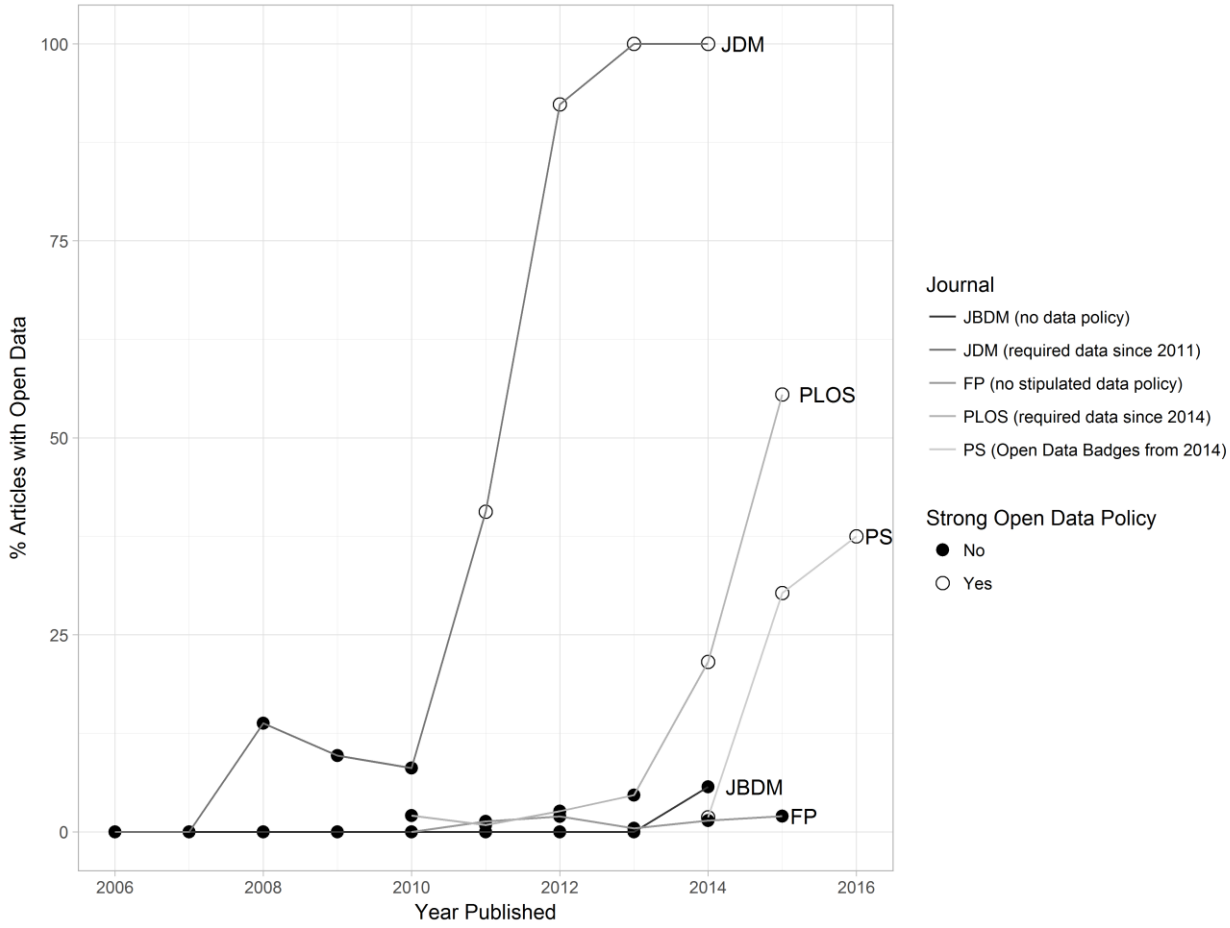


Figure 4.4

The percentage of articles per journal and year that had open data. A solid circle indicates that there was no (stipulated) open data policy at this point, and an open circle indicates that there was. The different line colors indicate the different journals. The journal abbreviations indicate the following: JBDM = Journal of Behavioral Decision Making, JDM = Judgment and Decision Making, FP = Frontiers in Psychology, PLOS = Public Library of Science, and PS = Psychological Science.

4.4.3 Open data and inconsistencies in significant vs. nonsignificant findings

In previous research we found that gross inconsistencies were more common in results reported as significant (1.56%) than as nonsignificant (0.97%), suggesting evidence for a systematic bias towards finding significance (Chapter 2). This finding can have several causes, ranging from deliberately rounding down nonsignificant *p*-values (see also John et al., 2012) to publication bias, which would primarily cause the *p*-values that are wrongly rounded *down* to be published. Because of this apparent emphasis on finding significant results, we looked in more detail at the difference between gross inconsistencies in results reported as significant and reported as nonsignificant.

We first tried to replicate our previous finding that there seems to be a systematic bias towards significant findings, using the aggregated data of Studies 1, 2, and 3. Interestingly, we

found no clear evidence for such a bias in the current data. Of all 56,716 results reported as significant, 1.26% was flagged as a gross inconsistency, as opposed to 1.23% of the 22,344 results reported as nonsignificant.[30]

Furthermore, we looked at whether the probability of data sharing was related to the type of gross inconsistencies in a paper. Specifically, we looked at the proportion of articles sharing data that 1) did not contain a gross inconsistency, 2) contained at least one gross inconsistency in general, 3) contained a gross inconsistency in a result reported as nonsignificant, and 4) contained a gross inconsistency in a result reported as significant. We speculated that if gross inconsistencies in favor of finding significant results as opposed to nonsignificant results would be intentional, authors would be reluctant to share data. We therefore expected that articles with gross inconsistencies, especially those in the direction of statistical significance, would be accompanied by open data less often than articles without any gross inconsistencies.

Interestingly, in the aggregated data of Studies 1, 2, and 3 we found no such pattern (see Table 4.13). Articles without gross inconsistencies shared data in 8.6% of the cases, whereas articles with gross inconsistencies shared data slightly more often: in 10.3% of the cases. We also found that articles with gross inconsistencies in the direction of finding a significant result shared data *more* often (9.9%) than articles with gross inconsistencies in the direction of non-significance (8.3%). This finding is not in line with the notion that authors are more reluctant to share data when their articles contain gross inconsistencies in favor of finding significant results.

We also looked at a special case of gross inconsistencies in favor of significance: *p*-values that were reported as significant, but upon recalculation turned out to be $p = .06$. This case most closely resembles the questionable research practice (QRP) of wrongly rounding down *p*-values as defined in (Agnoli et al., 2017; John et al., 2012). If such cases in our data were indeed the result of intentional QRPs, we would expect articles with such gross inconsistencies to be less likely to share data than articles without gross inconsistencies. Our findings seem to be in line with this notion (see Table 4.13). We found that articles that contained a *p*-value wrongly rounded down from $p = .06$ to $p < .05$ shared data in only 6.6% of the cases, as compared to articles without gross inconsistencies that shared data in 8.6% of the cases. Note that the sample sizes of these subgroup analyses are small, and these results should be interpreted with caution.

[30] Note that this does not add up to the total sample size of 79,784 extracted APA reported NHST results (Study 1: N = 6,482; Study 2: N = 52,376; Study 3: N = 20,926). This is because results reported as $p < .07$ could not be classified as significant or not significant, and these results were not included in this analysis.

Table 4.13

Categorization of all papers from Study 1, 2, and 3 with or without at least one (type of) gross inconsistency and whether they were accompanied by open data.

Articles that contain…			Data Available		% Articles with Data Available
			No	Yes	
No gross inconsistencies			5,473	516	8.6%
At least one gross inconsistency…			573	59	10.3%
	… in a result reported as n.s.		198	18	8.3%
	…in a result reported as sig.		402	44	9.9%
		…in a result where the recomputed *p*-value is .06	99	7	6.6%

4.5 Discussion

We conducted three retrospective observational studies to test the hypotheses that data sharing and data sharing policy are negatively related to statistical reporting inconsistencies. Overall, we found that on average the prevalence of statistical inconsistencies was in line with the estimates of previous research (see Chapter 2 for an overview). In Study 1, on average 9.3% of the *p*-values in an article were inconsistent and 1.1% grossly inconsistent, in Study 2 these numbers were 13.0% and 1.6%, respectively, and in Study 3, 9.3% and 1.1%, respectively. Contrary to what we hypothesized, we did not find consistent evidence that these inconsistencies were related to data sharing or data sharing policies. In Study 2, we did find that the probability of an inconsistency increased less steeply over time in PLOS after they installed a data policy, as compared to FP, that did not install such a policy. However, we did not find a similar pattern for gross inconsistencies, or for the other journals in Studies 1 and 3. Although we considered meta-analyzing the findings of our three studies, we decided not to, for two reasons. First, the results of three studies do not consistently point to a positive or negative effect. Second and most importantly, the three contexts are very different, which questions the use of combining them in one meta-analysis. Note that a random-effects meta-analysis with just three studies is generally also considered not to be very useful.

We ran several exploratory analyses and found some interesting results. First and foremost, we found that installing an open data policy seems to be highly effective: the

proportion of articles with open data increased rapidly after the journals started requiring or recommending open data, as compared to the prevalence of open data in journals without an open data policy over time. This is in line with previous research that shows evidence that journal policy can encourage desirable change in research practices (Giofrè et al., 2017; Kidwell et al., 2016). Even though these results seem promising, they should be interpreted with care. These findings are not based on experimental data but on observational data, which only allow for correlational conclusions.

Even though data availability increased after open data policies were introduced, we did find that a surprisingly high number of cases in which an article stated the data were available, whereas in fact they were not. We found that roughly one third of the articles in PLOS and FP that promised open data did not deliver. This is comparable to the findings of Chambers (2017, p. 86), and Kidwell et al. (2016). Kidwell et al. (2016) showed that of the articles from journals without badges that promised open data, only 40.5% actually had data available. Kidwell et al. also found that articles in Psychological Science with an Open Data Badge had a much higher probability of the data being available, usable, and complete. These data suggest that even though installing an open data policy increases the availability of open data, there needs to be an extra check at the journal to verify if open data statements are justified.

Finally, contrary to the findings in Chapter 2, we found that gross inconsistencies in this sample do not seem to be biased towards finding significant results. Furthermore, we found no evidence that articles with gross inconsistencies were less likely to have open data than articles without gross inconsistencies. Interestingly, we did find that articles were less likely to share data, when it contained a gross inconsistency in which a recalculated *p*-value of .06 was reported as < .05. This finding could indicate that some of the gross inconsistencies are intentionally wrongly rounded down *p*-values, which would lead to reluctance in sharing data. However, these findings are exploratory and based on a relatively small sample, so they should be interpreted with caution.

We recognize three main limitations in our studies. The first limitation is that our choice of retrospective observational designs limits the internal validity of the three studies, and prevents us from drawing causal conclusions. Because we did not randomly assign manuscripts to an “open data condition” and a control condition while keeping everything else constant, we were by definition not able to rule out alternative explanations for any relation between open data and reporting inconsistencies.

A second limitation is the lack of statistical power. Even though we downloaded a considerable number of articles for each study, the relatively low prevalence of inconsistencies dramatically decreases power to detect small effects. That said, we ran several power analyses that showed that if data sharing had a reasonable effect on the prevalence of

inconsistencies, we should have had enough power to detect that. This means that even if data sharing or data sharing policy decreases inconsistencies, the effect is probably not strong enough to be of much practical value. However, the situation was more problematic for detecting any effects on the prevalence of gross inconsistencies. Our power analyses in Study 3 revealed serious shortcomings of multilevel analysis to analyze low incidence rates (as with gross inconsistencies) when based on a small number of observations per level-2 unit (article, in our case). More specifically, in our power analysis we used the baseline probability for gross inconsistencies as found in previous research (1.2% in PS; Chapter 2), and found that in this case the Type I error does not equal .05 but approaches zero instead, and the power to detect extremely large effects may not even exceed .05. This problem holds for Studies 1, 2, and 3, and consequently we do not put too much trust in the results of the multilevel logistic analyses concerning gross inconsistencies. We decided to still include them in this chapter for the sake of completeness and because we preregistered these analyses. More generally, we recommend against using multilevel logistic regression analyses as a statistical method to analyze nested data characterized by a low incidence rate (e.g., less than 5%) in combination with level-2 units having few observations (e.g., eight observations per level-2 unit).

The third main limitation is that we used automated software to detect reporting inconsistencies. Even though statcheck was extensively validated (see Chapters 2 and 3), it will never be as accurate as a manual search. The main problem is that statcheck does not find all statistical results in a paper, due to variations in reporting style or problems in recognizing characters because of a journal's copy-editing process. It is possible that there is a systematic difference in the inconsistency rate between results that were or were not recognized by statcheck. For instance, maybe if researchers make an effort to report their results in APA style (which statcheck can detect), there is a lower probability of making a typo as compared to researchers who do not attempt to adhere to a strict reporting style. However, in statcheck's validity study there was no evidence for a systematic difference in reporting inconsistencies between results that were and were not picked up by statcheck, so we have no reason to assume that statcheck's estimates of the prevalence of inconsistencies is biased.

Taking the limitations into account, the results from these three studies are evidence against our hypotheses that data sharing and data sharing policies lead to fewer statistical reporting inconsistencies. We theorized that the precision needed to archive data in such a way that it is accessible and usable to others would also make typos and other errors in statistical reporting less likely. Additionally, we theorized that authors who are unsure about the quality of their analysis or know that there are errors in their work would be more reluctant to submit their work to a journal that requires data sharing. However, our data suggest that this is not the case; requiring data sharing in itself might not be enough to decrease the prevalence of statistical reporting inconsistencies in psychology.

Our findings are not directly in line with Wicherts et al. (2011), who found that reluctance to share data was related to, among other things, an increased rate of reporting inconsistencies. A meaningful difference between our studies is that we looked at whether data sets were published alongside the articles, whereas Wicherts et al. looked at (reluctance in) data sharing when explicitly requested. However, our findings are in line with those of Veldkamp et al. (2014) and Veldkamp, Hartgerink, Van Assen, and Wicherts (2017), who did not find support for their suggested "co-pilot' model in which they theorize that if multiple authors work on the analyses, the probability for reporting inconsistencies should decrease. Their rationale was that shared responsibility for the analysis and results section should (partly) eliminate human error and therefore increase accuracy of the reported results. However, they did not find a relation between co-piloting and the prevalence of statistical reporting inconsistencies. The combined evidence of our three studies and previous literature seems to point to the conclusion that strategies to increase more rigorous data management such as sharing data and collaborating on analyses is not enough to prevent statistical reporting inconsistencies. Even though this collection of findings is based on a limited set of journals, we see no immediate reason to expect differences in other journals. To find out which strategies could be effective in preventing statistical reporting inconsistencies, we need more research to investigate what causes them.

One way to help decreasing reporting inconsistencies is to use programs and apps such as statcheck (Epskamp & Nuijten, 2014; http://statcheck.io), or p-checker (Schönbrodt, 2015; http://shinyapps.org/apps/p-checker/) to quickly and easily check results for internal consistency. These programs can be used by authors themselves before submitting a paper in order to avoid mistakes in the published paper and having to file a correction. Similarly, journals themselves can also include these extra checks during peer review. The journal Psychological Science started using statcheck in their peer review process last year to prevent inconsistencies from ending up in the literature (http://www.psychologicalscience.org/publications/psychological_science/ps-submissions; retrieved on June 1, 2017), and the use of statcheck is recommended by the journals Stress & Health (Barber, 2017) and the new journal Advances in Methods and Practices in Psychological Science (http://www.psychologicalscience.org/publications/ampps/ampps-submission-guidelines; retrieved on June 1, 2017). Another solution to decrease the prevalence of reporting errors is to make use of Analytic Review (AR; Sakaluk et al., 2014), in which reviewers also check the analysis scripts and accompanying data files. The advantage of AR over automated programs is that a (human) reviewer can also check if the reported statistical analyses were the appropriate ones.

Even though we found no evidence that (recommended) data sharing is related to a decreased prevalence of statistical reporting inconsistencies, we still want to emphasize the importance of open data. Some of the greatest advantages of sharing data include, but are

not limited to, the possibility to run secondary analyses to answer new questions, verify analyses of published work or examine the robustness of the original analyses, and compute specific effect sizes for meta-analyses (see Wicherts, 2013). Stating that "data are available upon request", as is APA policy, is often not enough to ensure availability (Vanpaemel et al., 2015; Wicherts et al., 2006). On top of that, sharing data upon request is not robust to time: how likely is it that the data are actually still available after ten years? Or fifty? Or even longer? Vines et al. (2014) found that the odds of data actually being available upon request dropped by 17% per year. To ensure availability over time it is necessary to publish data in online repositories. An example of a platform for doing so is the Open Science Framework (http://osf.io). Availability of raw data does not guarantee usability or completeness, so it is desirable to build in checks or review of data sets. For instance, it is possible to publish your data in the Journal of Open Psychology Data, in which your data is reviewed to see if it is archived well. There have been concerns about data sharing pointing at issues such as privacy (Finkel et al., 2015), or the risk that "free riders" will take advantage of your painstakingly collected data (but see Longo & Drazen, 2016). These are valid concerns, but in most cases, it is easy to come up with solutions tailored to the situation. For instance, the majority of experiments in psychology do not concern sensitive data and can easily be anonymized, and there are options to publish data online privately, and only make it public after a pre-specified period of time in order to first publish findings from these data yourself. Moreover, there is evidence that data sharing is associated with an increased citation rate (Piwowar, Day, & Fridsma, 2007).

In this chapter, we used empirical methods to investigate one possible solution to the high prevalence of inconsistently reported statistical results. Reporting inconsistencies are only a small part of the problems related to the current "replication crisis" that psychology is facing (for an overview of these problems, see, e.g., Shrout & Rodgers, 2017). Even so, we think that it is useful to treat problems in our scientific system (no matter how small) as empirical questions that we can solve by applying the scientific method. Research that aims to do, such as this dissertation, adds to a growing body of literature on "meta-science" (Ioannidis et al., 2015; Munafò et al., 2017). Improving the quality of our research is a complex endeavor and we will need much more research to understand where the biggest problems lie, what caused them, and how we can solve them. Even though we still have a long way to go, it is encouraging to see that journal policies and research practices are changing to accommodate open science.

Chapter 5

Preventing Statistical Errors in Scientific Journals

This chapter is published as Nuijten, M. B. (2016). Preventing statistical errors in scientific journals. *European Science Editing*, *42(1)*, 8-10.

Abstract

There is evidence for a high prevalence of statistical reporting errors in psychology and other scientific fields. These errors display a systematic preference for statistically significant results, distorting the scientific literature. There are several possible causes for this systematic error prevalence, with publication bias as the most prominent one. Journal editors could play an important role in preventing statistical errors in the published literature. Concrete solutions entail encouraging sharing data and preregistration, and using the automated procedure "statcheck" to check manuscripts for errors.

In Chapter 2, we documented the prevalence of statistical reporting inconsistencies in more than 250,000 *p*-values from eight major psychology journals, using the new R package "statcheck" (Epskamp & Nuijten, 2015). The program *statcheck*: converts PDF and HTML articles to plain text files; extracts results of null hypothesis significance tests that are reported exactly according to APA style (American Psychological Association, 2010); recomputes the *p*-value based on its accompanying test statistic and degrees of freedom, and checks if the reported *p*-value matches the recomputed *p*-value, taking rounding of the reported test statistic into account. We found that in half of the papers at least one *p*-value was inconsistent with the test statistic and degrees of freedom. In most of these cases, the reported *p*-value was only marginally different from the recomputed *p*-value. However, we also found that one in eight papers (12.5%) contained gross inconsistencies that may have affected the statistical conclusions: in those cases, the reported *p*-value was significant, but the recomputed *p*-value was not, or vice versa. We found a higher prevalence of gross inconsistencies in *p*-values reported as significant, than *p*-values reported as nonsignificant, implying a systematic bias towards statistically significant findings.

This high prevalence of statistical errors in psychology papers is alarming, and there is evidence that this problem is not unique for psychology. Similar inconsistency rates have been found in, for instance, the medical sciences in general (Garcia-Berthou & Alcaraz, 2004) and psychiatry in particular (Berle & Starcevic, 2007). Even though small reporting errors might be inconsequential, wrongly reporting a *p*-value of .37 as .36 will probably not have serious effects, the apparent focus on significant results is worrying and can have far-reaching consequences. It may have added to the excess of (false) positive findings in science (Fanelli, 2010; Francis, 2014). There are several explanations for this high error prevalence. First, most of the inconsistencies could have been caused by mere sloppiness. Especially in psychology this is easy to imagine, since a single psychology paper on average already contains about ten statistical tests (Chapter 2). In the tangle of statistical output, it is imaginable that a *p*-value (or test statistic or degree of freedom) is copied incorrectly. Matters probably become worse because many researchers are not in the habit of double checking their own or their co-authors' analyses (who sometimes do not even have access to the raw data in the first place; Veldkamp et al., 2014). However, sloppiness alone does not explain the apparent systematic preference for significant findings.

A possible explanation for the excess of *p*-values wrongly reported as significant is publication bias: significant results have a higher probability to be published than nonsignificant results (Greenwald, 1975; Sterling, 1959; Sterling et al., 1995). It is imaginable that researchers just as often wrongly report a significant *p*-value as a nonsignificant *p*-value. However, because of publication bias, only the gross inconsistencies that wrongly present a *p*-value as significant are published, resulting in a systematic bias in favor of significant findings. Conversely, it is also possible that researchers *suspect* that their findings will not be

published if they do not find a significant effect, and because of this, they more often wrongly round down a nonsignificant *p*-value to obtain a significant finding, than vice versa. This would be in line with the finding of John et al. (2012), who found that 22% of a sample of over 2000 psychologists admitted to knowingly having rounded down a *p*-value to obtain significance, which would lead to an excess of false positive findings. Of course, it could also just be the case that researchers unknowingly maintain double standards concerning the checking of their results: they would inspect their results with more scrutiny when the result is unexpectedly nonsignificant, but not when it is significant.

I believe journal editors can play an important role in preventing, detecting, and/or correcting statistical errors in scientific literature. There are several concrete steps that could be taken to actively improve the state of the published literature.

A possible solution to the problem of statistical reporting errors is to promote data sharing. In previous research it has been found that if researchers were unwilling to share data of a certain paper, there was a higher probability that the paper contained reporting errors, often concerning statistical significance (Wicherts et al., 2011). This finding could illustrate that authors are aware of the inconsistencies in their paper and refuse to share their data out of fear to be exposed. An alternative explanation for this finding is that researchers who manage their data with more rigor both make fewer mistakes and archive their data better, which makes data sharing easier. In both cases the prevalence of reporting errors might decrease when journal editors would encourage data sharing.

Besides the possibility that authors themselves may become more precise in reporting their results if they have to share their data, encouraging data sharing has more benefits. If authors would submit their data and analysis scripts alongside their manuscript, it would allow for so-called analytic review (Sakaluk et al., 2014). In analytic review, peer reviewers or statistical experts verify if the reported analyses and results are in line with the provided data and syntax. Not only will this encourage authors to manage their data more carefully in order for a third party to understand it, statistical errors that were overlooked at first have a higher probability of being detected before publication.

Editors could decide to make data sharing mandatory, taking into account certain exceptions concerning privacy etc. (see, e.g., the policy of PLOS ONE). Another option is to simply reward authors who share data. For instance, the journal Psychological Science awards badges to papers that are accompanied by open data and also awards badges for open materials and preregistered studies. Although at first sight these badges might seem trivial, they can be considered a quality seal and have inspired many researchers to share their data.

Of course, researchers could still conceal deliberate rounding errors towards significance by manipulating the raw data before submitting them. However, falsifying research data like this is explicit scientific misconduct. Data from self-reports show that scientific fraud is much more uncommon than questionable research practices such as

wrongly rounding a *p*-value (John et al., 2012), so it seems implausible that encouraging data sharing will result in researchers hiding rounding errors by manipulating the raw data. In any case, there will always remain ways to commit fraud in science, but encouraging data sharing will definitely make it harder.

Another way to avoid reporting errors and to facilitate analytic review, is for editors of journals that adhere to APA reporting style to make use of statcheck (Epskamp & Nuijten, 2015). As described above, statcheck is a package for the statistical software R (R Core Team, 2014) that can automatically scan articles, extract statistical results reported in APA style, and recompute *p*-values. Editors could make it standard practice to use statcheck to automatically scan papers upon submission to check for statistical reporting inconsistencies. This takes almost no time; on average, statcheck can scan approximately 250 papers per minute. Since many journals already have an automatic plagiarism check, it is a small step of adding a check for reporting inconsistencies. Results that are flagged as problematic can then be corrected before publication. R and statcheck are both open source and freely available. For more information about statcheck and an extensive analysis of its validity, see Chapter 3. For instructions on how to install statcheck, see http://mbnuijten.com/statcheck.

The excess of results wrongly presented as significant is probably caused by publication bias. A promising way for editors to try to avoid publication bias is to encourage preregistration. Preregistration can take many forms, but in general the idea is that researchers write a detailed research (and analysis) plan *before* collecting the data. This research plan is then "registered" somewhere online (e.g., in a repository for clinical trials such as https://www.clinicaltrialsregister.eu), or even submitted to a journal. In the latter case, the research plan is peer reviewed, and if the plan meets the standards of the journal, the researchers can receive an "in principle acceptance", no matter what the results will be – given that they will adhere to the research plan (see, e.g., the guidelines for registered reports in the journals Cortex, Comprehensive Results in Social Psychology, and Perspectives on Psychological Science). This way, the decision to publish a paper cannot be influenced by whether the results were significant or not, avoiding the selective publishing of *p*-values wrongly rounded down as compared to the ones wrongly rounded up. On top of that, it takes away an incentive for researchers to deliberately report a nonsignificant *p*-value as significant.

Besides side-stepping publication bias and avoiding systematic reporting errors, preregistration also solves the problem of HARKing: Hypothesizing After the Results are Known (Kerr, 1998). When researchers are HARKing, they first explore the data to find interesting patterns, and then present these findings as having been predicted from the start. If a researcher performs a lot of exploratory tests, he or she is bound to find at least one significant result purely by chance. Reporting only the tests that were significant leads to an excess of false positive findings. However, if the research plan and hypotheses are registered beforehand, there is a clear distinction between confirmatory and exploratory tests in the

paper, which allows for a more reliable interpretation of the results (Wagenmakers et al., 2012).

To conclude, there is evidence for a high prevalence of statistical reporting inconsistencies in the scientific literature. Even though many of these inconsistencies are minor errors that are probably due to mere sloppiness, there is also a high prevalence of gross inconsistencies that may have affected the statistical conclusion, mainly in favor of statistical significance. Even though we can only speculate why there are more results wrongly presented as significant (deliberately rounding down, publication bias, less rigorous checks of findings in line with expectations, etc.) it remains a worrying finding, reflecting a systematic preference for "success" and leading to an excess of false positive findings in the literature.

There are several concrete steps that journal editors can take in order to avoid or reduce the number of reporting errors. For instance, editors could encourage data sharing and preregistration, or use the program statcheck to automatically check for inconsistencies during the review process. Besides decreasing the prevalence of reporting errors, these measures also reduce publication bias, HARKing, and other questionable research practices.

Statistical reporting errors are not the only problem we are currently facing in science but at least it seems like one that is relatively easy to solve. I believe journal editors can play an important role in achieving change in the system, in order to slowly but steadily decrease statistical errors and improve scientific practice.

Chapter 6

Discussion Part I

Part I of this dissertation focused on statistical reporting inconsistencies in psychology articles. Previous research showed that such reporting inconsistencies, in which the *p*-value did not match the reported test statistic and degrees of freedom, were highly prevalent in psychological research (Bakker & Wicherts, 2011; Caperos & Pardo, 2013). However, these studies were based on relatively small samples. In Part I we therefore set out to document the prevalence of reporting inconsistencies in a large sample of psychology articles, and to investigate possible solutions.

Checking reporting inconsistencies by hand is time-consuming and likely prone to human error. To solve these issues, we developed the R package "statcheck" (Epskamp & Nuijten, 2016), which automatically extracts statistics from articles and recalculates *p*-values to see if they match with the reported test statistic and degrees of freedom. Statcheck can detect results from *t*, *F*, *r*, χ^2, and *Z* tests, but only if they are reported completely (test statistic, degrees of freedom, and *p*-value), and according to APA style (American Psychological Association, 2010). In flagging inconsistencies, statcheck takes into account correct rounding of the test statistic, and it has an automated check for one-tailed tests that are identified as such. By default, statcheck assumes a significance level of .05.

We validated statcheck by comparing its results to the results of a manual coding of the same sample of articles (Wicherts et al., 2011). We found that statcheck detects roughly 60% of all reported NHST results. The results that were not detected were reported either in tables or in a manner that was inconsistent with the APA style. We found that the validity of statcheck in classifying the detected statistics in consistent and inconsistent results was high. The interrater reliability between statcheck and the manual coding was .76 for the inconsistencies and .89 for the gross inconsistencies (Chapter 2). Furthermore, statcheck's sensitivity (true positive rate) was between 85.3% and 100%, and its specificity (true negative rate) was between 96.0% and 100%, respectively, depending on the assumptions and settings. The overall accuracy of statcheck ranged from 96.2% to 99.9% (Chapter 3).

Using statcheck, we investigated the prevalence of inconsistently reported NHST results in 30,717 articles from 8 flagship psychology journals (Chapter 2). Statcheck detected over 250,000 NHST results from over 16,000 articles. We found that roughly half of these articles contained at least one inconsistency, while one in eight articles contained at least one gross inconsistency that concerned significance. At the level of the individual results, we found that, on average, 10.6% of the NHST results in an article were inconsistent, with 1.6% of the results being grossly inconsistent. Gross inconsistencies were more common in *p*-values reported as significant than *p*-values reported as nonsignificant, indicating evidence for a systematic bias in favor of finding "positive" results. We did not find such a systematic bias in a different sample in Chapter 4.

Some concerns about statcheck's validity were voiced by Schmidt (2016), who argued that statcheck falsely flags inconsistencies in results that have been corrected for multiple

testing, post-hoc testing, or possible violations of assumptions. Furthermore, he suspected that our approach to estimate the prevalence of such corrections in the literature as reported in Chapter 2 yielded an underestimate. Based on text searches in our full sample of articles from Chapter 2, we indeed concluded in Chapter 3 that we had previously underestimated the prevalence of corrections because of an error in Windows Explorer. However, we also found that such corrections did not affect our estimate of the prevalence of inconsistencies. Furthermore, we argued in Chapter 3 that there is no reason to report a corrected test statistic in a way that yields the type of inconsistency that statcheck rightly flags as being inconsistent. We also made recommendations on how to correctly report corrected test results, and indicated we see no reason to doubt our initial estimates about the high prevalence of reporting inconsistencies in psychology in Chapter 2.[31] Therefore, we recommend statcheck for use in self-checks, peer review, and research.

To lower the prevalence of reporting inconsistencies, it is important to consider their potential causes. Reporting inconsistencies could be due to random typos. This would decrease the reliability of results, but across the literature it would not cause systematic bias. However, in Chapter 2, we found that gross inconsistencies were more likely to occur in results reported as significant than the other way around (see also Hartgerink et al., 2016). This could be caused by several factors. Perhaps researchers report gross inconsistencies equally often in both directions, but because of publication bias, only the ones reported as significant end up in the literature. Alternatively, researchers could be using a "double standard" in (double) checking their results. If a result is reported as significant, they might be less likely to check its correctness, simply because it is in line with their hypothesis. This way, errors leading to significant results are less likely be corrected than errors leading to nonsignificant results. It is also possible that researchers deliberately wrongly round down *p*-values to obtain significant results. This behavior was classified as a questionable research practice (QRP), and it was found that 22% of the surveyed psychological researchers admitted to this practice (Agnoli et al., 2017; John et al., 2012; but see Fiedler & Schwarz, 2016).

The notion that gross inconsistencies could be the result of QRPs is in line with the findings of Wicherts et al. (2011). In their study, they found that researchers were less likely to share their data if their article contained a gross inconsistency. Again, this can be explained in several ways. It is possible that researchers knew of the inconsistencies in their work and did not want any other flaws to be discovered by sending their data. However, it could also be the case that researchers who are more rigorous in their data management, are more likely to share their data and less likely to commit a reporting error in the first place. In either case,

[31] But see the reply from Schmidt that "statcheck does not work" (Schmidt, 2017), and the summary of this discussion, including our reply in *Science* (Singh Chawla, 2017).

we speculated that if authors (have to) make their data available from the start, they will double-check their results to make sure there are no inconsistencies in their manuscript.

We conducted three retrospective observational studies to further investigate whether data sharing is related to fewer reporting inconsistencies (Chapter 4). In these studies, we compared the inconsistency rates between journals with or without stipulated data sharing policies. We hypothesized that journal policies about data sharing and data sharing itself would reduce inconsistencies. Against our expectations, we found no clear relation between data sharing or data sharing policies and reduced inconsistency rates. This is not in line with the findings of Wicherts et al. (2011), although one noticeable difference between our studies is that we did not explicitly request data from authors, but checked if data were published online. Our findings are in line with previous findings that shared responsibility for data analysis was not clearly related to a decrease reporting inconsistencies (the so-called "co-pilot model"; Veldkamp, Hartgerink, et al., 2017; Veldkamp et al., 2014). Both these findings and our findings in Chapter 4 strengthen our belief that strategies to increase more rigorous data management, such as data sharing policies and co-piloting, are not sufficient to decrease statistical reporting inconsistencies.

We did find, however, that data sharing policies are strongly related to actual data sharing, which might mean that incentivizing data sharing might be effective (see also Giofrè et al., 2017; Harper & Kim, 2017; Kidwell et al., 2016). This is potentially great news, because raw data are notoriously hard to obtain in psychology via personal communications (Vanpaemel et al., 2015; Vines et al., 2014; Wicherts et al., 2006).

6.1 Solutions

Based on Part I of this dissertation, we concluded that statistical reporting inconsistencies are prevalent in the psychological literature. In Chapter 5, I made several recommendations how editors can help in decreasing the prevalence of reporting inconsistencies, but for a large part these recommendations also hold for researchers themselves.

First, I see great potential in using statcheck for self-checks and in peer review. This already seems to be happening. At the time of writing, the R package has been downloaded over 8,000 times, since its publication on CRAN in November 2014 (Epskamp & Nuijten, 2014). The web app at http://statcheck.io has been visited over 20,500 times since its launch in September 2016. Furthermore, statcheck has been incorporated in the peer review process of the two prestigious journals, *Psychological Science* and the *Journal of Experimental Social Psychology*, and its use is recommended by several others, such as the new APS journal *Advances in Methods and Practices in Psychological Science*. Statcheck is also recommended in The Royal Society's research integrity statement (The Royal Society, 2017). Whether the use of statcheck can indeed prevent inconsistencies from being published, is an empirical

question. To test this, a good next step could be to conduct a randomized controlled trial in which one or several journals choose random periods in which they do or do not scan submitted articles with statcheck. Supposedly, reporting inconsistencies should be much lower (or even completely absent) in articles that were reviewed during a statcheck-period.

A second potential solution against reporting inconsistencies is to add "analytic review" to the general peer review process (Sakaluk et al., 2014). Here, researchers are required to submit their data and a syntax file together with their manuscript. This way peer reviewers can rerun the analyses to see if the right analyses were conducted and the results correctly reported. A potential downside of this suggestion is that it increases the burden on peer reviewers. However, this might be a worthwhile investment if analytic review indeed decreases reporting inconsistencies, and flawed statistical decisions in general.

Third, a recommendation specifically for authors is to use programs such as R Markdown (http://rmarkdown.rstudio.com/) to report statistical results. Programs such as R Markdown allow you to directly incorporate your analysis code in your paper, to automatically insert the statistical results into the paper. This should avoid human error in copying results from a statistical program to the manuscript.

In Chapter 5, I also suggested data sharing as a possible way to decrease reporting inconsistencies, based on the findings of Wicherts et al. (2011). However, considering the findings in Chapter 4, and those of Veldkamp et al. (2014) and Veldkamp, Hartgerink, et al. (2017), I do not think data sharing per se will solve the issue. That said, there are still many other problems in psychological science that may very well benefit from sharing data, materials, and analyses scripts (Nuijten, 2017). For instance, the same studies that showed that the wrong rounding of *p*-values was a practice psychologists often admitted to, also found high prevalence of a wide range of other questionable research practices (QRPs; Agnoli et al., 2017; John et al., 2012; but see Fiedler & Schwarz, 2016). Examples of such QRPs include (but are not limited to) failing to report all of a study's dependent measures, conditions, or control variables, and sometimes even failing to report entire experiments that failed to find the desired result. These practices all seem to be focused on one thing: reporting statistically significant results. One potential explanation for this focus on significance is publication bias; the phenomenon that significant findings have a higher probability of being published than nonsignificant findings (Greenwald, 1975). Both QRPs and publication bias cause an excess of significant findings and overestimated effects in the literature, and this becomes worse if the power in studies is low (Bakker et al., 2012; Button et al., 2013; Ioannidis, 2005, 2008). In Part II of this dissertation we will investigate how publication bias can affect replications and meta-analyses, if we can find study-specific factors that might be associated with increased risk for overestimated effects, and if there is reason to suspect publication bias and other biases affected the field of intelligence research.

Part II: Bias in Effect Sizes

Chapter 7

The Replication Paradox: Combining Studies Can Decrease Accuracy of Effect Size Estimates

This chapter is published as Nuijten, M. B., van Assen, M. A. L. M., Veldkamp, C. L. S., Wicherts, J. M. (2015). The replication paradox: combining studies can decrease accuracy of effect size estimates. *Review of General Psychology*, *19(2)*, 172-182.

Abstract

Replication is often viewed as the demarcation between science and non-science. However, contrary to the commonly held view, we show that in the current (selective) publication system replications may increase bias in effect size estimates. Specifically, we examine the effect of replication on bias in estimated population effect size as a function of publication bias and the studies' sample size or power. We analytically show that incorporating the results of published replication studies will in general not lead to less bias in the estimated population effect size. We therefore conclude that mere replication will not solve the problem of overestimation of effect sizes. We will discuss the implications of our findings for interpreting results of published and unpublished studies, and for conducting and interpreting results of meta-analyses. We also discuss solutions for the problem of overestimation of effect sizes, such as discarding and not publishing small studies with low power and implementing practices that completely eliminate publication bias (e.g., study registration).

Imagine that you want to estimate the effect size of a certain treatment. To this end, you search for articles published in scientific journals and you come across two articles that include an estimation of the treatment effect. The two studies can be considered exact replications because the population, designs and procedures of the included studies are identical. The only difference between the two studies concerns their sample size: one study is based on 40 observations (a small study; S), whereas the other study is based on 70 observations (a larger study; L). The following questions are now relevant: How do you evaluate this information? Which effects would you include to get the most accurate estimate of the population effect? Would you evaluate only the small study, only the large study, or both? And what if you would have come across two small or two large studies?

To get an idea about the intuitions researchers have about these questions, we administered a short questionnaire (see Appendix 1) among three groups of subjects, with supposedly different levels of statistical knowledge: second year's psychology students (N=106; paper survey administered during statistics tutorials; Dutch translation), social scientists (N=360; online survey), and quantitative psychologists (N=31; paper survey administered at the 78th Annual Meeting of the Psychometric Society). In the questionnaire we presented different hypothetical situations with combinations of small and large studies, all published in peer-reviewed journals, and asked which situation would yield the most accurate estimate of the effect of the treatment in the population. Accuracy was described in the questionnaire as "the closeness of the estimate to the population effect, inversely related to the bias of an estimate". We list the different situations and responses in Table 7.1.[32]

[32] For more details about the sample and procedure, the original survey, the Dutch translation of the survey, and the full data set, see the Open Science Framework page https://osf.io/973mb/.

Table 7.1

Results of the questionnaire to assess researchers' intuitions about the value of replication. Answers of 106 psychology students (PS), 360 social scientists (SS), and 31 quantitative psychologists (QP). S = Small published study with 40 observations; L = Large published study with 70 observations.

	"Which situation (A or B) yields the most accurate estimate of the effect of the treatment in the population?"		**Proportion of subsample that endorses the answer category**								
	Situation A	Situation B	Situation A more accurate			Situation B more accurate			Situation A and B equally accurate		
			PS	SS	QP	PS	SS	QP	PS	SS	QP
Question 1	L*	S	**.972**	**.857**	**.871**	.019	.036	.032	.009	.108	.097
Question 2	L*	L+S	.057	.045	.032	**.925**	**.839**	**.935**	.019	.117	.032
Question 3	L*	S+S	.340	.283	.258	**.566**	**.619**	**.710**	.094	.099	.032
Question 4	L	L+L	.000	.022	.032	**.943**	**.915**	**.935**	.057	.063	.032
Question 5	L+S*	S+S	**.943**	**.816**	**.839**	.038	.045	.032	.019	.139	.129

*The options that were selected most per subsample are printed in bold face. The correct answers (i.e., the scenarios that were shown to be most effective by our calculations) are indicated with a *. There is no * in Question 4, since both situations contain an equal amount of expected bias.*

The three groups showed the same pattern in all five situations: participants preferred to use as much information as possible, i.e., they preferred the situation with the largest total sample size. For instance, the majority (57% of the students, 62% of the social scientists, and 71% of the quantitative psychologists) preferred two small studies (total of 80 observations) over one large study (70 observations; Question 3). Second, most respondents believed that incorporating a small exact replication with a larger study in the evaluation (Question 2) would improve the accuracy of the estimate of the effect (93% of the students, 84% of the social scientists, 94% of the quantitative psychologists). So answers to questions 2 and 3 revealed two intuitions that are widely held among experts, social scientists, and students alike, namely, that (1) the larger the total sample size, the higher the accuracy, and (2) any replication, however small, improves accuracy. However logical these intuitions may appear at first sight, in this chapter we show that both intuitions are false in the current publication system.

In this article we first explain the origin of these intuitions. Second, we show that replications are not science's Holy Grail, because of the 'replication paradox'; the publication of replications by itself does not decrease bias in effect size estimates. We show that this bias depends on sample size, population effect size, and publication bias. Finally, we discuss the implications for replications (and other studies that would be included in a meta-analysis of the effect under investigation) and consider possible solutions to problems associated with the use of multiple underpowered studies in the current publication system.

7.1 Why Do We Want More Observations and More Studies?

Our intuitions are grounded in what we learned in our first statistics courses, namely that: the larger the sample size, the more information, the greater the precision (i.e., the smaller the standard error), and the better the estimate. A replication study can also be viewed as increasing the original sample size. Hence, intuitively, both increasing the number of observations and incorporating a replication study increases the precision and the accuracy of the estimate of the population effect. This line of thought is reflected in the fact that multiple-study papers have increasingly become the norm in major psychology journals (Giner-Sorolla, 2012), although many of these involve conceptual replications rather than direct replications (Pashler & Harris, 2012; see also Makel et al., 2012).

Furthermore, there is also a large and growing literature on the merits of replication studies. For example, replications are said to be able to protect science from fraud and questionable research practices (Crocker & Cooper, 2011) and clarify ambiguous results (Simmons et al., 2011). Replication is called "the gold standard for reliability" and "even if a small number of [independent replications] find the same result, then that result can be relied on" (Frank & Saxe, 2012). Finally, replications are supposed to uncover false positives that are the result of publication bias (Diekmann, 2011; Murayama, Pekrun, & Fiedler, 2013).

However, the above lines of reasoning do not take into account that publication bias may influence dissemination of both replication studies and original studies. We show how publication bias might limit the usefulness of replication studies and show why publication bias leads our intuitions and those of our colleagues astray (see Table 7.1). We first present evidence of the omnipresence of publication bias in science, and show analytically how publication bias affects accuracy of the effect size estimate of a single study. Thereafter, we discuss the implications of our findings for the accuracy of effect size estimates in meta-analyses that include replications.

7.2 Publication Bias and How It Affects Effect Size Estimates

7.2.1 Presence of Publication Bias

Publication bias is the phenomenon that studies with results that are not statistically significant are less likely to be published (Greenwald, 1975). A way to search for publication bias is by looking for an overrepresentation of statistically significant or "positive" findings given the typical power of the studies (Ioannidis & Trikalinos, 2007). If there was no publication bias, and all effects were truly non-null (further called "true effects" or "existing effects"), then the proportion of positive findings in the literature would be approximately equal to the average power (the probability that you reject the null hypothesis when it is false). Although the recommended power for a study is at least .80 (e.g., Cohen, 1988), the median power has been estimated to average around .35 across studies in psychology (Bakker et al., 2012)[33], the average power is .40-.47 across studies in behavioral ecology (Jennions & Moller, 2003)[34], and .21 across studies in neuroscience (Button et al., 2013)[35]. However, the rate of significant results is 95.1% in psychology and psychiatry, and 85% in neuroscience and behavior (Fanelli, 2010). These numbers are incompatible with the average power across studies in the respective fields and represent strong evidence for publication bias in these fields.

An excess of significant findings has been established in many fields (Bakker et al., 2012; Button et al., 2013; Fanelli, 2012; Francis, 2014; Ioannidis, 2011; Kavvoura et al., 2008; Renkewitz, Fuchs, & Fiedler, 2011; Tsilidis, Papatheodorou, Evangelou, & Ioannidis, 2012). The

[33] Estimated given a two independent samples comparison, assuming an effect size of d = .50 (based on estimates from meta-analyses) and a total sample size of 40, the median total sample size in psychology (Marszalek, Barber, Kohlhart, & Holmes, 2011).

[34] Based on 697 papers from 10 behavioral journals, assuming a medium effect size of r = .30. The authors report the estimated power for a small (r = .1), medium (r = .30), or large (r = .50) effect size. We report the power based on r = .30, because it is closest to the average estimated effect size in ecological or evolutionary studies of r = .18 - .19 (based on 44 meta-analyses; Jennions & Moller, 2002). The average power we report here is therefore likely to be an optimistic estimate.

[35] Based on data from 49 meta-analyses, using the estimated effect sizes in the meta-analyses as true effect sizes.

rate of positive findings seems to be higher in the "softer" sciences, such as psychology, than in "harder" sciences, such as space sciences (Fanelli, 2010). There is evidence that the rate of positive findings has stayed approximately the same from the 1950's (97.3% in psychology; Sterling, 1959) until the 1990s (95.6% in psychology and 85.4% in medical sciences; Sterling et al., 1995), and that it even has increased since the 1990s (Fanelli, 2012).

Several studies have combined the results of tests of publication bias tests from multiple meta-analyses from various scientific fields and found evidence for publication bias in these fields. For instance, there is evidence for publication bias in about 10% of the meta-analyses in the field of genetic associations (Ioannidis, 2011), in roughly 15% of the meta-analyses in psychotherapy (Niemeyer et al., 2012, 2013), in 20% to 40% of psychological meta-analyses (Ferguson & Brannick, 2012), in about 25%-50% of meta-analyses in the medical sciences (Sterne et al., 2000; Sutton et al., 2000), in 38%-50% of meta-analyses in ecology and evolution (Jennions & Moller, 2002), and in about 80% of meta-analyses in the field of communication sciences (Levine, Asada, & Carpenter, 2009). Although percentages of meta-analyses that are subject to publication bias do not seem to be impressively high, the power of publication bias tests was generally low in these meta-analyses. Hence, a failure to detect evidence for publication bias does not necessarily mean that there is no publication bias. A recent study established funnel plot asymmetry as a sign of publication bias in 82 meta-analyses (Fanelli & Ioannidis, 2013; see also Chapter 8).

Both the high prevalence of positive findings and the tests for publication bias in meta-analyses are not conclusive (but see Cooper et al., 1997; Franco et al., 2014 for direct evidence of bias in psychology and the social sciences), but together they make a strong case for a presence of publication bias in much of the scientific literature. Therefore, it is important to investigate how studies are affected by publication bias.

7.2.2 The Effect of Publication Bias on an Estimate from a Single Study.

We analytically derived the effect of publication bias on the effect size estimate in a published study with a two-independent samples design (see also Button et al., 2013; Gerber, Green, & Nickerson, 2001; Kraemer et al., 1998). We used several scenarios differing in the degree of publication bias, the samples sizes, and the underlying effect size. Effect sizes were expressed in Cohen's *d*, or the standardized mean difference (i.e., $d = (\mu_1 - \mu_2)/\sigma$), with $\sigma = 1$). In each scenario we tested H_0: $d = 0$ against H_1: $d > 0$ using a *z* test. We also derived the effect of publication bias in the case where σ is unknown, using a *t*-test. Because the results of the two analyses are very similar, we only report those of the simpler *z* test. The equations and results for the *t*-test can be found at the Open Science Framework page https://osf.io/rumwi/.

We assumed that all significant results were published ($\alpha = .05$) and that there was one underlying effect. Two additional parameters were sample size *N,* and *pub*, representing the proportion of nonsignificant results published. We assumed that all nonsignificant *p*-values

had the same probability of being published. Our assumptions on the probability of publication can also be interpreted differently, i.e., with *pub* as the probability of publication of a nonsignificant study *relative* to the probability of publication of a significant study, where the latter probability can be smaller than 1. We were interested in the bias in the effect size estimate as a function of *d*, *pub*, and *N*. Figure 7.1 shows a variant of the typical depiction of power (used in most statistics textbooks) in which we display the effect of publication bias. Specifically, it shows the effect of *d* and *pub* on the published effect size estimate. In the figure "H0" and "H1" are the regions of accepting and rejecting the null hypothesis, respectively; 1-β represents power, α is the Type I error, cv is the critical value of the *z* test, and *d* is the true population effect size. Without publication bias, available studies are drawn from the sampling distribution underlying *d* (H1). However, because of publication bias, nonsignificant results are less likely published, leading to an asymmetry of reported studies. Specifically, the dark gray area represents the proportion of studies with nonsignificant results that get published. The ratio of the lowered density (dark gray) to the regular density under H1 in the acceptance region equals *pub*, which equals .5 in Figure 7.1.

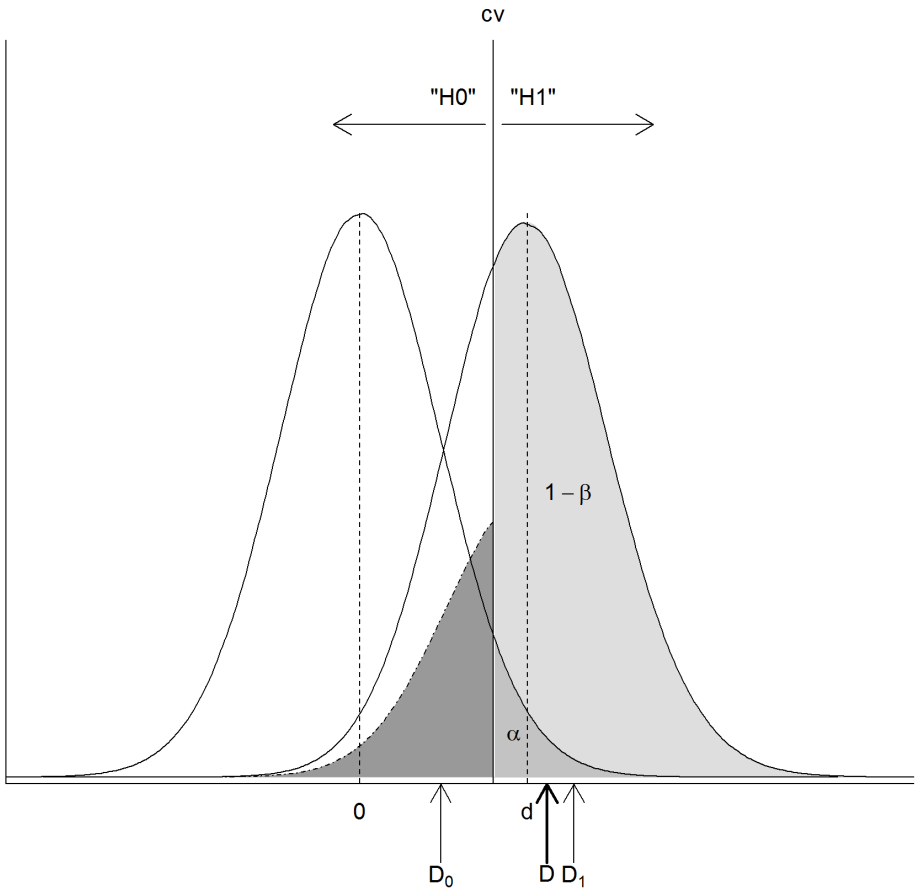


Figure 7.1

Schematic representation of the effect of publication bias on the published effect size estimate. "H0" and "H1" are the regions of accepting and rejecting H0, respectively, 1-β represents power, α is the Type I error, cv is the critical value of the test, and d *is the true effect size.* D_0 *and* D_1 *are the expected effect sizes conditional on the acceptance or rejection of* H_0*, respectively, and D is the expected value of the published effect size.*

To establish the bias in the effect size estimate, we calculated the difference between the actual effect size *d*, and the expected value of the published effect size estimate, *D*. The value of *D* consists of two components. The first component is the expected value of the published effect size given that the effect size was significant, D_1, i.e., the expected value of the light-gray area. The second component is the expected value of the published effect size given that it was nonsignificant, D_0, or the expected value of the dark-gray area. The overall estimate *D* is a weighted average of D_1 and D_0, weighted by the light-gray and dark-gray areas, respectively. The higher the publication bias, the fewer nonsignificant findings are published, and the less weight D_0 will receive. In that case the weighted average will depend more on D_1, and *D* will overestimate *d*, as illustrated in Figure 7.1. If *pub* = 1 (no publication bias), the estimate *D* is equal to the true *d*, and if *pub* = 0 (maximal publication bias), the estimate *D* is equal to D_1, which overestimates *d*. Appendix 2 contains the exact equations.

In our analysis of the effect of publication bias on the accuracy of the effect size estimate in a published study we varied sample size (*N)* to be either 20 or 35 observations per group (40 or 70 observations in total, as in our questionnaire). These sample sizes were chosen to reflect typical sample sizes in psychology (Marszalek et al., 2011; Wetzels et al., 2011). The population effect size, Cohen´s *d*, varied from zero to one. Finally, we chose values of *pub* equal to 0, .05, .25, .5, and 1. Values for *pub* of 0 and 1 reflect the two most extreme scenarios: total publication bias and no publication bias at all, respectively. The value .05 was based on an estimate of publication bias using the number of significant findings in the literature (see Appendix 3). We included the values .25 and .5 to reflect less severe publication bias. The dependent variable of our analysis is the bias in the effect size estimate, which is equal to the expected published effect size minus the true effect. The more bias in the effect size estimate, the less accurate the estimate. So in Figure 7.1, this amounts to the difference between *d* and *D*. Note that whereas this analysis renders the bias of the effect size estimate, the realized estimate will differ across studies and fields.

Figure 7.2 shows the effect of publication bias and population effect size on the bias in the effect size estimate in a single study with either 35 (left) or 20 observations per group (right). In both the large and the small study the same pattern appears. Both scenarios show that if the true effect size is sufficiently large, the bias approximates zero; the effect size estimate as it appears in the literature is equal to the true effect size. The nihil bias arises because for large enough effect sizes nearly all experiments are significant and therefore published. However, if the true effect size becomes smaller, more findings are nonsignificant and are not published. When that happens, bias or the overestimation of the effect generally increases.

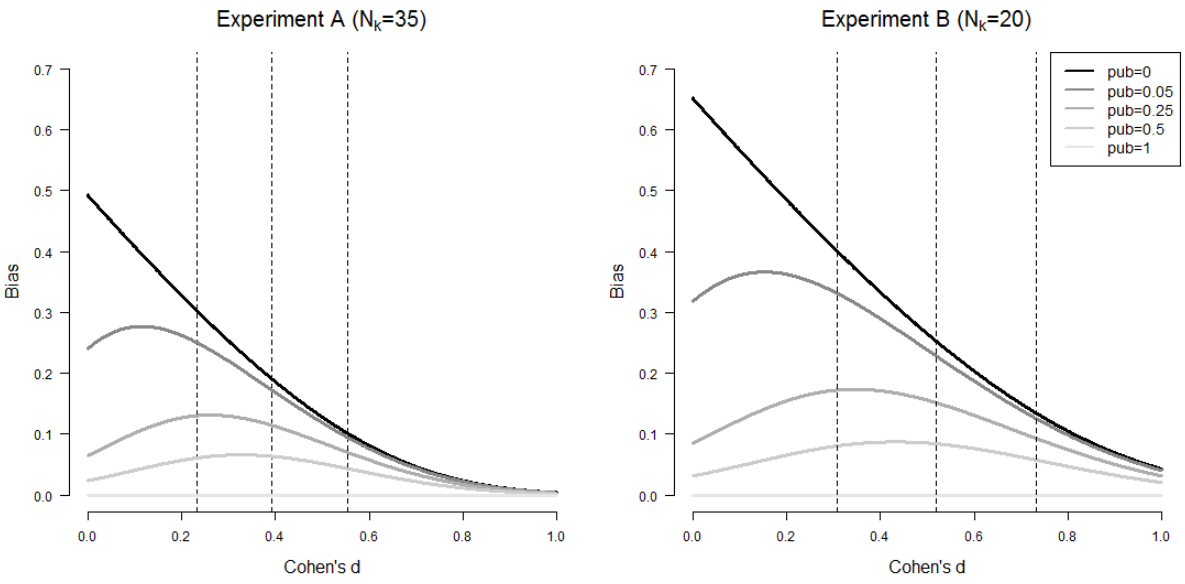


Figure 7.2

The effect of publication bias and population effect size (Cohen's d) on the bias in the effect size estimate in a single study with either 35 or 20 observations per group. The bias in the effect size estimate is equal to the published effect minus the true effect. The vertical, dotted lines indicate Cohen's d at a power of .25, .50, and .75, respectively.

Unsurprisingly, the magnitude of the bias depends on the severity of publication bias. If there is maximum publication bias (none of the nonsignificant results are published), the bias is the largest (black line in Figure 7.2). The bias decreases as more nonsignificant results are published. Without publication bias (results are published independent of their statistical significance), the bias in the effect size estimates disappears completely (lowest, light gray line in Figure 7.2). Formally, the (relative) bias compared to the situation where only significant results are published is a function of both *pub* and power (see Appendix 4 for the derivation of this equation):

$$Relative\ bias = \frac{1 - pub}{1 + pub\frac{\beta}{1-\beta}}$$

Equation 7.1

It follows from Equation 7.1 that bias already decreases dramatically for small values of *pub*, which is also apparent from the sharp drop in bias for *pub*=.05. For instance, consider a case in which *pub*=.05 and *d*=0. It follows that the obtained power is equal to α = .05. In this scenario we obtain a relative bias of (1-.05)/(1+.05*(.95/.05)) = .95/1.95 = .487, meaning that the bias is more than halved compared to the bias when *pub*=0. This is also apparent from Figure 7.2: in both the left and right panel it shows that at *d*=0 the bias in effect size estimate more than halves when *pub* increases from 0 to .05. Now consider a scenario where *pub* = .05

and power is .50 (middle vertical dotted line in Figure 7.2). Here we obtain a relative bias of (1-.05)/(1+.05*(.50/.50)) = .95/1.05 = .905, meaning that the bias is only slightly lower compared to the bias when *pub* = 0. It also follows from Equation 7.1 that relative bias for a certain value of *pub* is only dependent on power. Hence both figures in Figure 7.2 have exactly the same shape. However, absolute bias decreases when sample size increases, hence bias is more severe in the small published study (right figure) than in the large published study (left figure). The difference in bias between the two studies is greatest when publication bias is maximal, and diminishes as publication bias decreases.

Surprisingly, Figure 7.2 shows that bias sometimes first *increases* when population effect size *d* increases. This happens whenever a small proportion of nonsignificant studies is published (*pub*=.05, .25, .5) and power is low. This somewhat counterintuitive result is due to two opposing forces. The first force is the decrease in bias for *pub* = 0 (upper black line); as *d* increases, the average D_1 of the light gray area in Figure 7.1 gets closer to *d*, thereby decreasing bias. The other force is relative bias; if *pub* > 0 and *d* increases, then power increases and relative bias (1) increases. Bias is the product of these two forces (see also Appendix 4). The bump in the figures for *pub* > 0 arises because the increase in relative bias overrules the decrease in bias for the significant studies whenever power is small. In other words, bias increases because the proportion of significant studies, which result in bias, increases more than their bias decreases as *d* increases. For larger values of power, bias decreases monotonically in *d* because then relative bias increases relatively less (see Equation 7.1) than bias for *pub* = 0 decreases.

The results of the analysis of the effect of publication bias and true effect size on the accuracy on effect size estimate when using a *t*-test (when σ is unknown) show that the shape of the figure based on the results of the *t*-test is identical to the shape of Figure 7.2.[36] The difference is that bias is slightly higher for the *t*-test than for the *z*-test, given the same publication bias and true effect size, and this difference decreases in sample size or degrees of freedom of the *t*-test.

An often-proposed solution to the problems of publication bias is to perform multiple studies within an article (see, e.g., Murayama et al., 2013), or to add more replications (see, e.g., Nosek et al., 2012). However, this advice does not take into account that such multiple studies may suffer from the same bias in effect size estimation because of publication bias (Francis, 2012a). In the next paragraph we will therefore extend the known implications of publication bias on a single published study, to the implications of publication bias on scenarios with multiple published studies.

[36] Equations and results for the t test can be found at the Open Science Framework page https://osf.io/rumwi/.

7.3 Implications of Publication Bias on the Accuracy of Multiple Published Studies

In this paragraph we show that replication studies are not necessarily a solution to the problem of overestimated effect size. In fact, we will show that replication can actually *add* bias to an effect size estimate under publication bias. We analytically derived the bias for three possible replication scenarios: two large studies, two small studies, and a large and a small study, and compared the bias in the effect size estimate with the bias in a single large study.

Let A be the original study, and B the replication. If we have two studies, the combined (weighted) effect size estimate *D* equals

$$\frac{N_A D_A + N_B D_B}{N_A + N_B},$$

Equation 7.2

where N_A and N_B represent the sample size, and D_A and D_B the estimated effect size of A and B, respectively. The results for the bias of estimated effect size based on both studies are shown in Figure 7.3.

The left panel of Figure 7.3 shows the bias after combining two large studies (one large study and a large replication). The responses to the questionnaire indicate that most researchers believe that two large studies yield a more accurate estimate of effect size than only one large study. However, the bias of two large studies is exactly the same as the bias in just one large study; because the replication contains the same amount of bias as the original study, the weighted average (see Equation 7.2) of the two effect sizes will also contain the same amount of bias as the original study. Adding a replication to a single study will increase the *precision* or standard error of the estimate, but not its accuracy as long as there is publication bias.

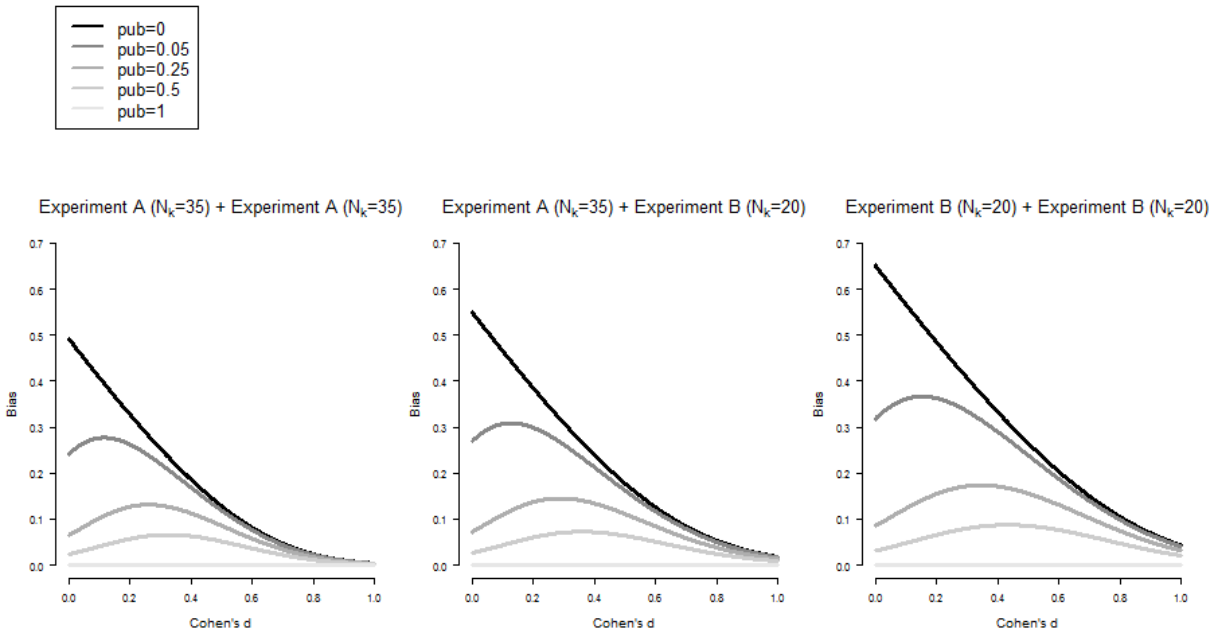


Figure 7.3

The effect of publication bias and population effect size (Cohen's d) on the bias in the effect size estimate in a replication scenario with either two large studies (left panel; identical to the bias in just one large study), one large and one small study (middle panel), or two small studies (right panel; identical to the bias in just one small study).

The middle panel of Figure 7.3 shows the bias in a large study combined with a small replication. According to the responses to the questionnaire, most researchers believe that a combination of one large and one small study yield a more accurate estimate than one large study. Again, this intuition is wrong when there is publication bias. Because a small study contains more bias than a large study, the weighted average (see Equation 7.2) of the effect sizes in a large and a small study is more biased than the estimate in a single large study.

The right panel of Figure 7.3 shows the bias in a combination of two small studies. The responses to the questionnaire indicate that researchers believe a combination of two small published studies yields a more accurate estimate than one large published study. This intuition is not correct. Our analytical results show that the bias in the total effect size estimate does not change if effect size estimates of replication studies of the same size as the original study are synthesized with the effect size estimate of the original study. This means that the comparison between one large and two small studies is equivalent to a comparison between one large and one small study. Hence, the bias is larger in the combination of two small studies than in one large study, even though the sample size of the combination is larger than that of the large study.

In summary, in none of the three replication scenarios did the bias in the effect size estimate decrease by synthesizing the published replication with the large original published study. This means that both intuitions (1) the larger the total sample size, the higher the accuracy, and (2) any replication, however small, improves accuracy, are false when publication bias exists.

7.4 General Implications

Our examples and questionnaire refer to situations in which a published study is combined with a published exact replication. Our analysis shows that synthesizing a published original study with a published replication study generally does not decrease bias in the effect size estimate, yet may even increase bias if the replication study is smaller (in terms of sample size) than the original study. Our analysis has implications for more general situations such as combining effect size estimates of (i) an original study and a *larger* replication study (ii) published conceptual replication studies, (iii) conceptual replication studies within one single published article, (iv) many published studies on the same phenomenon, as in meta-analysis, and (v) for determining whether an effect exists or not.

In the light of recent calls for high-powered replication studies (see, e.g., Brandt et al., 2014), we encounter more and more situations in which the replication study is actually larger than the original study. In those cases, the combined effect size estimate will have less bias than the effect size estimate of just the smaller, original study. Note, however, that in these cases incorporating the smaller original study in the estimation increases bias. Hence, evaluating only the large replication study would provide the most accurate effect size estimate (see also Kraemer et al., 1998).

The conclusion of our analysis holds for any situation in which two or more published effect sizes are combined to obtain an overall effect size (in a meta-analysis), when there is publication bias. This principle generally holds for all sample sizes, and any number of studies. The smaller the study, the larger the bias. So just like combining one small study with one larger study will increase bias in the effect size estimate, combining multiple smaller studies with multiple larger studies will also increase bias, as opposed to combining only large studies.

The same problem applies to situations in which conceptual (published) replications are combined to estimate one underlying (or average) effect size. If both the original study and its conceptual replication estimate the same population effect size and are subject to publication bias, both effect sizes will be inflated, and combining the two studies to obtain a new effect size will result in an overestimation of the population effect size, exactly in the same way as in our analysis. Similarly, the overestimation increases as the studies become smaller.

Multi-study papers are similarly affected by the paradox. Multiple studies within a single paper are also susceptible to publication bias (Francis, 2012b, 2012c, 2013a; Francis,

Tanzman, & Matthews, 2014), which means that an overall effect size based on the effects within one multi-study paper will be inflated as well. Our analysis generalizes straightforwardly to situations in which many published effect size estimates are combined, as in meta-analysis, which are also affected by publication bias (see, e.g., Fanelli & Ioannidis, 2013; Ferguson & Brannick, 2012; Chapter 8). Here, too, overestimation gets worse whenever more small or underpowered published studies are included. What is even more problematic in meta-analysis is that *precision* of the effect size is increased (i.e., standard error of the estimate is decreased) by including more studies, thereby providing a false sense of security in the combined (biased) effect size estimate.

Publication bias also affects analyses used to establish whether an effect exists or not. It has been argued that replication may uncover false positives (e.g., Diekmann, 2011; Open Science Collaboration, 2012; Simmons et al., 2011), but this only holds if studies with nonsignificant results are accessible to researchers (see also Ferguson & Heene, 2012). Similarly, it has been argued that even though multi-study papers can inflate the effect size estimate, they can still decrease the rate of false positives (Murayama et al., 2013). The reasoning is that it is implausible that a research team generates, say, five false positive findings, since on average 5/.05 = 100 studies are needed to obtain five false positives. However, a problem in this argument is that the Type I error is typically much larger than .05, because of the use of so-called questionable research practices (QRP). For instance, Simmons et al. (2011) show that Type I error may even increase to .5 or higher after simultaneous use of some QRPs that are often used by researchers (John et al., 2012; Simmons et al., 2011). Assuming a Type I error of about .5, five positive findings are no longer implausible, since only about ten studies need to be run. Both publication bias and QRP affect effect size estimates of smaller studies more than larger studies (Bakker et al., 2012; Fanelli & Ioannidis, 2013; Chapter 8). This means that even if the goal is not to obtain an overall effect size, but to determine whether an effect exists, multiple underpowered published studies can still distort conclusions.

Does the problem of overestimation of population effect size also hold for unpublished research? We have to distinguish two different types of unpublished studies. First, there are unpublished studies, statistically significant or not, of which the results were subject to biases such as QRP. These biases result in overestimation of population effect size, even when a study's outcome was not statistically significant (Bakker et al., 2012). This implies that incorporating these unpublished studies into meta-analyses may not decrease bias in effect size, particularly if their sample size is similar or smaller to those of published studies. Furthermore, this implication begs the question of the validity of publication bias tests that compare the effects of published and unpublished studies. These tests suggest there is no publication bias if the average effect sizes of published and unpublished studies are similar. Although this publication bias test addresses the effect of publication or not, a nonsignificant

difference between the effects of published and unpublished studies does not imply that the published studies do not yield an overestimated effect size. Ferguson and Brannick (2012, p.126) even concluded that unpublished studies should not be included in meta-analyses, because searches for unpublished studies may be ineffective and unintentionally biased, and these studies may be inherently flawed. The second type of unpublished studies concerns studies that are not affected by biases such as QRP. Incorporating these studies into meta-analysis should generally decrease bias. However, these studies cannot or can hardly be distinguished from those unpublished studies affected by QRP as long as none of these studies are preregistered (see below). Because it is also unknown what proportion of unpublished studies is affected by QRP, it is impossible to tell to what extent unpublished studies yield overestimated effect sizes, both absolutely and relative to published studies.

7.5 Discussion

At the beginning of this article we presented results from a questionnaire that showed that psychology students, social scientists, and experts have the intuition that a published replication, independent of its sample size, improves accuracy of an estimated effect size. We also presented quotes from the published literature suggesting that replications are considered a tool to uncover false positives and to strengthen belief in true positives. We have shown that these intuitions do not hold in a publication system with substantial bias against nonsignificant results. The present system seems to be of this type, although some signs of improvement have recently emerged (e.g., Klein et al., 2014; Open Science Collaboration, 2012). We investigated the effect of replication on the bias in effect size estimate as a function of publication bias, sample size, and population effect size. We found that synthesizing a published original study with a published replication study can even add bias if the replication study's sample size is smaller than that of the original study, but only when there is publication bias. One implication of these findings is that replication studies are not necessarily the ultimate solution to false positives in the literature, as is sometimes implied, but should be evaluated with caution in the current publication system. Our results also hold more generally, i.e., for published conceptual replication studies, conceptual replication studies within one single published article, and many published studies on the same phenomenon, as in meta-analysis.

Our findings are based on the assumption that publication bias affects replication studies in the same way as it affects original studies. However, it is possible that this is not or no longer the case. For instance, publication bias might affect replications even more strongly than it affects original studies. Even though more and more psychologists have started to emphasize the advantages of replication studies, papers containing only one of more replications may still have a low probability of getting published (Giner-Sorolla, 2012; Makel, Plucker, & Hegarty, 2012; Neuliep & Crandall, 1990, 1993). Replications with nonsignificant

results are easily dismissed with the argument that the replication might contain a confound that caused the null finding (Stroebe & Strack, 2014).

On the other hand, it is also possible that publication bias affects replications in the opposite way in some fields. That is, replications could have a *higher* chance of getting published if they contain nonsignificant results while a seminal study contains significant results, because this would be a controversial and thus an interesting finding. In that case, the next study would be controversial again if it were significant. What could follow is an alternation of effect sizes in opposite directions that eventually converge to – possibly – the true effect size. This is known as the Proteus phenomenon (Ioannidis & Trikalinos, 2005). If the Proteus phenomenon holds in practice, biased effect size estimates will cancel each other out over time and the overall effect size estimate will be close to unbiased (De Winter & Happee, 2013). Although the Proteus phenomenon may lead to unbiased effect size estimation, neglecting to publish studies with nonsignificant results is a very inefficient scientific enterprise with problems for statistical modeling of effect sizes (Van Assen, Van Aert, Nuijten, & Wicherts, 2014b, 2014c). Furthermore, even though there are occurrences of the Proteus phenomenon in some fields (Ioannidis, 2011), in psychology the vast majority of studies test if an effect is significantly different from zero, rather than if an effect is significantly different from a previously estimated effect (Fanelli, 2010, 2012; Van Assen, Van Aert, et al., 2014b).

Our analysis also assumes that there are no QRPs that affect the estimated effect size. Considering the seemingly widespread prevalence of QRPs (see, e.g., John et al., 2012), this might not be a realistic assumption. QRPs will likely also result in overestimation of effect sizes. Direct or close replication studies have generally less room for QRPs, since design, procedure, and measures are fixed by the original study. Hence less overestimation of effect size because of QRPs can be expected in direct replication studies. We must stress, however, that there exist only few studies of the effects of QRPs on effect size estimation, alone or in combination with publication bias (but see Bakker et al., 2012). Problematic is that QRPs are not well-defined and most likely have diverse effects on effect size estimation (cf. Lakens, 2015).

There are several potential solutions to the problem of overestimation of effect sizes. The first solution is to only evaluate studies (and replications) with high precision or sample size (Stanley, Jarrell, & Doucouliagos, 2010) or, equivalently, high power. As our results showed, studies with high power will contain less bias in their effect size (see also Bakker et al., 2012; Button et al., 2013; Ioannidis, 2008; Kraemer et al., 1998). A related strategy is not only to evaluate, but also to conduct studies and replications with high power (Asendorpf et al., 2013; Brandt et al., 2014). Each of the studies with high power has little bias, and combining them will increase the precision of the final estimate. A complication with this solution, however, is that the power calculations cannot be based on the (previously)

published effect size, because that published effect size is likely to be overestimated (see also Tversky & Kahneman, 1971). In order to perform an unbiased power calculation, the published effect size needs to be corrected for publication bias (Perugini, Galucci, & Constantini, 2014; Van Assen, Van Aert, & Wicherts, 2014; Vevea & Hedges, 1995).

A second solution is to eliminate publication bias altogether: without publication bias there is no bias in the effect size estimate. Many researchers have emphasized the importance of eliminating publication bias, and there are many proposals with plans of action. For instance, it has been proposed to split up the review process: reviewers should base their decision to accept or reject an article solely on the introduction and method section to ensure that the decision is independent of the outcome (Chambers, 2013; De Groot, 1956/2014; Newcombe, 1987; Smulders, 2013; Walster & Cleary, 1970). A related method to eliminate publication bias is to evaluate submissions on their methodological rigor and not on their results. There are journals that evaluate all submissions according to these standards (see for instance PLOS ONE), journals with special sections for both "failed and successful" replication attempts (e.g., Journal of Experimental Social Psychology, Journal of Personality and Social Psychology, Psychological Science; Brandt et al., 2014), or websites like Psych File Drawer (http://psychfiledrawer.org) on which researchers can upload replication attempts. Furthermore, there have been large scale, preregistered replication attempts of different psychological experiments (Klein et al., 2014; Open Science Collaboration, 2012; see also Wagenmakers et al., 2012). However, even though these proposals and solutions show a high motivation to eliminate publication bias, finding and implementing the best strategy will take time.

What can we do with studies that are already published, and that most likely were subject to publication bias? Following upon others (e.g., Banks, Kepes, & Banks, 2012), we recommend publication bias analyses on past (as well as future) meta-analytic studies in an attempt to evaluate whether publication bias affected the estimated effect size in a field. Many different procedures exist that test for signs of publication bias (see, e.g., Banks et al., 2012; Rothstein et al., 2005). A weakness of statistical procedures that test for publication bias, such as the rank correlation test (Begg & Mazumdar, 1994), Egger's test (Egger, Davey Smith, Schneider, & Minder, 1997), the trim and fill method (Duval & Tweedie, 2000a, 2000b), or Ioannidis and Trikalinos' test for an excess of significant findings (Ioannidis & Trikalinos, 2007; for an extensive discussion about this test and its usage see, e.g., Ioannidis, 2013; Morey, 2013; Simonshon, 2013; Vandekerckhove, Guan, & Styrcula, 2013), is that their statistical power is usually low for meta-analyses with a typical number of studies. Consequently, when these procedures do not signal publication bias, publication bias may still be present and the meta-analysis' effect size estimate biased. On the other hand, these tests could also signal publication bias whenever there is none (a Type I error). When this happens

in a multi-study paper, the test would falsely imply that the author left out one or more studies, which may have unwarranted harmful consequences for the author.

Another option besides testing for publication bias is estimating an effect size that is robust against publication bias or one that is corrected for it. An often used procedure is the trim and fill method (Duval & Tweedie, 2000a, 2000b). However, the trim and fill method does not perform well with heterogeneous meta-analyses (Moreno et al., 2009; Terrin, Schmid, Lau, & Olkin, 2003) and its performance also depends strongly on assumptions about why studies are missing (Borenstein et al., 2009). Another procedure that can be used to obtain unbiased effect sizes in the presence of publication bias is selection models (Copas, 2013; Hedges & Vevea, 1996, 2005; Vevea, Clements, & Hedges, 1993; Vevea & Hedges, 1995; Vevea & Woods, 2005). Selection models use the estimated or a priori probability that a study with a certain *p*-value is published, to estimate the influence of publication bias and to calculate an adjusted effect size. Selection models can deal with heterogeneous effect sizes (Hedges & Vevea, 2005), but may require many studies (e.g., 100 or more) to perform well (Field & Gillett, 2010). Furthermore, selection models are difficult to implement and depend on sophisticated choices and assumptions (Borenstein et al., 2009). A third procedure is to obtain an unbiased effect size by using only studies with statistically significant effects (Hedges, 1984; Simonsohn et al., 2014; Van Assen, Van Aert, & Wicherts, 2014). Van Assen et al. (2014) show that their procedure, called *p*-uniform, provides unbiased effect size estimates, even with the relatively small number of eight studies in a meta-analysis, when the population effect size is homogenous. *P-uniform* also outperformed the standard fixed-effects meta-analysis, the trim and fill method, and the test of excess significance, when publication bias was present. Although we recognize the merits of all aforementioned procedures for testing and correcting for publication bias, they often lack power and/or require rather strong assumptions we believe these procedures do not provide the ultimate solution to problems resulting of publication bias.

Although we cannot establish the exact influence of publication bias on effect sizes estimates in published scientific articles, evidence suggests that publication bias affects many fields. To solve the problem of overestimated effect sizes, mere replication is not enough. Until there are ways to eliminate publication bias or correct for overestimation because of publication bias, researchers are wise to only incorporate and perform studies with high power, whether they are replications or not.

7.6 Appendix 1: The survey including introduction text

The aim of this research is to examine how researchers value exact replications. More precisely, using five questions we assess your evaluation of **the effect of exact replication on the accuracy of the estimation of a population effect**. Accuracy is the closeness of the estimate to the population effect, and is inversely related to the bias of an estimate.

Introduction to questions: please read carefully

Imagine yourself being in the following situation. You want to estimate the effect of a treatment. To estimate this effect, you carry out a literature search. You only include **articles published in scientific journals** in your search. Additionally, you only include **exact replications** in your search. That is, the population, designs and procedures of the included studies are identical; the only difference between the exact replications may be their sample size. After your search you use the available empirical evidence to estimate the treatment effect in the population.

In the questions below you are asked to compare two situations. Your task in each question is to answer the question **'Which situation yields the most accurate estimate of the effect of the treatment in the population?'**. In both situations the same treatment effect is estimated. Hence, the question can also be formulated as **'Which situation would you prefer when your goal is to obtain an accurate estimate of the effect of the treatment in the population?'**.

A situation either involves one *published scientific article* (that is, no exact replications were found) or two *published scientific articles*. A published article is based on either **40** (**S**mall sample size) or **70** (**L**arge) observations. In the five questions below each situation is summarized by one or two letters. For instance, 'L' indicates that only one article was found with a sample size of 70. And 'L+S' indicates two studies were found that were exact replications of each other, one with 70 and the other with 40 observations.

Instruction for answering the questions

The table below contains both situations A and B of the questions (first columns) and the answers to the questions (last three columns). Answer the question by crossing *precisely one* of the three answering categories. For instance, consider Question 0 in the first row. Question 0 compares situation A and situation B, both with a small sample of 40 participants. The cross in the last column indicates that the respondent believes that both situations yield an equally accurate estimate of the effect of the treatment in the population.

Questions

Which situation (A or B) yields the most accurate estimate of the effect of the treatment in the population?

	Question		Answer		
	Situation A	Situation B	Situation A more accurate	Situation B more accurate	Situation A and B equally accurate
Question 0	S	S			X
Question 1	L	S			
Question 2	L	L+S			
Question 3	L	S+S			
Question 4	L	L+L			
Question 5	L+S	S+S			

S = Small study with 40 observations; L = Large study with 70 observations

Thank you for your participation. Any questions or remarks about this research can be sent to Michèle Nuijten (m.b.nuijten@tilburguniversity.edu).

7.7 Appendix 2: Calculation of the Effect of Publication Bias and True Effect Size on the Accuracy on Effect Size Estimate When Using a *z*-test

The following equations show the influence of the proportion of nonsignificant results published (*pub*) on the accuracy of the effect size estimate in a single study, using a *z*-test comparing the means of two independent samples, with σ = 1 (see also Figure 7.1 for a schematic representation of these equations):

1) What is the critical value *cv* of the test?

$$cv = 1.645 \cdot \sqrt{2/N,}$$

where *N* is the number of observations per group.

2) What is the z-value z_1 of the critical value under the alternative hypothesis?

$$z_1 = (cv - d) \cdot \sqrt{N/2},$$

where *d* is the standardized true mean difference between the groups. The probability that $Z > z_1$ is the power of the test, 1-β.

3) What is the expected value D_1 of the mean difference, conditional on a rejection of H_0?

$$D_1 = \frac{f(z_1)}{(1-\beta) \cdot \sqrt{N/2}} + d,$$

where $f(z_1)$ is the density of the standardized normal distribution at z_1. The formula is based on the fact that the expected value of a truncated standardized normal distribution, truncated at probability *p*, equals $f(z_p)/(1-p)$.

4) What is the expected value D_0 of the mean difference, conditional on acceptation of H_0?

$$D_0 = d - \frac{f(z_1)}{\beta \cdot \sqrt{N/2}}$$

Note that $\beta D_0 + (1-\beta)D_1 = d$, as it should.

5) What is the expected value *D* of the estimate of *d*?

$$D = \frac{pub\ \beta D_0 + (1-\beta)D_1}{pub\ \beta + (1-\beta)}$$

The derivations of our results using a *t*-test comparing the means of two independent samples are presented in an online Appendix at Open Science Framework: https://osf.io/rumwi/.

7.8 Appendix 3: Estimation of the Amount of Publication Bias in the Literature

We can make a rough estimate of the amount of publication bias in the literature based on the number of significant findings in the literature. We used the following equations (Van Assen, Van Aert, & Wicherts, 2014):

$$P(\text{"}H_1\text{"}|\text{published}) = \frac{P(\text{"}H_1\text{"} \cap \text{published})}{P(\text{published})} = \frac{P(\text{"}H_1\text{"} \cap \text{published})}{P(\text{"}H_0\text{"} \cap \text{published}) + P(\text{"}H_1\text{"} \cap \text{published})}$$

$$= \frac{(1-\beta)P(H_1) + \alpha\, P(H_0)}{pub[\beta\, P(H_1) + (1-\alpha)P(H_0)] + (1-\beta)P(H_1) + \alpha\, P(H_0)},$$

where $P(\text{"}H_1\text{"})$ and $P(\text{"}H_0\text{"})$ are the proportion of significant and nonsignificant findings in the literature respectively, $P(H_1)$ and $P(H_0)$ are the proportion of effects that are truly non-null or null, respectively, α represents Type I error, β represents Type II error (and (1-β) represents power). Furthermore, *pub* < 1 represents the relative proportion of nonsignificant findings that are published, i.e. proportions of significant and insignificant findings that get published are assumed to be *q* and × *q*, respectively.

Following Ioannidis (2005), we assume that $P(H_1)$ is .50, which is perhaps an optimistic assumption, considering the exploratory nature of much psychological research. Furthermore, assume a power of .50 and α = .05. If we insert these values into the equation, and we assume that *pub* is .05, we get the following:

$$P(\text{"}H_1\text{"}|\text{published}) = \frac{.5 * .5 + .05 * .5}{.05[.5 * .5 + (1-.05).5] + .5 * .5 + .05 * .5} = .88.$$

This result is in line with the research of Fanelli (2010) who found that between 84% and 91.5% of the papers in social and behavioral sciences report positive results. This would mean that the proportion of nonsignificant findings published lies around .05.

Of course, this estimate of the amount of publication bias depends heavily on our assumptions. For instance, we could also consider a scenario in which α is not the nominal .05, but as high as .50. Simmons et al. (2011) indeed report that the actual α may increase from .05 to .5 when researchers employ several questionable research practices (QRP). When redoing our analysis with α = .5, with assuming these QRP will also boost power from .5 to .9, we obtain 88% reported positive results for *pub* = .32. To conclude, even when assuming scientists heavily use QRP, publication bias is estimated to be substantial.

7.9 Appendix 4: Calculation of Relative Bias in Effect Size Estimate

We can calculate the relative bias in effect size estimate compared to the situation where only significant results are published. Subtracting *d* from $D = \frac{pub\beta D_0 + (1-\beta)D_1}{pub\beta + (1-\beta)}$ yields the bias. Denote the bias for *pub* = 0, which equals$D_1 - d$, by *q*. Note that $D_0 - d = -\frac{1-\beta}{\beta} q$, since *d* is the weighted average of D_0 and D_1, with Type II error and power as weights, respectively. Generally, for *pub* ≥ 0, bias $D - d$ can then be rewritten as

$$\frac{pub\beta D_0 + (1-\beta)D_1}{pub\beta + (1-\beta)} - \mathrm{d} = \frac{-pub(1-\beta)(D_1 - \mathrm{d}) + (1-\beta)(D_1 - \mathrm{d})}{pub\beta + (1-\beta)} = \mathrm{q}\frac{1 - \mathrm{p}ub}{1 + \mathrm{p}ub\frac{\beta}{1-\beta}},$$

where $\frac{1-\mathrm{p}ub}{1+\mathrm{p}ub\frac{\beta}{1-\beta}}$ denotes relative bias. This formula for relative bias also holds for the t-test.

Chapter 8

Standard analyses fail to show that US studies overestimate effect sizes in softer research

This chapter is published as Nuijten, M. B., Van Assen, M. A. L. M., Van Aert, R. C. M., & Wicherts, J. M. (2014). Standard analyses fail to show that US studies overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences*, *111(7)*, E712-E713.

Fanelli and Ioannidis (2013) have recently hypothesized that scientific biases are worsened by the relatively high publication pressures in the United States (US) and by the use of "softer" methodologies in much of the behavioral sciences. They analyzed nearly 1200 studies from 82 meta-analyses and found more extreme effect sizes in studies from the US, and when using soft behavioral (BE) versus less soft biobehavioral (BB) and nonbehavioral (NB) methods. Their results are based on non-standard analyses, with $\sqrt[4]{\left|log_{10}\left(\frac{d_{ij}}{\overline{d_j}}\right)\right|}$ as the dependent variable, where d_{ij} is the effect size (log of the odds ratio) of study *i* in meta-analysis *j*, and $\overline{d_j}$ is the summary effect size of meta-analysis *j*. After obtaining the data from Fanelli, we performed more standard meta-regression analyses on d_{ij} to verify their conclusion that effect sizes and publication bias differ between methods and US vs. other countries. For our analyses we used the R package metafor (Viechtbauer, 2010).

First, we ran 82 mixed-effects meta-analyses:

$$d_{ij} = \alpha^j + \beta^j_{US} US_{ij} + \beta^j_{SE} SE_{ij} + \beta^j_{US.SE} US_{ij} SE_{ij} + \varepsilon_{ij}.$$

Equation 8.1

We multiplied d_{ij} by -1 if the primary researchers expected a negative effect. $US_{ij} = 1$ if the primary study was conducted in the US, and 0 otherwise. SE_{ij} is the study's standard error, where a positive β^j_{SE}signifies publication bias (tantamount to Egger's test (Egger et al., 1997)). Next, we ran two mixed-effects meta-meta-regressions on the 82 $\widehat{\beta^J_{US.SE}}$, both with and without method (NB, BB, or BE) as a moderator. The goal was to examine whether the regression weights from the 82 meta-analyses differed between methods, and whether they deviated from zero when averaged over the three methods.

In the meta-meta-regression, method had no effect on $\widehat{\beta^J_{\mathrm{US.SE}}}$ ($\chi^2_{(2)} = 2.271, p = .32$). The overall effect of $\widehat{\beta^J_{US.SE}}$ in the intercept-only model was also not significant ($-.251; z = -.765, p = .44$), meaning that publication bias was not different for the US and other countries.

Because there was no overall $US_{ij}SE_{ij}$ interaction, we reran the 82 meta-analyses without this interaction, and then again analyzed both $\widehat{\beta^J_{US}}$ and $\widehat{\beta^J_{SE}}$ with meta-meta-regressions. Figure 8.1 shows the distributions of $\widehat{\beta^J_{US}}$ and $\widehat{\beta^J_{SE}}$. There was no effect of method on $\widehat{\beta^J_{US}}$ ($\chi^2_{(2)} = 3.464, p = .18$), and no overall effect of US ($-.006; z = -.176, p = .86$). Hence, contrary to Fanelli and Ioannidis, using standard analyses we found no evidence of higher effect sizes in the US for any of the three methods. There was also no effect of method on $\widehat{\beta^J_{SE}}$ ($\chi^2_{(2)} = 5.060, p = .08$), but the overall positive effect of SE ($.537; z = 3.88, p < .001$) signifies publication bias across all methods.

To conclude, we failed to find that US studies overestimate effect sizes in softer research. It is rather surprising that Fanelli and Ioannidis did find an effect of *US*, because the distribution of $\widehat{\beta_{US}^{J}}$ is almost centered on zero (see Figure 8.1, left panel). We found no effect of US and no effects of 'softness' of methods using standard analyses. However, we found overall publication bias for all methods. Hence, the conclusions of Fanelli and Ioannidis are not robust to method of analysis.

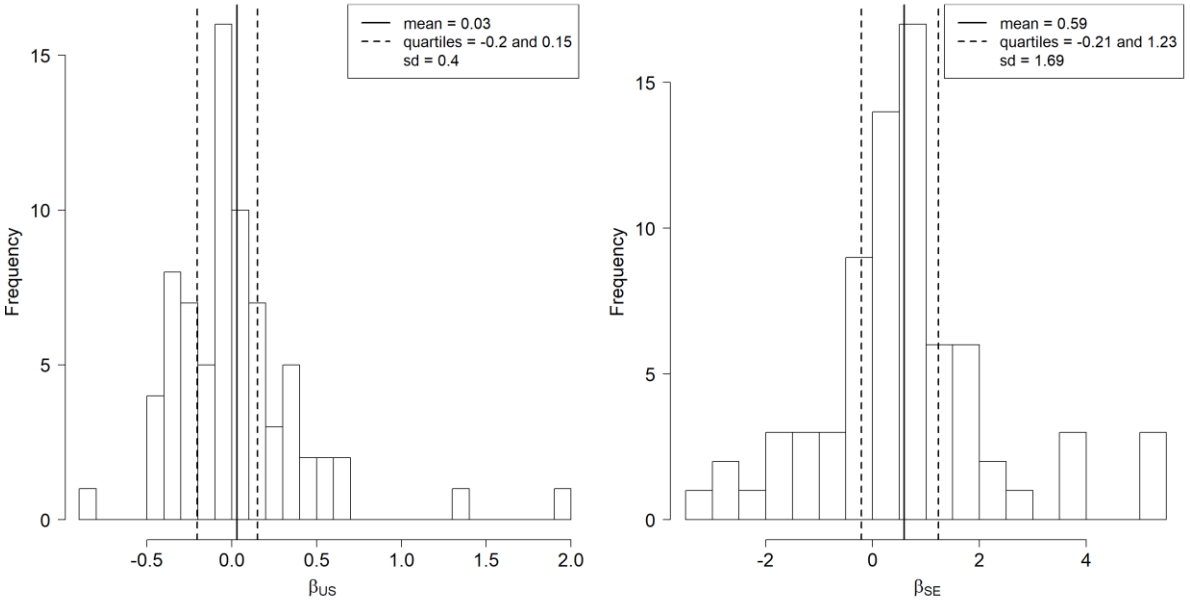


Figure 8.1

Histograms of the effect of US and SE on effect size.

Chapter 9

Effect Sizes, Power, and Biases in Intelligence Research: A Meta-Meta-Analysis

This chapter will be submitted as Nuijten, M. B., Van Assen, M. A. L. M., Augusteijn, H. E. M., Crompvoets, E. A. V., & Wicherts, J. M. (n.d.). Effect sizes, power, and biases in intelligence research: a meta-meta-analysis.

We thank John P. A. Ioannidis and Daniele Fanelli for their comments on earlier versions of this paper.

Abstract

We analyzed 2,439 effect sizes from 131 meta-analyses in intelligence research to estimate the average effect size, median power, and evidence for bias in this field. We found that the typical effect size in this field was a Pearson's correlation of .26, and the median sample size was 60. We calculated the power of each primary study by using the corresponding meta-analytic effect as a proxy for the true effect. The median power across all studies was 48.8%, with only 29.8% of the studies reaching a power of 80% or higher. We documented differences in average effect size and median power between different subfields in intelligence research (correlational studies, studies of group differences, experiments, toxicology, and behavior genetics). Across all meta-analyses, we found evidence for small study effects in meta-analyses, highlighting potential publication bias. The evidence for the small study effect being stronger for studies from the US than for non-US studies (a US effect) was weak at best. We found no clear evidence for the decline effect, early extremes effect, or citation bias across meta-analyses. Even though the power in intelligence research seems to be higher than in other fields of psychology, this field does not seem immune to the problems of replicability as documented in psychology.

Mounting evidence suggests that the literature in psychology and related fields paints an overly positive picture of effects and associations because of a set of biases in how researchers design and conduct studies, and in how they analyze and report research results. Many published findings cannot be replicated in novel samples (Klein et al., 2014; Open Science Collaboration, 2015), many meta-analyses highlight selective reporting of results depending on significance (Button et al., 2013; Fanelli et al., 2017; Niemeyer et al., 2012, 2013), and the number of confirmed hypotheses in the literature is incompatible with the generally low statistical power of psychological studies (Bakker et al., 2012; Fanelli, 2010; Francis, 2014; Marszalek et al., 2011). It is argued that the main cause for this "replicability crisis" (Baker, 2016a) is a combination of publication bias and strategic use of flexibility in data analysis (Ioannidis, 2005; Munafò et al., 2017). Publication bias is the phenomenon where statistically significant results have a higher probability of being published than non-significant results (Greenwald, 1975). Moreover, it is suspected that many researchers try out multiple analysis strategies to search for a significant finding, and only report the ones that "worked" (Bakker et al., 2012; John et al., 2012; Simmons et al., 2011), which increases false positive rates and generally inflates estimates of genuine effects. Because such biases might negatively affect the trustworthiness of published findings, it is important to assess their severity in different bodies of literature. In this chapter, we investigated patterns of bias in the field of intelligence research.

Intelligence research provides a good field to study effect size, power, and biases, because it encompasses a wide range of fields using different methods that still focus on measures of the same construct. Intelligence is among the most well-known constructs in psychology and has been investigated extensively from various angles since the development of the first successful intelligence tests in the early 20th century (Binet, 1905; for reviews, see, e.g., Hunt, 2010; Mackintosh, 2011; Ritchie, 2015). Individual differences in intelligence and cognitive ability tests have been related to many relevant outcomes, correlates, and (potential) causes, in the contexts of education, health, cognitive development and aging, economic outcomes, genes, and toxic substances (e.g., adverse effects of lead or alcohol exposure). Intelligence research is a multidisciplinary field with links to behavior genetics, educational sciences, economics, cognitive psychology, neuroscience, and developmental psychology. These different types of research use different methods and involve different effect sizes, and hence might differ in how strongly they are affected by potential biases (Ioannidis, 2005). For instance, effect sizes are expected to be fairly large in research relating one type of cognitive test (e.g., fluid reasoning tasks) compared to other related cognitive test (e.g., spatial ability tasks), because of the well-established phenomenon of the positive manifold (e.g., Van Der Maas et al., 2006). Conversely, research that attempts to improve intelligence by certain interventions might show smaller effects in light of longstanding challenges in raising intelligence (e.g., Spitz, 1986). Similarly, some research methods in the

study of intelligence are more challenging in terms of data collection (e.g., neuroscientific measures, twin designs in behavior genetics, or controlled interventions) than other research methods (e.g., studies that establish correlations between existing measures in readily accessible samples), thereby creating variation in sample sizes that play a key role in power and (over)estimation of effects and associations.

9.1 Patterns of Bias

One way to investigate bias in science is by analyzing patterns in effect size estimates in meta-analyses (see, e.g., Fanelli et al., 2017; Fanelli & Ioannidis, 2013; Jennions & Moller, 2002). Here, we analyzed 2,439 effect sizes from 131 meta-analyses in intelligence research to estimate the average effect size, median power, and evidence for bias in this field. Specifically, we looked at five types of bias that have been found to be problematic in other research fields (Fanelli et al., 2017): small study effect, US effect, decline effect, early-extremes effect, and citation bias. We discuss these biases in detail below.

9.1.1 Small Study Effect

A small study effect occurs when (published) studies with smaller sample sizes yield larger average effect sizes than those with larger sample sizes (Sterne & Egger, 2005). A small study effect can have several causes. One possible cause of a small study effect is publication bias. Smaller studies generally contain more sampling error, which means that the effect size estimates can vary widely. Effects in smaller studies need to be larger in order to reach significance thresholds than effects in larger studies. If mainly the statistically significant effects are published, small studies with overestimated effects will be overrepresented in the literature. In a meta-analysis, such a small study effect is readily visible by verifying whether the effects in primary studies can be predicted by the studies' precision (typically the standard error).

It is important to note that a small study effect does not necessarily signify bias. For instance, a small study effect can also arise because of true heterogeneity in which underlying effects happen to be related to studies' precision. For instance, in a clinical meta-analysis, the study size may be related to intensity of the intervention, because more strongly afflicted patients are both rare and receive more extensive treatments than less afflicted patients. A small study effect can also arise purely by chance. For an overview of alternative explanations of the small study effect, see Sterne et al. (2011).

9.1.2 US Effect

Studies from the US may have a higher probability of reporting overestimated effects (Fanelli & Ioannidis, 2013). The suggested explanation for this "US effect" is that the publish-or-perish culture is stronger in the US than in other countries (van Dalen & Henkens, 2012), which makes US researchers more inclined to taking advantage of flexibility in data analysis

(Simmons et al., 2011) and selecting only (studies with) significant findings to submit for publication.

Patterns of overestimation in US studies can be analyzed in several ways. One possibility is looking at differences in deviation of effects from the overall meta-analytic effect size in US and non-US studies (Fanelli & Ioannidis, 2013; Munafò, Attwood, & Flint, 2008). Another option is to investigate whether US studies generally find larger effects than non-US studies, controlled for sample size (Fanelli et al., 2017; Fanelli & Ioannidis, 2014). Finally, it is possible to analyze if small study effects are stronger in US studies than non-US studies, potentially indicating stronger publication bias in the US (Doucouliagos, Laroche, & Stanley, 2005; Nuijten, Van Assen, Van Aert, & Wicherts, 2014). Again, note that bias is again only one of the possible explanations of a US effect. It is imaginable that there are other factors at play that cause US studies to report larger effects or display stronger small study effects (Fanelli & Ioannidis, 2013).

9.1.3 Decline Effect

Studies that are published earlier in a research line may be more likely to report larger effects, relative to later studies (Fanelli et al., 2017; Ioannidis, 1998; Ioannidis, Ntzani, Trikalinos, & Contopoulos-Ioannidis, 2001; Song et al., 2010; Stern & Simes, 1997). A decline effect can be caused by decreasing publication bias over time in that field; this is also called time-lag bias (Trikalinos & Ioannidis, 2005). When time-lag bias occurs, manuscripts with significant effects may take less time until completion and publication than manuscripts with non-significant effects (Stern & Simes, 1997).

A decline effect can also occur because of true heterogeneity in which underlying effects happen to be related to when in a research line these effects were examined. For instance, imagine a psychology experiment that makes use of deception to hide the true goal of the study from the participants. When such an experiment is replicated over time, it is imaginable that participants become familiar with the type of experiment and the deception loses its credibility (Schooler, 2011). For an overview of alternative explanations of a decline effect, see Trikalinos and Ioannidis (2005).

9.1.4 Early-Extremes Effect

Alternatively to the decline effect, studies that are published earlier in a research line may be more likely to report more extreme effects in any direction, relative to later studies (Ioannidis & Trikalinos, 2005). One explanation is that early in a research line such extreme findings in opposing directions are deemed controversial and therefore more publishable. This phenomenon seems to be most problematic in fields where findings can follow each other rapidly, for instance in genetic association studies, as opposed to research fields in which data collection takes longer, such as clinical trials (Ioannidis & Trikalinos, 2005).

9.1.5 Citation Bias

Studies with larger effects may be more likely to be cited than studies with small, non-significant effects (Christensen-Szalanski & Beach, 1984; Jannot, Agoritsas, Gayet-Ageron, & Perneger, 2013). Citation bias can cause effects to look more important or undisputed than they really are when taking into consideration all relevant evidence.

We will investigate the five patterns of bias above in different types in intelligence research, because they can differ substantially in terms of research questions, methodology, and ease with which samples can be collected. This may lead to differences in average effect size, power, and bias. As we expected some relevant differences in severity of patterns of bias, we distinguished between five different subtypes of intelligence research in our analyses: correlational, group differences, experiments, toxicology, and behavior genetics. We give more details about these subtypes below.

9.2 Method

9.2.1 Sample

We searched for meta-analyses about IQ and intelligence on the 29th of August, 2014, on ISI Web of Knowledge, using the search string "TOPIC: (IQ OR intelligence) AND TOPIC: (meta-analysis)". This rendered 638 records. From these 638 records, we excluded 6 duplicate ones, and 71 records in which the article was not available in our university library. We then looked at the content of the articles and excluded 186 articles that were not a quantitative meta-analysis, and we excluded 102 articles that were meta-analyses, but not about intelligence or IQ. We operationalized intelligence by including IQ tests and other cognitive maximum performance tests that were featured in Carroll's (1993) seminal review of the intelligence literature.

All effect sizes retrieved from the meta-analyses were based on independent samples both within and between meta-analyses (below we indicate how we avoided overlap between meta-analyses). Because a meta-analysis could report different intelligence test results in the same samples and/or for different types of cognitive tests, we selected effect sizes based on the type of measure used in the primary studies. If the meta-analysis reported results for Full Scale IQ (FSIQs), we included only those FSIQs. If the meta-analysis only reported Verbal IQs or Performance IQs, we selected one of these depending on which set of primary studies was largest. If no IQ measure was presented in the meta-analysis, we chose the largest set that used a cognitive test (or a set of similar cognitive tests) that is strongly associated with the general factor of intelligence (McGrew, 2009). Thus, whenever the meta-analysis lacked IQ measures, we included studies that used the same cognitive test (e.g., the use of Raven's progressive matrices test), or highly similar tests that were labeled in the same manner in the meta-analytic article (e.g., all used fluid reasoning tasks). Because of the positive manifold

between cognitive tests and fairly high correlations between broad cognitive factors (McGrew, 2009), this strategy ensures inclusion of measures bearing on intelligence, while also creating less heterogeneity within meta-analyses that would have been present had we included entirely different (types of) cognitive tests.

One article contained two independent meta-analyses, so we included both. Next, we excluded 139 meta-analyses because they did not contain sufficient data (or no data at all) to calculate the effect sizes and standard errors for the primary studies (102), used non-standard meta-analytic methods (e.g., multi-level models based on individual level data or unweighted analyses; 32), or contained fewer than three unique primary studies (9). Our final sample consisted of 131 meta-analyses, consisting of 2,443 unique primary studies,[37] and over 20 million participants. See Figure 9.1 for a schematic overview of the exclusion criteria and meta-analysis selection. A list of all included meta-analyses can be found online at https://osf.io/dqc2b/.

[37] For four primary studies, we were not able to calculate the effect size, so these were excluded from our analyses.

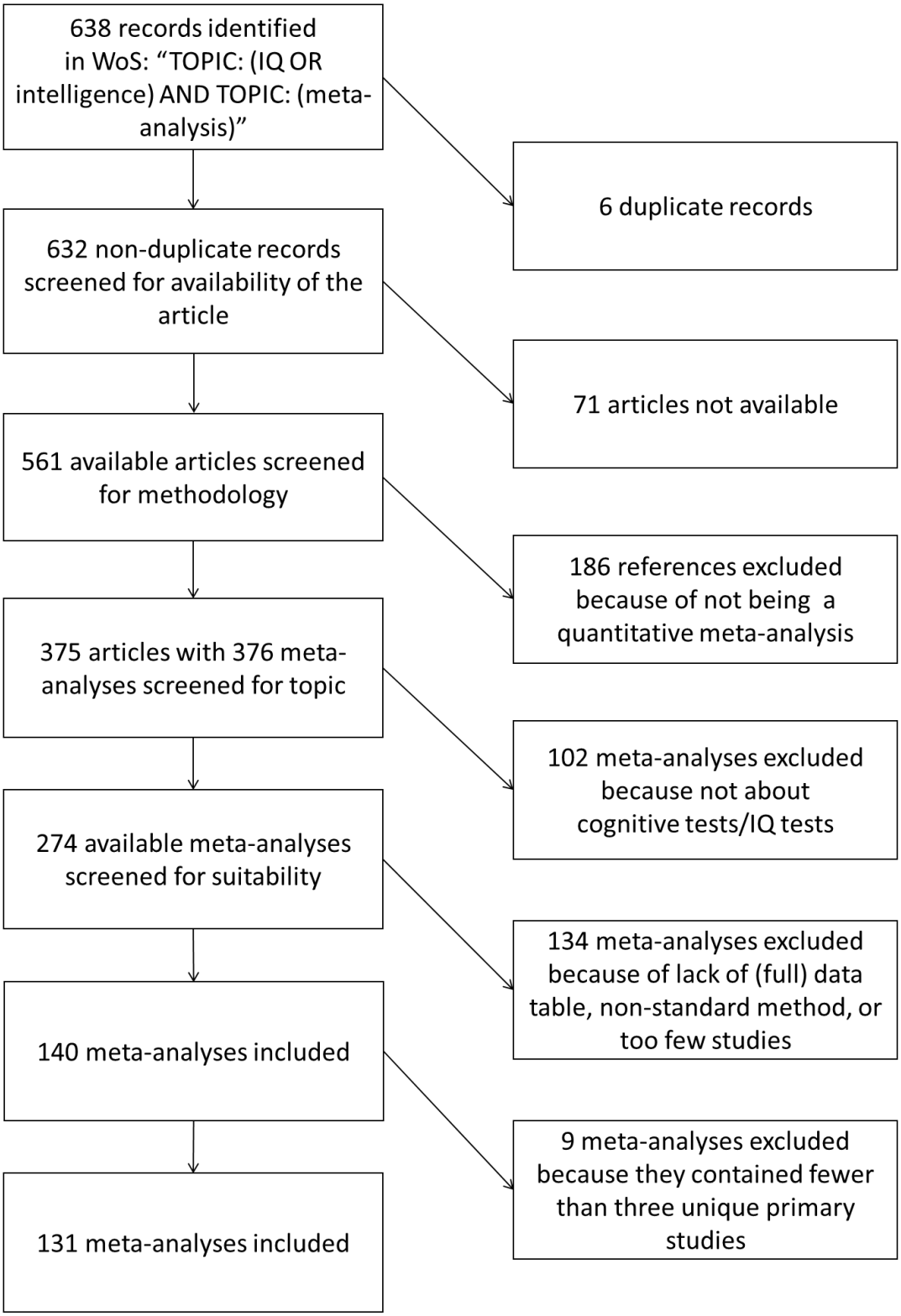


Figure 9.1

Schematic representation of the strategy to search and select meta-analyses to include in our study.

9.2.2 Procedure

9.2.2.1 *Variables*

For each meta-analysis, we coded several variables. Firstly, we coded whether primary studies were unique in our sample, to avoid dependencies between meta-analyses. If a study appeared in more than one meta-analysis, we removed it from the meta-analysis with the most primary studies. This way, we ensured that the number of effect sizes of the individual meta-analyses would remain as large as possible. Furthermore, for each unique primary study, we recorded the effect size that was included in the meta-analysis and its standard error (SE). Often, the meta-analysts calculated the effect size and its SE of a primary study themselves. Analyzing data and reporting results are error prone (see e.g., Bakker & Wicherts, 2011; Gotzsche et al., 2007; Mathes, Klaßen, & Pieper, 2017; Nuijten et al., 2016; Petrocelli, Clarkson, Whitmire, & Moon, 2012). To minimize the risk of copying erroneously calculated or reported effect sizes and SEs, we calculated the effect sizes and SEs ourselves using data reported in the meta-analysis, where possible.[38] Effect sizes could often be calculated with statistics such as means and standard deviations or frequency tables, and we could often calculate the SE using sample sizes or confidence intervals. If there was insufficient information available to calculate the primary studies' effect size and SE, we copied them directly from the meta-analysis. Where possible, we also recorded the primary studies' total sample size, and the sample size per condition. After a first round of data collection, all effect size computations and effect size retrievals from meta-analytic articles were checked by a second coder to avoid errors and to correct any errors that emerged.

Next, for each primary study, we coded several additional variables that were relevant for our analyses. We recorded the year in which the primary study was published, and we coded the relative order in which primary studies were published within a meta-analysis. Here, studies published in the same calendar year were considered as published at the same time. Furthermore, we coded the country in which the corresponding author of the primary study was based when the study was published. Based on this information, we created a binary variable to indicate if a study was from the US (1) or not (0). We also coded the number of citations the primary study received (coded in March 2015). All information about publication year, country, and citations was extracted from Web of Knowledge. We did not code whether a primary study was published in a peer reviewed journal or not.

Finally, we categorized the meta-analyses in five different types of research: correlational, group differences, experiments and interventions, toxicology, and (behavior) genetics (see Table 9.1). Correlational studies refer to studies that lack any manipulation or

[38] In these cases, we did not record the effect size and SE reported by the authors. It would be an interesting additional study to estimate how much reported and recalculated effect sizes and SEs differed, but this is beyond the scope of this study.

treatment in which a measure of intelligence was correlated with another individual difference variable that was measured on a continuous scale. The effect sizes in such studies are typically Pearson's correlations. Examples of such studies include studies relating IQ to personality (Cohn & Westenberg, 2004), brain size (McDaniel, 2005), or self-rated intelligence (Freund & Kasten, 2012). Studies into group differences compare existing (non-manipulated) groups and typically use Cohen's *d* or raw mean IQ differences as the key effect size. Examples include studies comparing mean IQs between males and females (Irwing & Lynn, 2005) or mean IQs between healthy controls and people diagnosed with schizophrenia (Aylward, Walker, & Bettes, 1984). Experiments and interventions are studies that attempt to improve IQ of either healthy or unhealthy groups. Effect sizes are typically standardized mean differences and examples include studies investigating the effect of interventions improving cognitive development in institutionalized children (Bakermans-Kranenburg, van IJzendoorn, & Juffer, 2008), or the effect of iron supplementation on cognition (Falkingham et al., 2010). Studies of toxic effects on IQ entail observational studies or randomized clinical trials in which the toxic effects relate to side effects of a certain drug treatment. Examples include studies investigating potential harmful effects of lead exposure on IQ (Carlisle, Dowling, Siegel, & Alexeeff, 2009), or prenatal cocaine exposure on children's later IQ (Lester, LaGasse, & Seifer, 1998). Finally, behavior genetic studies link intelligence to genetic variations or estimate heritability using twin designs. Examples include studies about the heritability of cognitive abilities (Beaujean, 2005) or studies linking specific genetic variants to general cognitive ability (Zhang, Burdick, Lencz, & Malhotra, 2010).

We chose these five categories to distinguish between substantively different types of research questions and their associated research designs, while retaining a sufficient number of meta-analyses in each type. We ordered the types in increasing complexity of the methodology. Correlational studies and studies about group differences usually do not require special populations, and often make use of convenience samples. In experimental research, the methodology increases in complexity, because participants have to be randomly assigned to carefully constructed conditions. Toxicological studies are mainly correlational (i.e., observational) or quasi-experimental, but require special populations, which makes them logistically much more challenging. Finally, behavior genetic studies are arguably the most complex in methodology, and often require special groups (especially in twin-designs). The five study types were independently coded by MN and JW. The initial interrater reliability was a Cohen's $\kappa = .868$. Any coding discrepancies were solved through discussion by the coders.

Table 9.1

The number of included meta-analyses and primary studies split up in the five different types, reflecting substantive differences in research questions and methodology.

Type of Research	Explanation	# Meta-analyses	# Unique primary studies
1. Predictive validity & correlational studies	(a) Selected IQ test is correlated with other, continuous measurement of psychological construct; (b) test-retest correlation	31	781
2. Group differences (clinical & non-clinical)	Correlation IQ test & categorical, demographical variables or clinical diagnoses (e.g., male/female, schizophrenia y/n)	59	1,249
3. Experiments & interventions	Studies in which participants are randomly assigned to conditions to see if the intervention affects IQ	20	185
4. Toxicology	Studies in which IQ is correlated to exposure to possibly harmful substances	16	169
5. (Behavior) genetics	Genetic analyses & twin designs	5	59

9.2.2.2 ***Effect size conversion***

For our analyses, we converted the effect sizes in all meta-analyses to a single type of effect size. For most of the meta-analyses, the effect size we extracted or calculated based on available data was either a Cohen's *d* (79 meta-analyses; 60.3%) or a correlation (*r*; 42 meta-analyses; 32.1%), so converting to one of these effect sizes seemed most convenient. We chose to convert all effect sizes to *r*, because it makes more conceptual sense to express a *d* in *r* than vice versa. If one expresses a *d* in *r*, the resulting point biserial correlation gives exactly the same information as *d*, but if one expresses an *r* in *d*, the *d* loses information (for more information, see Rosenthal & DiMatteo, 2001). For the meta-analyses of patterns of bias, we subsequently converted all *r*'s to Fisher's *Z* values, because the standard error then only depends on the sample size and not on the correlation itself (see also Sterne, Becker, & Egger, 2005).

The direction in which the meta-analytical hypothesis was formulated can affect whether the primary effect sizes are reported as positive or negative. To correct for any influence of the direction of the hypothesis, we used a procedure called "coining", following Fanelli et al. (2017). In this procedure, we assumed that if the meta-analytic (mean) effect size was negative, the expected direction of the primary effect sizes was also negative. In these cases, we multiplied all primary effect sizes within the meta-analysis by -1. This procedure would be risky if the meta-analyses did not have a specific hypothesis and investigated null effects. If that were to be the case, we would expect some effects to be positive and some negative, due to sampling variation. Multiplying all negative effects by -1 would then lead in

an overall overestimation of the effect sizes. To avoid this, we checked all meta-analytic articles for that yielded a negative average meta-analytic effect, and concluded that in all of these cases the result was in line with the expectations of the meta-analysts. This meant that if a recoded primary study yielded a negative outcome, this study showed an effect contradicting the hypothesis of the meta-analysts. The following analyses all use coined primary effect sizes, unless stated otherwise. All our data and analysis scripts for both confirmatory and exploratory analyses are freely available from https://osf.io/z8emy/.

9.3 Effect Sizes in Intelligence Research

We were able to convert the effect sizes from 2,439 primary studies to Fisher Z values. For four primary studies, we were not able to convert the effect sizes, because information on sample sizes was missing. Figure 9.2 shows the distribution of the 2,439 primary effect sizes, converted back to Pearson's correlations to facilitate interpretation. The unweighted mean effect size in the 2,439 primary studies was a Pearson's correlation of .25 (SD = .23), with a minimum of -.94, and a maximum of .95. We also calculated the average effect size per type using a random effects meta-analysis across all primary studies of the same type. This resulted in a (weighted) average Pearson's correlation of .26. The sample size in the primary studies varied widely, from 6 participants to over 1,530,000. The median total sample size per primary study was 60.

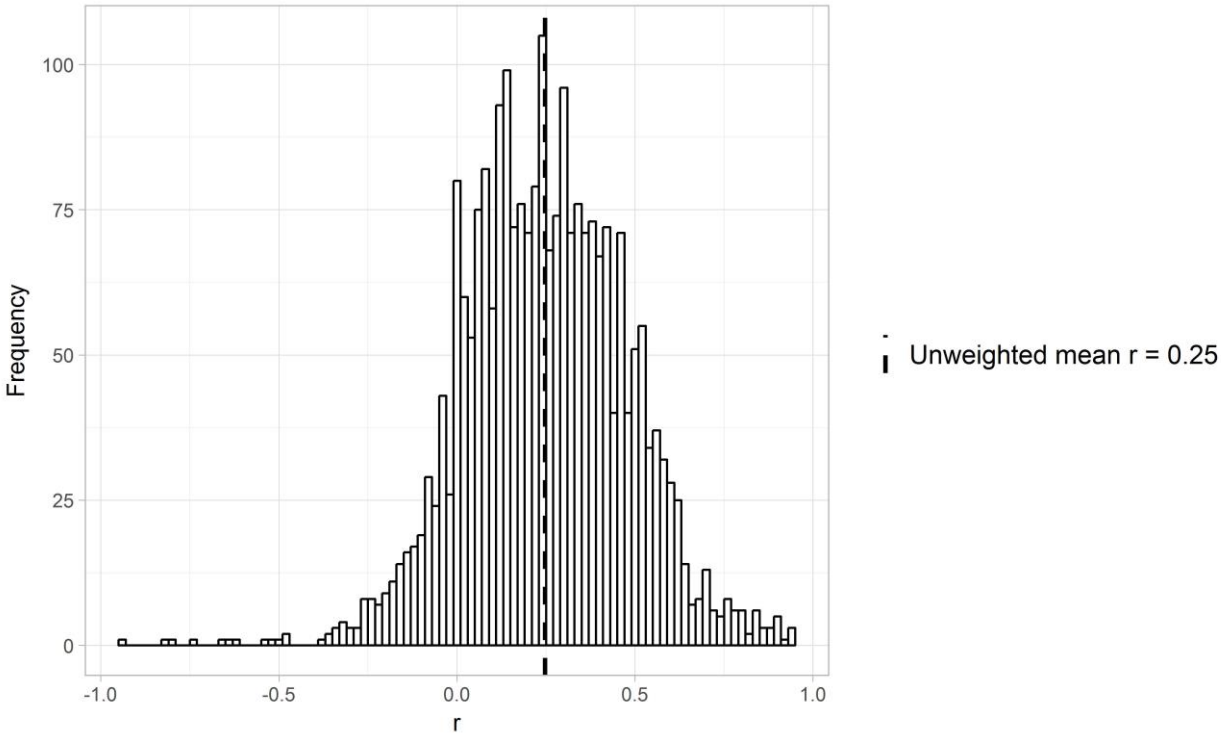


Figure 9.2

Histogram of the effect sizes of 2,439 primary studies about intelligence. All effect sizes were converted from Fisher's Z to Pearson's correlation to facilitate interpretation.

We also looked at the sample sizes and effect sizes for the five types of research, separately, and found some clear differences between them (see Table 9.2). First, the majority of meta-analyses and primary studies concern either research about group differences in intelligence (59 meta-analyses, 1,247 primary studies, or correlational research (31 meta-analyses, 779 primary studies), in which intelligence is related to other continuous psychological constructs. The fact that certain research types occur more often in the literature (at least as included in meta-analyses), might also explain why there are more meta-analyses in these fields. However, we also noted that some meta-analyses seemed to overlap substantially. For instance, in our sample we included 12 meta-analyses about the cognitive abilities in schizophrenia patients. This can be a sign of redundancy in the meta-analyses that are produced in this field, as has been found in medicine research (Ioannidis, 2016).

Interestingly, in all different research types we found relatively low median sample sizes, considering the average effect sizes in those fields. This suggests that intelligence research might be generally underpowered. Note, however, that median sample sizes also vary considerable across the fields, with those of behavioral genetics (169) being much larger than for the other four types (49-65). The meta-analytic effect size also differs across the five types. We will come back to this in the next section where we estimate the power across all meta-analyses and for the different types of research, separately.

Table 9.2

Descriptive statistics of the primary studies split up in five types of studies and in total. The random effects meta-analytic summary effect was calculated across all primary studies per subtype, using random effects meta-analysis. We calculated the power of each primary study to detect the summary effect in the corresponding meta-analysis. We reported the median of all power estimates per subtype.

	# Meta-analyses	**# Unique primary studies**	**Total N**	**Median total N**	**Median unweighted Pearson's r**	**Meta-analytic summary effect (r)**	**Median power**
1. Predictive validity & correlational studies	31	779	367,643	65	0.26	.28	53.5%
2. Group differences (clinical & non-clinical)	59	1,247	19,757,277	59	0.26	.28	56.6%
3. Experiments & interventions	20	185	24,040	49	0.18	.19	22.9%
4. Toxicology	16	169	25,720*	60	0.15	.16	22.3%
5. (Behavior) genetics	5	59	30,545	169	0.07	.12	8.9%
Total	**131**	**2,439**	**20,205,225**	**60**	**0.25**	**.26**	**48.8%**

* One of the meta-analyses reported two studies with non-integer total sample sizes. It seems that the authors wanted to correct their sample sizes to ensure they did not count the same observations twice. Here, we rounded the total sample size.

9.4 Power in Intelligence Research

Low power leads to several problems. First, if a study is underpowered, the chance that it detects a true effect decreases. Second, in a set of studies containing both null effects and genuine effects, lower power increases the chance that a significant study represents a false positive (Ioannidis, 2005). Third, when a significant finding in an underpowered study does reflect a true effect, it is likely to be overestimated (Button et al., 2013). These problems occur even when all other research practices are ideal, and there is strong evidence that they are not. Researchers have a strong focus on reporting significant results (Franco et al., 2014; LeBel et al., 2013). To obtain significant results they seem to make strategic use of flexibility in data analysis, also referred to as "researcher degrees of freedom" (Agnoli et al., 2017; John et al., 2012; Simmons et al., 2011; but see Fiedler & Schwarz, 2016). Underpowered studies are particularly vulnerable to such researcher degrees of freedom, both because they probably will not find a significant effect in the first place, but also because effect sizes are particularly strongly affected by researcher degrees of freedom in a study with low power (Bakker et al., 2012).

Several studies found that research in psychology is underpowered (Button et al., 2013; Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Despite repeated recommendations to change this, there seems to have been no overall improvement (Fraley & Vazire, 2014; Hartgerink et al., 2017; Marszalek et al., 2011; Maxwell, 2004; Stanley, Carter, & Doucouliagos, 2017; Szucs & Ioannidis, 2017; but see Maddock & Rossi, 2001; Rossi, 1990).

Here, we estimated the median power in intelligence research, based on the power of each primary study to detect the corresponding meta-analytic effect. First, for each meta-analysis we calculated the average Fisher's Z using a random effect meta-analysis, and used it to represent the true population effect size. To facilitate our calculations, we converted all 131 meta-analytic effects to Pearson's correlations. We then calculated the power of each primary study to detect the meta-analytic effect in the corresponding meta-analysis with a t-test for correlation, assuming $\alpha = .05$ and two-sided tests, using the R package "pwr" (Champely, 2017). Note that the meta-analytic effect size estimates are probably inflated, precisely because of selective reporting of significant findings and researcher degrees of freedom, that inflated the effect estimate (Francis, 2013b; Nuijten, Van Assen, Veldkamp, & Wicherts, 2015; Pereira & Ioannidis, 2011). Furthermore, in random effects meta-analyses small studies receive relatively more weight than in fixed effect meta-analyses. If small studies are more likely to contain overestimated effects, the meta-analytic effect size will be larger in

a random effects meta-analysis than in a fixed effects meta-analysis, again inflating our power calculations (Borenstein et al., 2009).[39]

We found a median power of 48.8%, which is well below the recommended power of 80% (Cohen, 1988). The median power per type of research (see Figure 9.3) differed substantially. Studies on group differences and correlational research showed the highest median power (56.6% and 53.5%, respectively). Studies in behavior genetics had the lowest median power (8.9%), even though their median sample size was much larger than for the other four types of research.

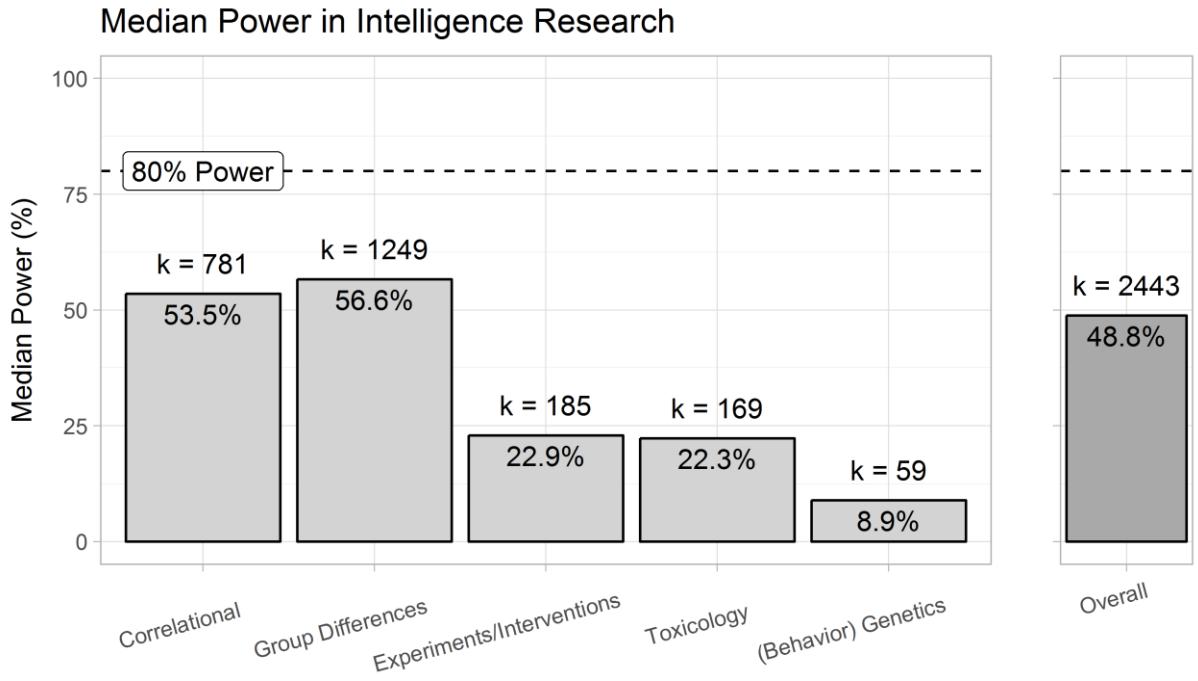


Figure 9.3

The median power in different subtypes of intelligence research and intelligence research as a whole. The true effect sizes were approximated with random effects meta-analyses. The number of studies per type of research is indicated with the letter "k".

It has been argued that estimates of the average power across an entire field lack nuance (Nord, Valton, Wood, & Roiser, 2017) and could paint the possibly misleading picture that *all* studies in a field are underpowered, which is not necessarily true. Indeed, across the 2,439 primary studies, we found that the power varied widely (see Figure 9.4). Overall, less than one third (29.8%) of all primary studies included in our sample reached the recommended power of 80% or higher (Cohen, 1988). Splitting up per research type, we

[39] As a robustness analyses, we also estimated power based on fixed effect meta-analyses, and found lower overall estimates. When assuming fixed effects, median power decreased slightly to 45.6%. See the appendix for details.

found substantial differences in power distributions. The percentage of studies that reached 80% power or higher was 29.4% (correlational), 35.1% (group differences), 14.6% (experiments), 12.4% (toxicology), and 16.9% (behavior genetics).

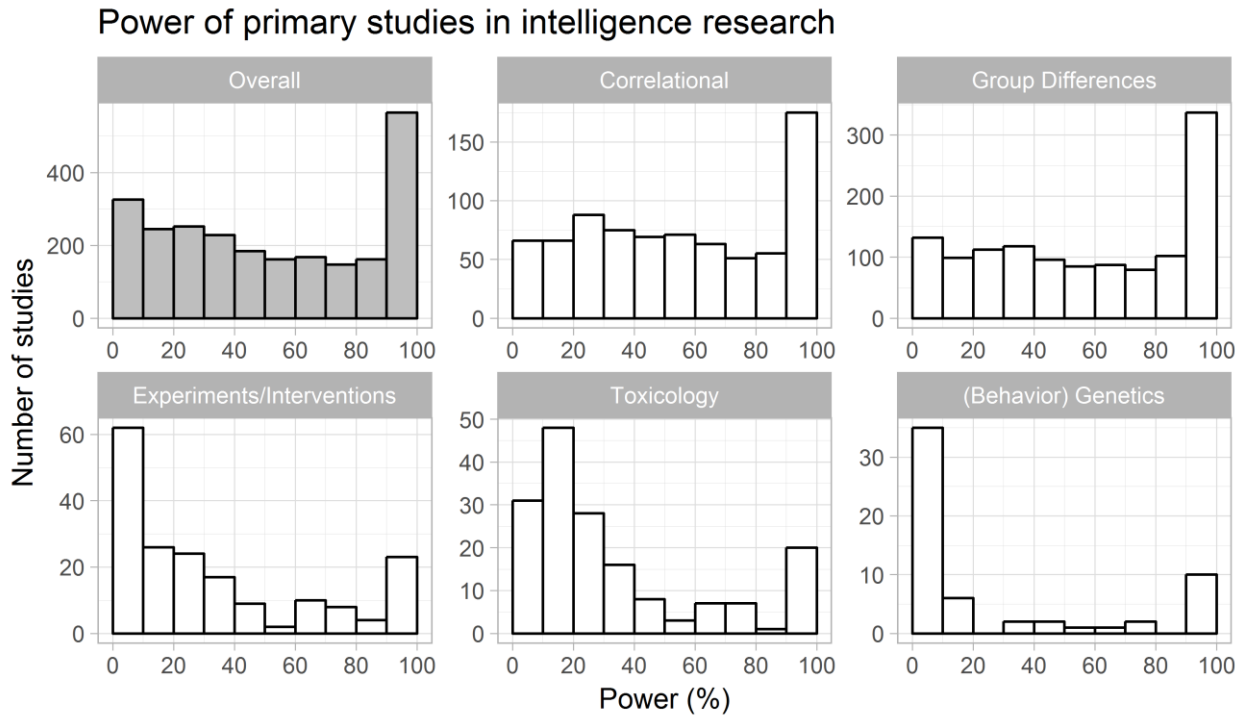


Figure 9.4

Estimated power of 2,439 primary studies from 131 meta-analyses in intelligence research, split up per research type and overall. We calculated the power of a primary study with a specific sample size to detect the meta-analytic effect (assuming random effects) in the corresponding meta-analysis, assuming α = .05 and two-sided tests. The percentage of studies that reached 80% or higher was 29.8% (overall), 29.4% (correlational), 35.1% (group differences), 14.6% (experiments), 12.4% (toxicology), and 16.9% (behavior genetics).

The power distributions of experimental research and behavior genetics in Figure 9.4 show high peaks at a power below 10%, indicating that extremely underpowered studies are prevalent in these fields. In experimental research, the median sample size was 49 in total, which means that in a standard experimental design with one treatment group and one control group, there are only 25 subjects per cell. For such a design to reach 80% power, the true Cohen's d needs to be at least .81 (considered to be a large effect); this corresponds to a true correlation of about .37, which is arguably unrealistic in interventions often intended to increase IQ. In genetic association studies, the true effects under investigation are also likely to be very small (Plomin & Deary, 2014). Even though the median sample size in this field was larger than in the other subfields (median N = 169), it is still probably not large enough to detect true effects with sufficient power. As an illustration, to achieve 80% power with a

sample size of 169, the true correlation between a specific gene and intelligence needs to be .21, which again is probably unrealistic (although other behavior genetic tests such as heritability being larger than zero require smaller samples).

Overall, power in intelligence research seems to be low. As we discussed above, studies with low power are more at risk to overestimate effect sizes when biases are present (e.g., publication bias, or researcher degrees of freedom; Bakker et al., 2012; Nuijten et al., 2015). In the sections below, we investigate whether different biases are likely to have affected effect size estimates in intelligence research.

9.5 Bias-Related Patterns in Effect Sizes

We investigated whether the following five specific bias-related patterns were present in intelligence research: small study effect, US effect, decline effect, early-extremes effect, and citation bias. Evidence for these biases can be analyzed with multilevel weighted regression analyses that take into account that primary studies are nested within meta-analyses (Fanelli et al., 2017; Fanelli & Ioannidis, 2013). A downside to this method is that they require strong statistical assumptions that are difficult to meet with these data. A more straightforward way to analyze if there is evidence for these biases is via two-step meta-regressions (Fanelli et al., 2017; Fanelli & Ioannidis, 2014; Nuijten et al., 2014). Here, bias-related patterns are investigated for each individual meta-analysis, and this information is then combined across all meta-analyses. We used this analytical strategy here.

9.5.1 Two-Step Meta-Regressions

Within individual meta-analyses, we investigated each of the biases via meta-regression. All meta-regressions we estimated were of the following general form:

$$Fisher's\ Z_{ij} = a^j + b^j Predictor_{ij} + \varepsilon_{ij},$$

Equation 9.1

where the dependent variable *Fisher's* Z_{ij} was the coined effect size of primary study *i* in meta-analysis *j*, weighted by its standard error, a^j is the intercept, $Predictor_{ij}$ was a study-level predictor for *Fisher's Z*, and b^j indicates the unstandardized regression coefficient of $Predictor_{ij}$ in predicting *Fisher's Z*. All meta-regressions were estimated using the rma() function in the R package metafor (Viechtbauer, 2010). We assumed random effects models, and we used the Paule-Mandel estimator for random effects because it has the most favorable properties in most situations to estimate variance in true effect size between studies (Langan, Higgins, & Simmonds, 2017; Veroniki et al., 2016). Table 9.3 shows a summary of each of the meta-regressions we estimated for the separate biases.

After running these meta-regressions for each of the meta-analyses, we obtained estimates of the bias (and their SEs) in the separate meta-analyses. To combine this information across meta-analyses, we then ran another meta-analysis to obtain a weighted

average of all obtained regression coefficients b^j. At this meta-meta-level, we again used the Paule-Mandel estimator for random effects. We assumed random effects models at both levels, because it is highly unlikely that the same population effect underlies (1) every study within the meta-analyses, and (2) every meta-analysis in the meta-meta regression (Borenstein, Hedges, Higgins, & Rothstein, 2005b). In a previous discussion about estimating bias in meta-analyses (Fanelli & Ioannidis, 2013, 2014; Nuijten et al., 2014), Fanelli and Ioannidis (2014) argued that choosing a random effects model at both levels unnecessarily reduces power, and they advocated the use of a fixed effect models within each of the meta-analyses to decrease the amount of random fluctuation in the estimates. However, we argue that the choice for a fixed effect or random effects model is a theoretical choice, not a statistical one (see also Borenstein, Hedges, Higgins, & Rothstein, 2005a).

In this Chapter, we ran a substantial number of significance tests. To correct for multiple comparisons, we applied a Bonferroni correction based on the number of predictors (5 patterns of bias) for our main meta-meta-regressions based on Equation 9.1, resulting in a significance level of .01.

Table 9.3

Overview of the meta-meta-regressions we estimated in this paper to investigate different predictors for effect size that could potentially indicate bias. We estimated these bias-related patterns in five separate analyses.

Type of bias	"Predictor" in $\boldsymbol{Fisher's\ Z_{ij} = a^j + b^j Predictor_{ij} + \varepsilon_{ij}}$	Estimate of the mean parameter across meta-analyses [95% CI]	Heterogeneity of the estimate (SE)
1. Small study effect	Standard error of primary study's effect size (SE)	0.68 [0.44; 0.92]	τ^2 = 0.71 (0.24)
2. US effect	US*SE	0.47 [0.01; 0.93]	τ^2 = 0.21 (0.75)
3. Decline effect	Order of publication	0.001 [-0.003; 0.004]	τ^2 = 0.00 (0.00)
4. Early-extremes effect*	$deviation = \lvert Fisher's\ Z_{ij} - Fisher's\ Z_j \rvert$, $deviation = a^j + \boldsymbol{b^j PublicationOrder_{ij}} + \varepsilon_{ij}$	-0.001 [-0.005; 0.002]	τ^2 = 0.00 (0.00)
5. Citation bias	Citations per year	0.001 [-0.001; 0.003]	τ^2 = 0.00 (0.00)

* We estimated the presence of early-extremes using a different dependent variable; instead of predicting the primary study' effect size itself, we predicted the deviation of the primary study effect size from the meta-analytic effect.

9.5.2 Results Bias Analyses

The results of the five meta-meta-regressions are shown in Table 9.3 and in Figure 9.5. In the Figure, each panel shows a histogram of the estimated meta-regression coefficients for each type of potential bias. The vertical dashed line indicates the meta-analytic weighted average of these coefficients. We found significant evidence for a small study effect across meta-analyses. We also found that the small study effect was stronger in studies from the US than for non-US studies, but this effect was not significant anymore when we corrected for multiple testing. We found no evidence for a decline effect, early-extremes, or citation bias across meta-analyses. We discuss the results in more detail below.

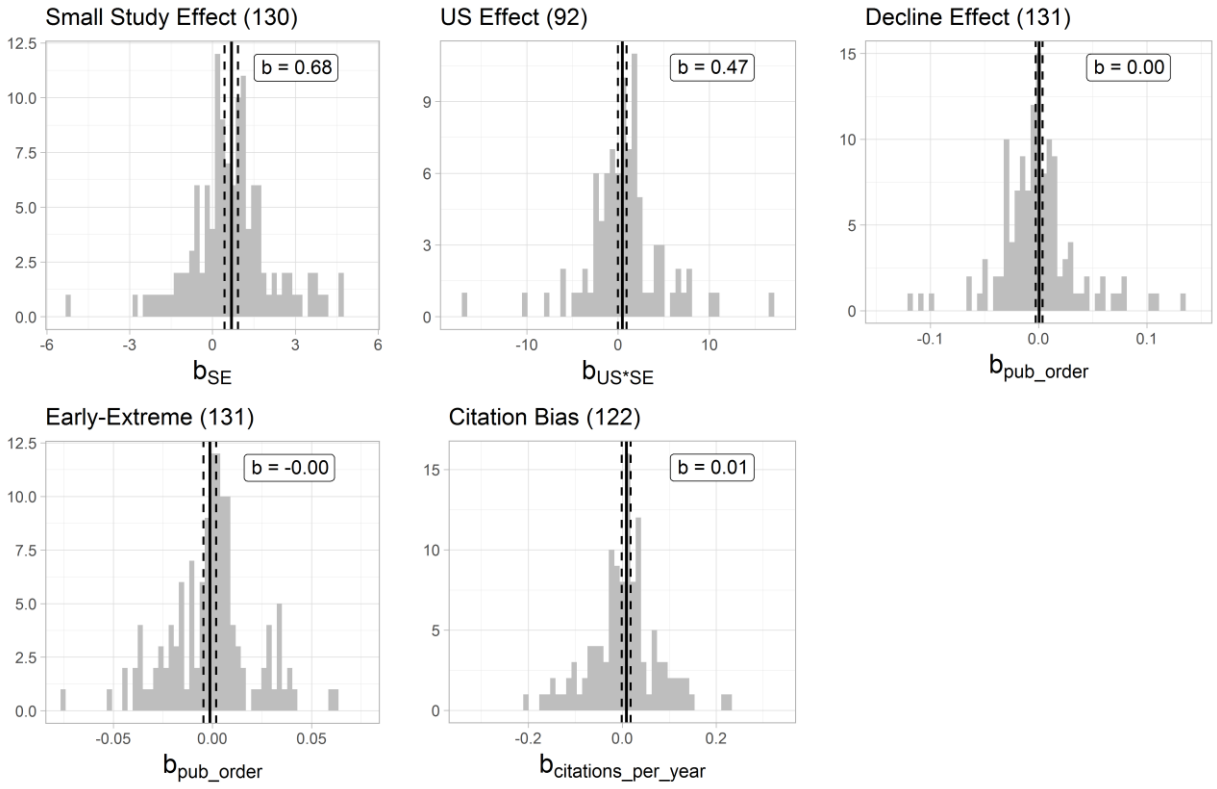


Figure 9.5

Histograms of estimated meta-regression coefficients for five different bias patterns. The vertical solid line indicates the meta-analytic weighted average of the coefficients, the estimate is also depicted in the plots. The dashed lines indicate the 95% confidence interval. We truncated the x-axes at 3 times the standard deviation of the b-coefficients for the sake of readability. The complete distributions can be found in the Appendix. In the titles of the histograms, in parentheses, we indicated the number of meta-analyses for which we could estimate this bias. In the small study histogram and US effect histogram we removed an outlier to improve readability of the plot.

9.5.2.1 *Small Study Effect*

We excluded one meta-analysis from this analysis, because there was too little variation in the standard errors of the primary studies to estimate a small study effect. Across the remaining 130 meta-analyses, we found a significant overall small study effect, b_{SE} = 0.68, SE = 0.12, Z = 5.57, p < .001, 95% CI = [0.44; 0.92], I^2 = 46.5%, τ^2 = 0.71 (SE = 0.24).[40] In 17 of the meta-analyses (13.1%) we found a significant small study effect (α = .05). Because this regression test has low power when meta-analyses include few studies (k < 10), it is advised to retain a significant level of α = .10 (see, e.g., the example in Sterne & Egger, 2005), in which case 19 meta-analyses (14.6%) show a significant small study effect. We ran a robustness analysis including only meta-analyses with at least ten primary studies, and still found consistent evidence for a small study effect (see Appendix). We did not find consistent differences in the small study effect between different types of studies. See the Appendix for details.

Concretely, the overall small study effect across meta-analyses means that if the total sample size of a study increases from 50 to 100 observations, Fisher's Z decreases from .24 to .21 (corresponding to r = .23 and r = .21, respectively). If the sample size increases to 1,000 observations, Fisher's Z would decrease to .16 (r = .16). Figure 9.6 shows how the effect size decreases when sample size increases in each of the individual meta-analyses (thin grey lines) and overall (thick black line). We chose to express the effect size in Pearson's r rather than Fisher's Z to facilitate interpretation, and to set natural bounds on the y-axis.

[40] We found one meta-analysis with an extreme small study effect (b_{SE} = 41.79, SE = 42.50). Removing this outlier from the overall meta-meta-regression did not affect the overall estimate of the small study effect: b_{SE} = 0.68 (SE = 0.12, Z = 5.57, p < .001, 95% CI = [0.44; 0.92]), so we decided to not remove this meta-analysis from further analyses.

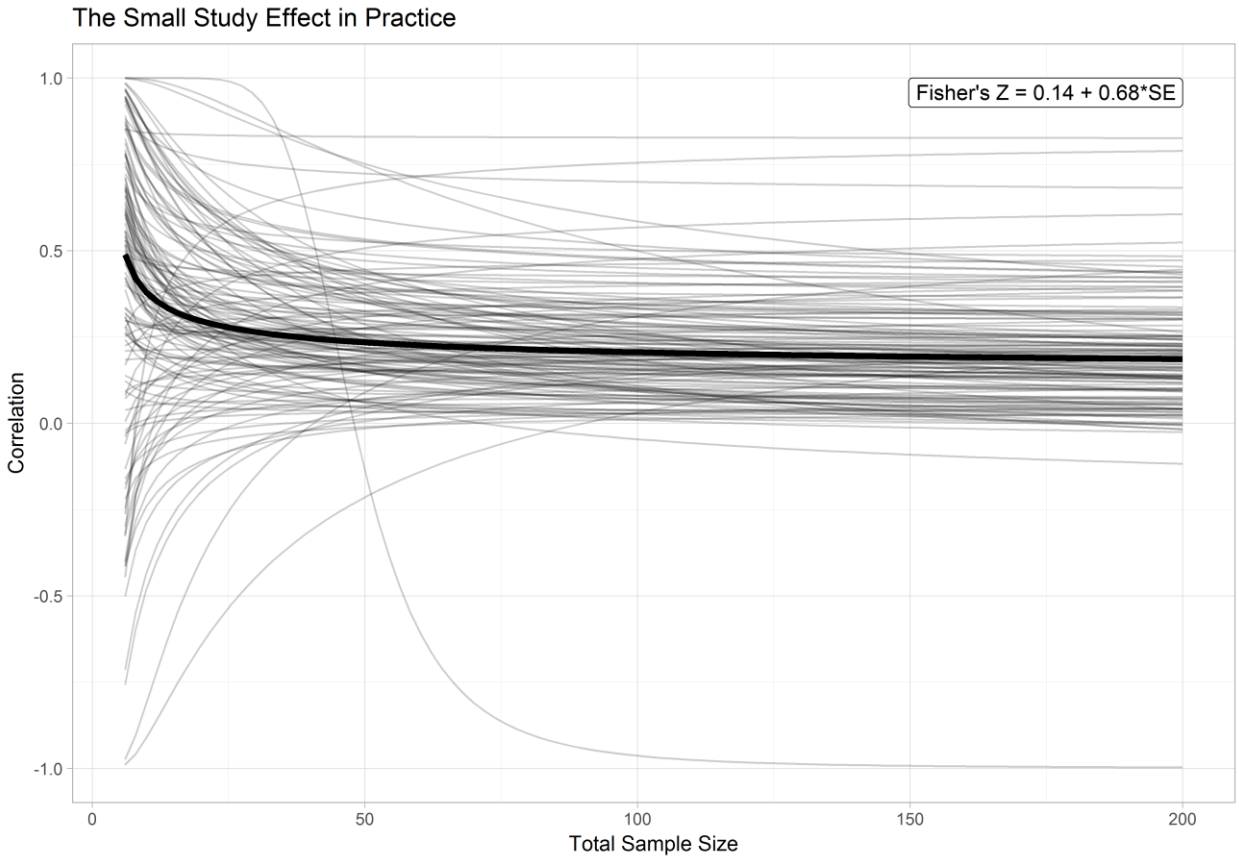


Figure 9.6

Illustration of the small study effect in intelligence research. The solid black line shows the expected Pearson's correlation for studies of different total sample sizes (starting at N = 6, the smallest sample size in our study), estimated across all meta-analyses. The thin, grey lines show the expected small study effect within each of the meta-analyses.

9.5.2.2 *US Effect*

In 39 meta-analyses, the meta-regression could not be estimated because there was not enough data bearing on the US effect. In most cases, this meant that either almost none or almost all primary studies in a meta-analysis were from the US. For the model to fit, at least two studies had to be from the US if the rest was not, or vice versa. When we summarized the estimated US effect across the remaining 92 meta-analyses, we found a positive overall estimate of the interaction between ES and US on effect size, but this result was not statistically significant after we corrected for multiple testing (Bonferroni corrected α = .05/5 = .01), b^{j}_{US*SE} = 0.47, SE = 0.24, Z = 1.99, p = .047, 95% CI = [0.01; 0.93], I^2 = 4.47%, τ^2 = 0.21 (SE = 0.75).[41] Of the 92 meta-analyses in which we could estimate a US effect, 5 (5.4%) showed

[41] We found one meta-analysis with an extremely negative interaction estimate (b^{j}_{US*SE} = -89.12, SE = 49.71). Removing this outlier from the overall meta-meta-regression did not strongly affect the overall estimate of the

significant evidence for a US effect (α = .05). In our analysis of the US effect, we did find some differences in the US effect between different study types, but study type was not a significant moderator. See the Appendix for details. To conclude, the evidence for a US effect in the intelligence literature is weak and inconsistent (see also Chapter 7; Fanelli et al., 2017).

9.5.2.3 *Decline Effect*

When we combined all 131 regression coefficients, we found no overall evidence for a decline effect, $b^{j}_{PubOrder}$ = 0.001, SE = 0.001, Z = 0.279, p = .781, 95% CI = [-0.003; 0.004], I^2 = 63.8%, τ^2 = 0.00 (SE = 0.00). In six cases (4.6%) we found significant evidence for a decline effect against α = .05. We found some evidence that the decline effect was moderated by study type, but any differences between types were small.[42] All estimates for the decline effect per study type can be found in the Appendix.

It is possible that the decline of effects over time is not linear, but that there is an overall "winner's curse", and only the first effect is overestimated (Trikalinos & Ioannidis, 2005). To test this, we again used a meta-meta-regression approach in which we predicted effect size with a dummy-coded variable indicating if a study in a meta-analysis was published first or not. We excluded one meta-analysis due to convergence problems. We used a random effects meta-analysis to summarize all 130 obtained $b^{j}_{FirstPublished}$ coefficients, and we found no overall evidence for a winner's curse, $b^{j}_{FirstPublished}$ = 0.02, SE = 0.01, Z = 1.56, p = .119, 95% CI = [-0.006; 0.049], I^2 = 32.5%, τ^2 = 0.00 (SE = 0.00).[43] We also tested a less extreme version of the winner's curse, by using 1/(publication order) as a predictor for effect size, but this analysis also did not show evidence for an overall decline effect. See the Appendix for details. In sum, we conclude that there is no clear evidence for an overall decline effect or a winner's curse in intelligence research.

9.5.2.4 *Early-Extremes*

When we summarized all obtained 131 coefficients, we found no evidence for an early-extremes effect (Bonferroni corrected α = .05/5 = .01), $b^{j}_{PubOrder}$ = -0.001, SE = 0.002, Z = -0.804, p = .422, 95% CI = [-0.005; 0.002], I^2 = 89.4%, τ^2 = 0.00 (SE = 0.00). Of the 131 estimations for $b^{j}_{PubOrder}$, eight (6.1%) were significantly smaller than zero (α = .05), indicating evidence for an early-extremes effect. As a robustness test, we also used 1/(publication order)

---

overall interaction effect between US*SE: b^{j}_{US*SE} = 0.48 (SE = 0.23, Z = 2.12, p = .034, 95% CI = [0.04; 0.93]), so we decided to not remove this meta-analysis from further analyses.

[42] We found a significant (α = .05) overall decline effect in the 5 meta-analyses in behavior genetics ($b^{j}_{PubOrder}$ = -0.023, SE = 0.011, Z = -2.18, p = .029). Conversely, we found a significant, though very small, "increase" effect in the 59 meta-analyses in which they analyzed group differences ($b^{j}_{PubOrder}$ = 0.003, SE = 0.002, Z = 1.99, p = .011).

[43] When we ran this analysis again, including SE as a control variable, we did find evidence for an overall winner's curse. However, this effect was small ($b^{j}_{FirstPublished}$ = 0.032), and it is hard to interpret substantively why a winner's curse would only show if we control for SE. See the Appendix for details.

as a predictor for absolute effect size, but this analysis also did not show evidence for an overall early-extremes effect. Nor did we find any difference in the early-extremes effect for different study types. See the Appendix for details. Based on our results, we conclude that there is no meta-analytic evidence for the presence of an early-extremes effect across the intelligence literature.

9.5.2.5 *Citation Bias*

In five meta-analyses, we could not estimate the model, because we had insufficient information about the citation rates.[44] When we summarized all 126 obtained regression coefficients, we did not find evidence for overall citation bias, $b^{j}_{CitPerYear}$ = 0.001, SE = 0.001, Z = 1.276, p = .202, 95% CI = [-0.001; 0.003], I^2 = 58.3%, τ^2 = 0.00 (SE = 0.00). We found significant evidence for citation bias in 10 of the remaining 126 meta-analyses (7.9%; α = .05). We ran additional robustness analyses including several control variables, and consistently found no clear evidence for citation bias, and no differences in citation bias between study types (see the Appendix for details).

9.6 Robustness Checks and Exploratory Analyses

In an exploratory analysis, we found that across all intelligence meta-analyses, sample size seemed to increase with publication order. In other words, within a meta-analysis, studies that were published earlier had smaller samples than those published later. However, this effect was qualified by substantial heterogeneity, hence it may not generalize to all lines of intelligence research (see the Appendix for details). As this change in sample size over time might be related to any change in effect size over time, we also ran the analyses for the decline effect and early-extremes effect again, including standard error of the primary study as a control variable. This did not affect our overall conclusions that there is no overall evidence for either bias patterns.

Furthermore, we ran several robustness analyses when we estimated citation bias. Citation bias could be related to journal impact factor and sample size of the study (Jannot et al., 2013), so we ran several additional analyses included different combinations of these control variables. In none of these robustness analyses we found evidence for citation bias (see the Appendix for details).

Finally, for all five bias patterns (small study effect, US effect, decline effect, early-extreme effect, and citation bias), we tested whether any heterogeneity in the estimated effects could be explained by study type. This was only the case for the decline effect, although the differences between study types were very small and only reached statistical

[44] In four meta-analyses, all but one primary studies were from China, and we could not find information about the number of citations. The remaining meta-analysis did not synthesize primary studies, but investigated scores on the SAT-M in different years, which were not cited.

significance when we did not correct for multiple testing. For the sake of completeness, we report all estimates of bias-related patterns for the separate types in the Appendix.

9.7 Discussion

In this study, we analyzed 2,439 effect sizes from 131 meta-analyses about intelligence, based on over 20 million participants. We found that the typical effect size in this field was a Pearson's correlation of .26. This is slightly higher than the average effect in psychology (r = .24; Bakker et al., 2012).[45] We found relevant differences between the subtypes of intelligence research. Specifically, we found that the types of studies that were least complex in terms of methodology, were most often conducted and also found the largest effect sizes; in correlational research and research about group differences we found an average effect size of *r* = .28. In less prevalent - and arguably more complex - types of research the effect size was lower and decreased rapidly from *r* = .19 in experimental research, to *r* = .16 in toxicological studies, and *r* = .12 in behavior genetics. Given the typical effect sizes, the sample sizes in for all study types were relatively small, with an overall median of 60.

Both small effect sizes and small sample sizes increase the risk that a study finds a false positive (Ioannidis, 2005). These problems are largest in toxicological and behavior genetic studies of intelligence. Another risk factor identified by Ioannidis (2005) is the "popularity" of a field; when more research teams are involved in a scientific field, the competition increases and there is stronger pressure to report statistically significant results. Given our results, this would be most problematic in correlational studies and studies about group differences. Another risk factor for false positives is flexibility in research design and data analysis (Ioannidis, 2005). It is not immediately clear for which subfield in intelligence research flexibility would be highest. Indeed, it has been argued that the social sciences in general have a lot of flexibility in design and analysis (Fanelli, 2010; Fanelli & Ioannidis, 2013), which might mean that the risk of false positives might be high across the entire field of intelligence. At the same time, however, the measures used in intelligence research are typically quite well established and standardized, allowing less flexibility in operationalizing the key variable of interest. However, this is just one of many potential degrees of freedom (Wicherts et al., 2016) that might create biases in research into intelligence.

If a study finds a significant effect, the probability that it is a false positive increases when power is lower (Button et al., 2013; Ioannidis, 2005). Any overestimation of the effect in underpowered studies is aggravated by publication bias and the opportunistic use of researcher degrees of freedom (Bakker et al., 2012; Kraemer et al., 1998; Nuijten et al., 2015). We estimated the power of each primary study in our sample by using the meta-analytic

[45] Based on an average total sample size of 40 and d = .50.

effect size as a proxy for the true effect size, and we found that the median power was 48.8%. Less than a third of all studies (29.8%) reached the recommended power of 80% or more. Again, we found relevant differences between subfields; power was lowest in experimental research, toxicology and behavior genetics, although all subfields suffered from low power.

Based on these findings, we expected to find evidence for overestimated effects across intelligence research, but we did not have strong expectations for differences in bias patterns between subfields. We investigated five bias-related patterns frequently discussed in the literature: the small study effect, US effect, decline effect, early-extremes effect, and citation bias. We found evidence for a small study effect across the intelligence literature: smaller studies seemed to yield higher effect sizes, which could be a sign of publication bias, especially given the overall low power we found. All five subfields showed consistent evidence for a small study effect, and we did not find evidence that the small study effect was stronger in any of these fields.

The evidence that the small study effect is stronger for US studies (the US effect) was weak and inconsistent. This is in line with previous findings that the US effect does not seem robust against method of analysis (Chapter 7; Fanelli et al., 2017). We also did not find consistent evidence for an overall decline effect, early-extremes effect, or citation bias.

Compared to other fields, the potential problems in intelligence research do seem less severe. First, the median power in intelligence research seems higher than the median power estimated in neuroscience (8-31%; Button et al., 2013), psychology (between 12% and 44%; Szucs & Ioannidis, 2017; Stanley et al., 2017) , behavioral ecology and animal research (13–16% for a small effect and 40–47% for a medium effect; Jennions & Moller, 2003), and economics (18%; Ioannidis, Stanley, & Doucouliagos, 2017), but slightly lower than social-personality research (50% for r = .20; Fraley & Vazire, 2014). Second, we did not find clear trends in effect sizes over time, which might indicate that the field of intelligence research is less susceptible to time-lag biases such as the decline effect or the early-extreme effect (Trikalinos & Ioannidis, 2005). This is in line with the theory that such biases would mainly affect research fields in which results can be rapidly produced and published, which might not apply to the majority of studies about intelligence (Ioannidis & Trikalinos, 2005). Finally, citation bias seems to be a problem in medical research (Jannot et al., 2013), and there is some evidence that it also affects social sciences in general (Fanelli et al., 2017), but in intelligence research in specific we find no evidence that larger effects are cited more often.

In our study, we were limited to meta-analyses that actually included the full data table, which was very often not the case (viz. in 81 meta-analyses). It is imaginable that these meta-analyses contained stronger and/or other patterns of bias. It could be the case that meta-analysts who go through the effort of presenting the full data in their paper are more rigorous in their work. This could then mean they may also have tried harder to find all primary studies (published and unpublished), which would have decreased overall bias in the

meta-analysis. Furthermore, not all studies in the intelligence literature end up being included in a meta-analysis. That said, our sampling scheme provided us with a fairly large and diverse set of over 2,400 studies, providing a broad overview of the field.

Another limitation in our study is that investigating patterns in effect sizes is an indirect way of assessing potential patterns of bias. Many of the patterns we found can have several underlying causes. For instance, a small study effect can also arise through deliberate design choices and power analyses; if researchers expect to find a large effect, they can decide to collect a smaller sample. However, it appears that researchers seldom use formal power analyses to determine sample size (Bakker, Hartgerink, Wicherts, & van der Maas, 2016; Tressoldi & Giofre, 2015; Vankov, Bowers, & Munafò, 2014). Furthermore, we found the majority of the included studies to be underpowered. Another reason why a small study effect could arise is because of true heterogeneity in effect sizes (Sterne et al., 2011). It is possible that larger studies typically investigate small effects, and smaller studies large effects. For instance, in a clinical setting, participants in smaller studies may have been specifically selected to increase the chance that the treatment is effective. Conversely, in a larger sample, it might be more difficult to thoroughly administer the treatment, and effects might be smaller. For a full overview of alternative explanations, see Sterne et al. (2011).

Another limitation of our study is that we did not conduct a formal power analysis for our analyses. That means that our meta-meta-regressions might be underpowered (see Fanelli & Ioannidis, 2014) and any significance tests on our data need to be interpreted with care. In future research, it would be valuable to garner an even larger sample of meta-analyses, conduct formal power analyses for the (preferably preregistered) meta-meta-regressions, and consider other options for modelling the different types of bias.

When interpreting our current results, it is also important to take into account that these are patterns of potential bias that are aggregated over meta-analyses. Even though we found evidence for an overall small study effect, this does not mean that each meta-analysis in intelligence research shows this problem. Conversely, even though we did not find consistent evidence for an overall US effect, decline effect, early-extremes effect, or citation bias, this does not mean that these problems never occur in intelligence research. Furthermore, there are other types of scientific biases that we did not investigate here. For instance, previous studies showed evidence for a “grey literature bias” (Dickersin, 2005; Fanelli et al., 2017; Glass, Smith, & McGaw, 1981; McAuley, Pham, Tugwell, & Moher, 2000; Song et al., 2010). Here, unpublished literature, such as PhD theses or conference proceedings, typically report smaller effects than research published in peer reviewed journals, which could be a possible indicator for publication bias. Another type of bias we did not investigate is “industry bias”, where sponsorship from a company may be related the size and direction of published effects (Fanelli et al., 2017; Lexchin, Bero, Djulbegovic, & Clark, 2003). These might be interesting patterns to investigate in future research.

Based on our findings, we conclude that intelligence research shows signs that publication bias may have caused overestimated effects. Specifically, we found that power is often too low and in general smaller studies yielded larger effects. This is in line with the notion that publication bias and perhaps also researcher degrees of freedom in the analysis of data and reporting of results may have led to overestimated effects. Even though there might be several alternative explanations for these results, we argue that it is safe to assume that intelligence research is not immune to the problems in psychology, although the problems in intelligence seem to be less severe as compared to other fields. Even so, the field of intelligence research is not immune to biases and there is still room for improvement, particularly in experimental and behavior genetic studies where power remains low.

There are several strategies to improve the reliability of primary studies and meta-analyses (Asendorpf et al., 2013; Brandt et al., 2014; Munafò et al., 2017). One strategy is to increase power. One way to do this, is to increase sample size. Especially in intelligence research concerning correlations, non-clinical group differences, and interventions, there is no immediate reason why it is not possible to obtain larger samples. We suspect that the generally small sample sizes we observed are in part caused by the fact that researchers often base their sample size on rules of thumb or intuition, rather than formal a priori power analyses. This often leads to vastly underpowered designs (Bakker et al., 2016). When researchers do run a power analysis to determine the sample size, they need to take into account that published effect sizes from previous research are probably overestimated and might lead to overly small sample sizes. Ways to deal with this are correcting the observed effect sizes for publication bias (Anderson, Kelley, & Maxwell, 2017; van Assen, van Aert, & Wicherts, 2015; Vevea & Hedges, 1995) , calculating lower bound power (Perugini, Galucci, & Constantini, 2014), or base a power analysis on the smallest effect size of interest (Ellis, 2010). In studies where it is more difficult to obtain larger samples (for instance in research with special populations), multi-lab collaborations might be a solution. Examples of such collaborations are the embrace of consortia in genetics (Davies et al., 2015), or multi-lab (replication) efforts in psychology and biomedical sciences (Klein et al., 2014; Nosek & Errington, 2017; Open Science Collaboration, 2015).

Power also increases when the reliability of the measurements increases. This might be one explanation why intelligence research seems to have higher power than other fields in psychology. Intelligence research is over a hundred years old and has provided several replicable findings, including the positive manifold (Van Der Maas et al., 2006) and the hierarchical structure of individual differences (McGrew, 2009), heritability of intelligence (Plomin, DeFries, Knopik, & Neiderhiser, 2016), relative stability of individual differences over the life span (Deary, Whalley, Lemmon, Crawford, & Starr, 2000), and many important results concerning the predictive power of intelligence tests in educational and socioeconomic contexts (Neisser et al., 1996; Strenze, 2007). This extensive literature has also offered many

excellent measures of general intelligence and sub-domains of intelligence that show high reliabilities. Although less reliable measures might lower effects and associations in other fields (e.g., Vul, Harris, Winkielman, & Pashler, 2009), cognitive measures used in intelligence research are typically quite reliable, thereby offering relatively larger effects and associations.

Besides increasing power, another way to decrease overestimation is to eliminate publication bias. One way to avoid publication bias is to encourage (pre)registered reports (Chambers, 2013; Wagenmakers et al., 2012). Here, researchers submit a detailed research and analysis plan to a journal, *before* executing a study. This registration is peer reviewed and possibly amended during the planning phase. If the theoretical and methodological quality is high enough, researcher can earn an "in principle acceptance" and the researcher(s) can set out to collect the data and analyze the data as planned. This implies that if the registered plan is adhered to, the manuscript will be published regardless of the results. Another advantage of these registered reports, is that there is no more room for undisclosed flexibility in data analysis. Data exploration is still possible, but has to be mentioned explicitly in the paper. Eliminating publication bias does not only increase accuracy of effect size estimates, it can also be shown that it is more efficient in terms of the number of studies that have to be conducted to estimate an effect with a certain level of precision (van Assen, van Aert, Nuijten, & Wicherts, 2014a).

A final, more general recommendation is to increase transparency (Nosek et al., 2015; Nosek & Bar-Anan, 2012; Nuijten, 2017; Wicherts, 2013). If data, materials, and analysis scripts were available (which is often not the case; see, e.g., Krawczyk & Reuben, 2012; Vanpaemel et al., 2015; Wicherts & Bakker, 2012; Wicherts et al., 2006), it would be possible to reanalyze existing data to correct any mistakes, test the robustness of findings for different analytical choices, or even investigate new research questions.

In conclusion, intelligence research seems to be affected by low power and publication bias, which leads to systematically overestimated effects. Even though other scientific fields might be affected by these problems more strongly than intelligence research, we think that increasing power, eliminating publication bias, and promoting transparency can greatly improve the field of intelligence.

9.8 Appendix

In this Appendix we provide additional information on the analyses we described in Chapter 9, and also discuss the results of several robustness analyses.

9.8.1 Additional Variables

One variable that we did code, but did not use in our final analysis was “similarity” of the hypothesis of the primary study to that of the meta-analysis. For instance, a meta-analysis about the IQ of patients with schizophrenia could include similar primary studies that also specifically investigated the IQ of schizophrenia patients, but the meta-analysis could also include non-similar primary studies that focused on different characteristics of schizophrenia patients, but as an extra control also recorded their IQ.

9.8.2 Power Based on Fixed Effect Meta-Analyses

In our power estimates we approximated the true effects of each meta-analysis using random effects meta-analyses. We noted that assuming random effects may have inflated our power estimates, if there is a small study effect; in random effects meta-analyses small studies receive more weight than in fixed effect meta-analyses. If small studies are more likely to contain overestimated effects, the meta-analytic effect size will be higher in a random effects meta-analysis than in a fixed effects meta-analysis, inflating our power calculations (Borenstein et al., 2009). As a robustness analysis, we estimated power again, by approximating the true effects with fixed effect meta-analyses. Figure 9.7 shows the median power in intelligence research when the true effect sizes were approximated with fixed effects meta-analyses (FE). As expected, the power estimates were somewhat lower after assuming fixed effects instead of random effects, although differences were small and differences in median power between fields show a similar pattern compared with power estimates based on random effects (see Figure 9.3 in the main text). The only difference was for behavior genetics: the median power was slightly higher when assuming fixed effects (10.5%) than when assuming random effects (8.9%).

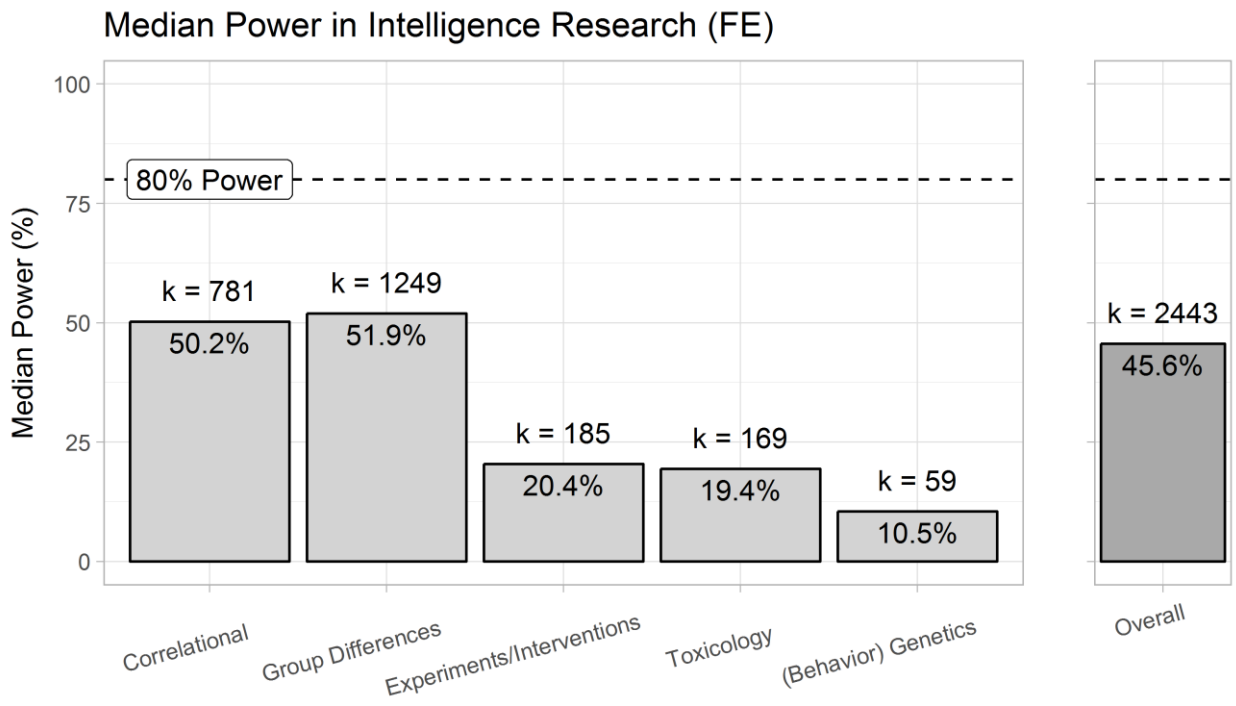


Figure 9.7

The median power in different subtypes of intelligence research and intelligence research as a whole. The number of studies per type of research is indicated with the letter "k".

We also looked at the distribution of power in the primary studies when power was estimated based on fixed effect meta-analyses (see Figure 9.8 and Figure 9.4 in the main text). Overall, the distributions look quite similar to the ones estimated based on random effects meta-analyses. As expected though, fewer studies reached 80% power or higher (see also Figure 9.9) when assuming fixed effects rather than random effects. Again, only in behavior genetics, we found that slightly more studies reached 80% power or higher when assuming fixed effects (18.6%) rather than random effects (16.9%).

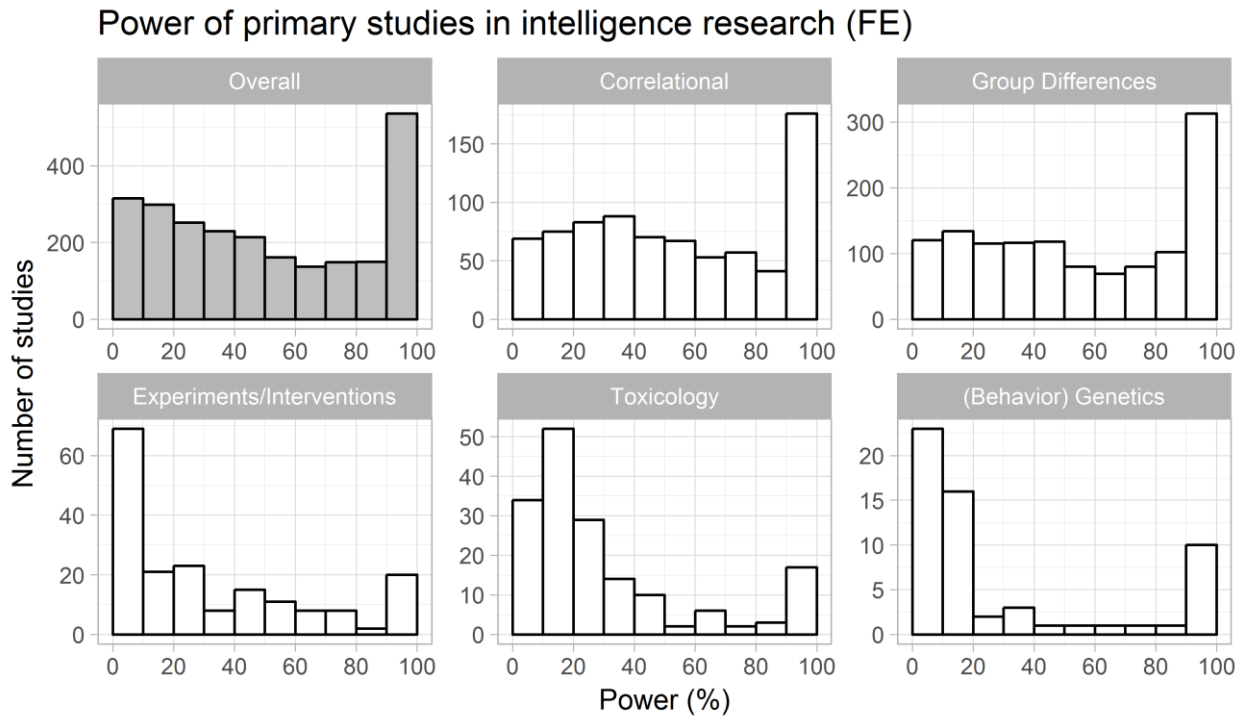


Figure 9.8

Estimated power of 2,439 primary studies from 131 meta-analyses in intelligence research, split up per research type and overall. We calculated the power of a primary study with a specific sample size to detect the meta-analytic effect (fixed effects) in the corresponding meta-analysis, assuming α = .05 and two-sided tests.

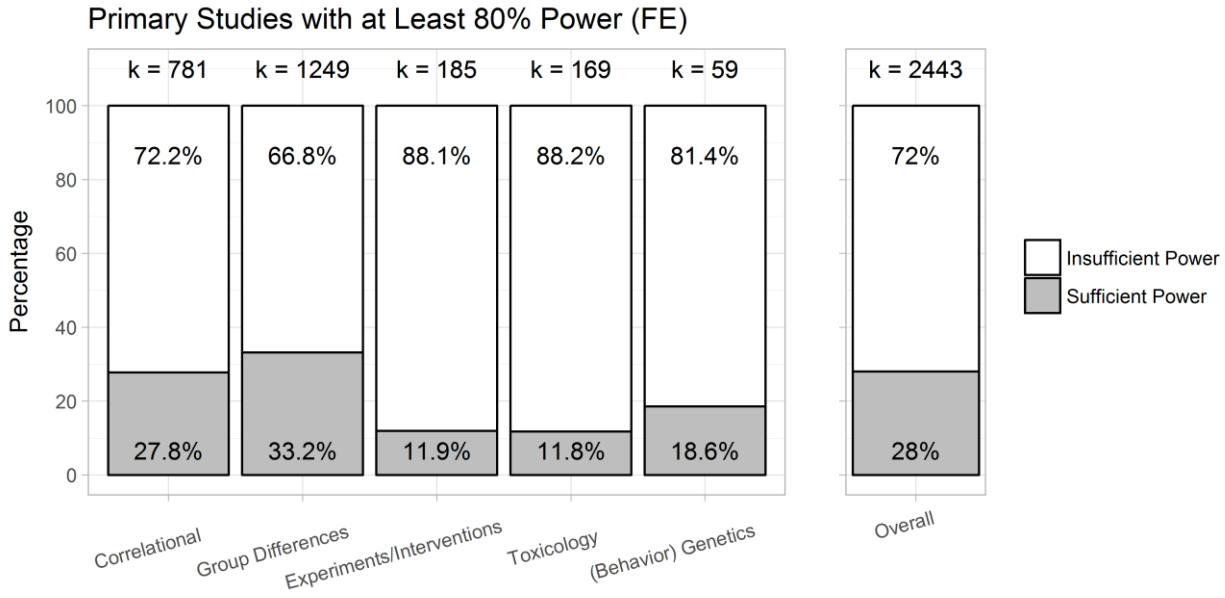


Figure 9.9

The percentage of primary studies in intelligence research with sufficient power (80% or more), split up per type of research methodology, and overall. The true effect sizes were approximated with fixed effect meta-analyses. The number of studies per type of research is indicated with the letter "k".

9.8.3 Overview Bias Patterns

Figure 9.10 shows the histograms of estimated meta-regression coefficients for the five different bias patterns. Contrary to Figure 9.7 in the main text, we did not truncate the x-axis, nor did we remove outliers.

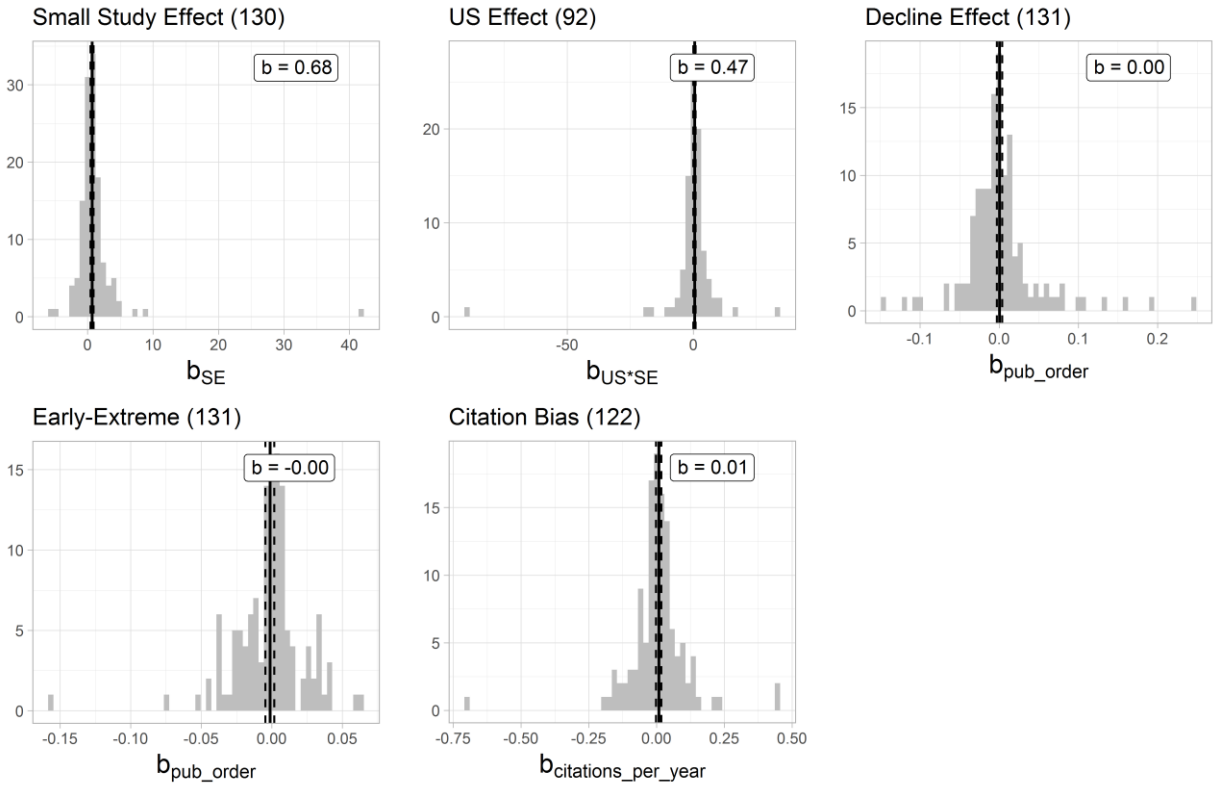


Figure 9.10

Histograms of estimated meta-regression coefficients for five different bias patterns. The vertical solid line indicates the meta-analytic weighted average of the coefficients, the estimate is also depicted in the plots. The dashed lines indicate the 95% confidence interval. Here, we did not truncate the x-axis, nor did we remove outliers. In the titles of the histograms, in parentheses, we indicated the number of meta-analyses for which we could estimate this bias.

9.8.4 Small study effect

9.8.4.1 *Heterogeneity*

In estimating the overall small study effect, there was moderate to high heterogeneity due to variance in true effects, I^2 = 46.5%, τ^2 = 0.71 (SE = 0.24, $Q(129)$ = 220.55, p < .001). We speculated that the heterogeneity might be caused by differences in the small study effect across different types of research. We therefore ran the meta-meta-regression over the regression coefficients again, but including type of study as a moderator. We found no evidence for any differences in the small study effects between different type of studies: $Q_M(4)$ = 1.05, p = 0.903.

9.8.4.2 *Small Study Effect per Type*

Even though type of study did not seem to moderate the small study effect, we still depict the small study estimates for the separate types in Figure 9.11 below. The estimates show that all five study types showed consistent evidence for a small study effect. We did not formally test these coefficients.

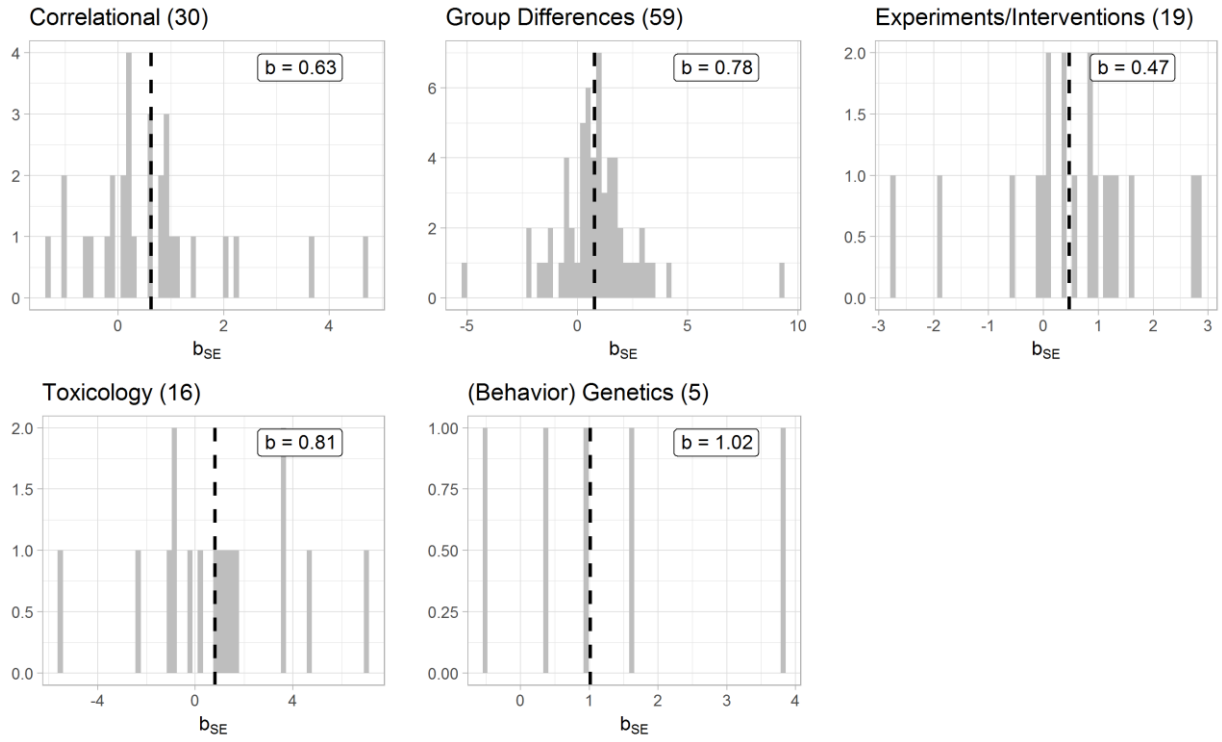


Figure 9.11

Histograms of estimated meta-regression coefficients for the small study effect, split up per research type. The vertical dashed line indicates the meta-analytic weighted average of the coefficients, the estimate is also depicted in the plots. In the titles of the histograms, in parentheses, we indicated the number of meta-analyses for which we could estimate this bias.

9.8.4.3 *Robustness Analysis: Only Include Meta-Analyses where k ≥ 10*

It is recommended to only test a small study effect if the meta-analysis includes ten or more studies, because otherwise the power of this test is too low (Sterne et al., 2000). With fewer than ten primary studies, results of the test for the small study effect for these meta-analyses are not very reliable if looked at individually. However, for our main analysis we did not focus at individual results, but at an aggregated estimate of all estimated small study effects. In this case, estimates from meta-analyses with very few primary studies will influence the overall result less, because they were weighted by their standard error.

As a robustness test, we estimated the overall small study effect again, excluding 50 meta-analyses with fewer than 10 primary studies. In one of the remaining 80 meta-analyses

we could again not estimate the effect because there was too little variation in the primary effect sizes and their standard errors. We now found 11 meta-analyses (13.8%) with significant evidence for a small study effect when retaining α = .05. This increased to 12 meta-analyses (15%) when we retained a less strict α of .10. When we summarized the 80 regression coefficients, we still found evidence for a small study effect across all included meta-analyses, b_{SE} = 0.61, SE = 0.13, Z = 4.59, p < .001, 95% CI = [0.35; 0.87].

9.8.5 US Effect

9.8.5.1 *Heterogeneity*

In estimating the overall US effect, we found very low heterogeneity due to variance in true effects, I^2 = 4.47%, τ^2 = 0.21 (SE = 0.75), $Q(91)$ = 95.50, p = .353.

9.8.5.2 *US Effect per Type*

We depict the US effect estimates for the separate study types in Figure 9.12 below. The estimates of the five study types show different patterns. In correlational research, research on group differences, and behavior genetics, the coefficients are in line with the US effect: the small study effects are stronger for US studies. In experiments and toxicological research, we found the opposite pattern: the small study effect was stronger for non-US studies. We did not formally test these coefficients.

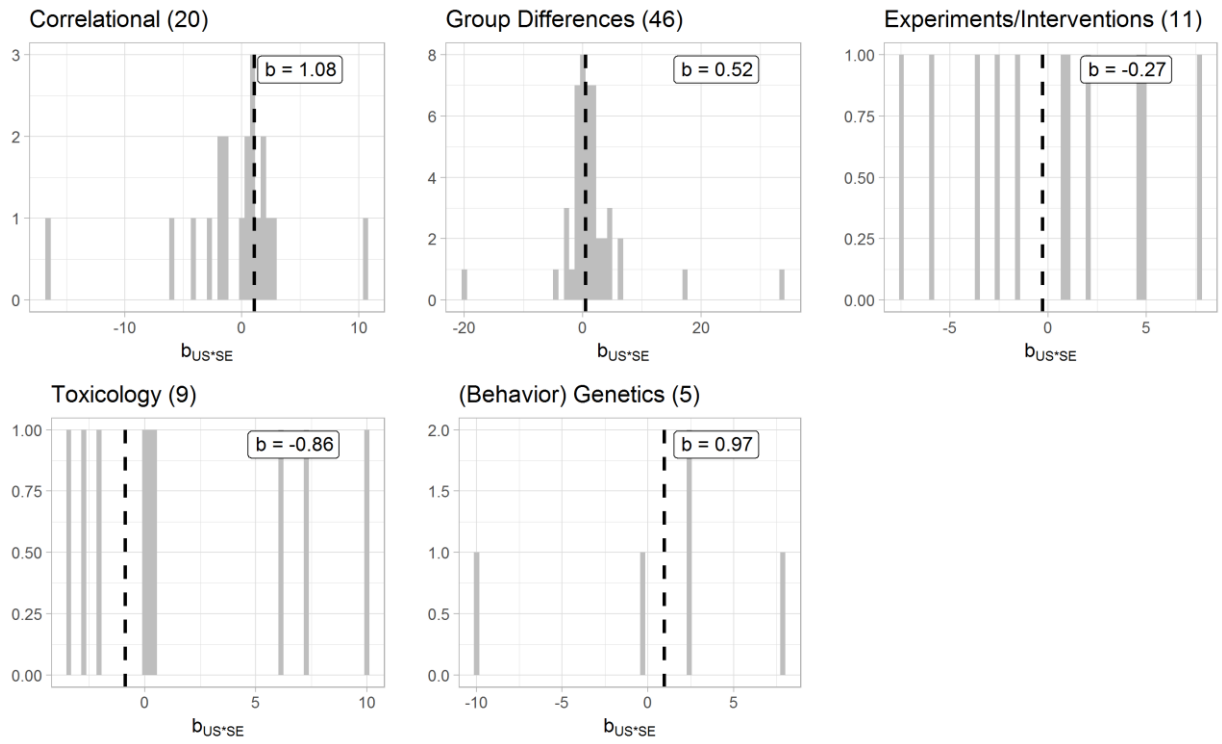


Figure 9.12

Histograms of estimated meta-regression coefficients for the US effect, split up per research type. The vertical dashed line indicates the meta-analytic weighted average of the coefficients, the estimate is also depicted in the plots. In the titles of the histograms, in parentheses, we indicated the number of meta-analyses for which we could estimate this bias.

9.8.6 Decline Effect

In our main analysis, we found no evidence for a decline effect across all meta-analyses. We ran several additional analyses to estimate the robustness of our estimate.

9.8.6.1 *Heterogeneity*

In estimating the overall decline effect, we found moderate to high heterogeneity due to variance in true effects, I^2 = 63.8%, τ^2 = 0.00 (SE = 0.00), $Q(130) = 118.17$, $p < .001$.

9.8.6.2 *Decline Effect per Type*

Study type seemed to be a moderator for the decline effect, Q_M (4) = 9.79, p = .044. We therefore looked at the overall decline effect for each type of research separately. Figure 9.13 shows all estimates of the decline effect for the five study types. We found some evidence for an overall decline effect in the 5 meta-analyses in behavior genetics ($b^j_{PubOrder}$ = -0.023, SE = 0.011, Z = -2.18, p = .029). Conversely, we found weak evidence for an "increase" effect in the 59 meta-analyses in which they analyzed group differences ($b^j_{PubOrder}$ = 0.003, SE

= 0.002, *Z* = 1.99, *p* = .011). Even though these effects reached formal significance, they were small.

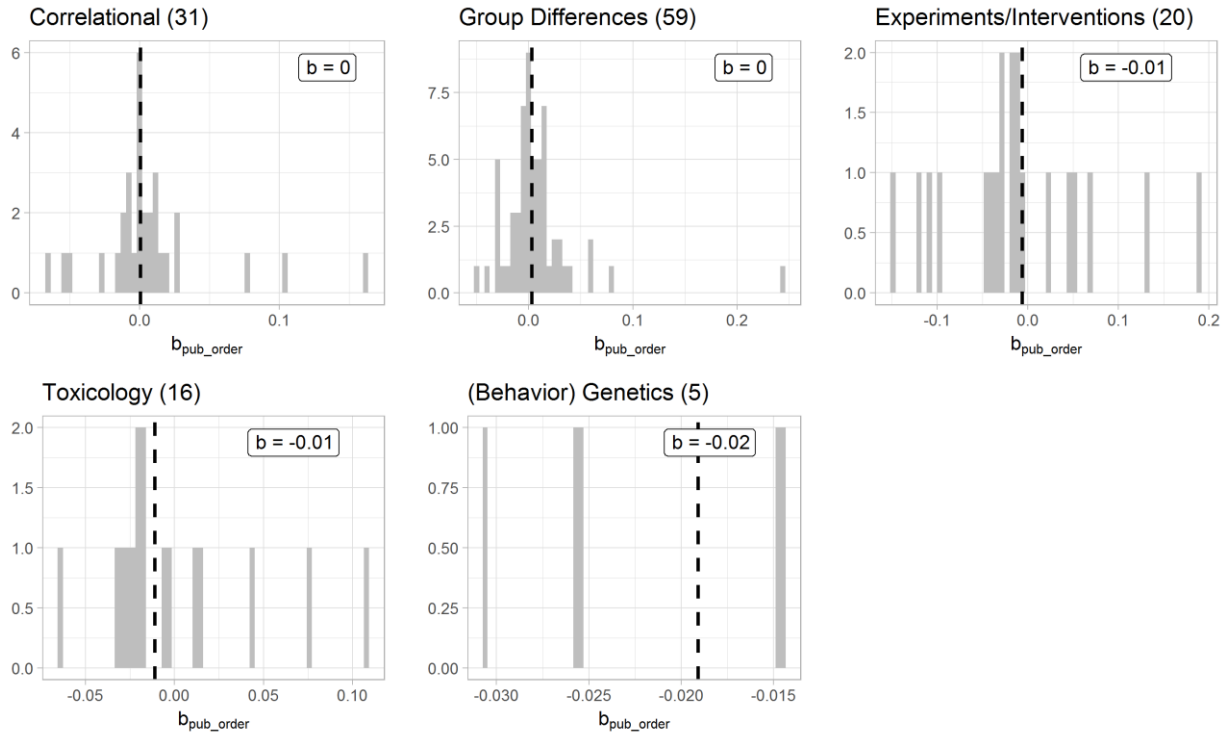


Figure 9.13

Histograms of estimated meta-regression coefficients for the decline effect, split up per research type. The vertical dashed line indicates the meta-analytic weighted average of the coefficients, the estimate is also depicted in the plots. In the titles of the histograms, in parentheses, we indicated the number of meta-analyses for which we could estimate this bias.

9.8.6.3 ***Sample Size over Time***

It is imaginable that the precision of the studies is positively related to the order in which the studies are published; the start of a line of research could be characterized by smaller, more explorative studies, and once an idea becomes more established, larger sample sizes are needed. We tested this notion by fitting the following regression model for each of the meta-analyses:

$$N_{ij} = a^j + b^j_{PubOrder} PublicationOrder_{ij} + \varepsilon_{ij},$$

Equation 9.2

where *N* indicates the total sample size of study *i* in meta-analysis *j*. A positive value for $b^j_{PubOrder}$ would indicate that larger studies are usually published later, relative to the rest of the studies in the meta-analysis.

We then wanted to summarize all obtained coefficients in a random effects meta-analysis to estimate if there was an overall trend in sample size with respect to publication

order. However, in several cases, the standard error of $b^j_{PubOrder}$ was so large, that the meta-meta-analysis did not converge. We therefore excluded all cases in which the standard error was larger than 1000, which excluded 4 meta-analyses. We summarized the remaining 127 coefficients, and found an increasing trend in sample size, $b^j_{PubOrder}$ = 3.73, SE = 0.93, Z = 3.99, p < .001, 95% CI [1.90; 5.56], suggesting that sample size generally indeed increased when research progressed in research lines in intelligence research. We found high heterogeneity due to variance in true effects, I^2 = 89.5%, τ^2 = 43.97 (SE = 13.95), $Q(126)$ = 226.23, p < .001, which implies that this trend likely does not generalize to all research lines in intelligence research. We estimated the model again with type as a moderator to see if that would explain the heterogeneity, but we did not find evidence corroborating this notion, $Q_M(4)$ = 1.5695, p = .814.

9.8.6.4 *Decline Effect, Controlled for SE*

To control for our finding that sample size increased over time, we tested the decline effect again for each of the meta-analyses, but this time we added standard error of the study as a control variable to the meta-regressions:

$$Fisher's\ Z_{ij} = a^j + b^j_{PubOrder} PublicationOrder_{ij} + b^j_{SE} SE_{ij} + \varepsilon_{ij},$$

Equation 9.3

where SE_{ij} indicates the standard error of the effect size in study i in meta-analysis j. Again, the coefficient of interest was $b^j_{PubOrder}$, which, when negative, reflects a decline effect controlled for the precision of the study.

In two meta-analyses the model could not be fit. This time, in only three meta-analyses (2.3%) we found statistically significant evidence for a decline effect, compared to the six significant cases (4.6%) we found when we did not control for SE. We used a random effects meta-analysis to summarize all 129 obtained $b^j_{PubOrder}$ coefficients, and again found no overall evidence for a decline effect, $b^j_{PubOrder}$ = 0.001, SE = 0.002, Z = 0.606, p = .545, 95% CI = [-0.003; 0.006]. Again, the heterogeneity due to variance in the true effects was high, I^2 = 80.2%, τ^2 = 0.0003 (SE = 0.0001), $Q(128)$ = 273.81, p < .001. We ran the random effects meta-analysis on $b^j_{PubOrder}$ again, and included a moderating effect for type. Type did not seem to be a moderator, $Q_M(4)$ = 8.60, p = .072, so we did not continue estimating the decline effect for the different types of meta-analyses separately.

9.8.6.5 *Winner's Curse*

9.8.6.5.1 Heterogeneity

In estimating the overall winner's curse, we found moderate heterogeneity due to variance in true effects, I^2 = 32.5%, τ^2 = 0.004 (SE = 0.003), $Q(129)$ = 157.63, p = .044. We speculated that this heterogeneity may have been caused by the different types of research, so we ran the meta-meta-regression again, including type as a moderator. We found that type

did not seem to be a moderator, $Q_M(4) = 6.99$, $p = .136$, so we did not investigate the winner's curse for the different types of meta-analyses separately.

9.8.6.5.2 Winner's Curse, Controlled for SE

As we speculated above, it is imaginable that the first studies in a line of research are systematically smaller than subsequent studies. To control for this possibility, we tested the winner's curse again for each of the meta-analyses, but this time we added standard error of the study as a control variable to the meta-regressions:

$$Fisher's\ Z_{ij} = a^j + b^j_{FirstPublished} FirstPublished_{ij} + b^j_{SE} SE_{ij} + \varepsilon_{ij},$$

Equation 9.4

where SE_{ij} indicates the standard error of the effect size in study i in meta-analysis j. Again, the coefficient of interest was $b^j_{FirstPublished}$, which, when positive, reflects a winner's curse controlled for the precision of the studies.

In four meta-analyses the model could not be fit. This time, we found evidence for a winner's curse in nine meta-analyses (7.1%). We used a random effects meta-analysis to summarize all 127 obtained $b^j_{FirstPublished}$ coefficients, and this time we did find weak overall evidence for a winner's curse, $b^j_{FirstPublished} = 0.032$, SE = 0.016, $Z = 1.99$, $p = .047$, 95% CI = [0.000; 0.063]. Even though this finding is formally significant at $\alpha = .05$, we did not correct for multiple testing and this result needs to be interpreted with care.

Again, when fitting the winner's curse, we found moderate heterogeneity due to variance in the true effects, $I^2 = 32.2\%$, $\tau^2 = 0.0072$ (SE = 0.0041), $Q(126) = 163.58$, $p = .014$. We ran the random effects meta-analysis on $b^j_{FirstPublished}$ again, and included a moderating effect for type. We found no evidence that study type was a moderator, $Q_M(4) = 2.74$, $p = .602$, so we did not continue estimating the winner's curse for the different types of meta-analyses separately.

9.8.6.6 *Decline Effect: 1/Order*

Instead of using publication order as predictor, we also estimated the decline effect with 1/(publication order). This reflects a nonlinear decline of effect sizes over time, with the decline decreasing in publication order. This is a less extreme version of the winner's curse.

In seven of the 131 meta-analyses (5.3%) we found evidence for a decline effect, if we take 1/(publication order) as a predictor. This analysis did show a stronger decline effect than when taking "publication order" as a predictor, $b^j_{1/(publication\ order)} = 0.018$, SE = 0.018, $Z = 1.00$, $p = .317$, 95% CI = [-0.018; 0.054], but the effect was small. In fitting the decline effect, we found moderate heterogeneity, $I^2 = 44.5\%$, $\tau^2 = 0.105$ (SE = 0.005), $Q(130) = 174.46$, $p = .006$. We ran the random effects meta-analysis on $b^j_{1/(publication\ order)}$ again, and included a moderating effect for type. We found no evidence that study type moderated the effect, $Q_M(4) = 6.26$, p

= .180, so we did not continue estimating the decline effect for the different types of meta-analyses separately.

9.8.7 Early-Extremes Effect

9.8.7.1 *Heterogeneity*

In estimating the overall early-extremes effect, we found high heterogeneity due to variance in true effects, I^2 = 89.4%, τ^2 = 0.00 (SE = 0.00), $Q(130) = 300.38$, $p < .001$. We ran the random effects meta-analysis on $b^j_{FirstPublished}$ again, and included a moderating effect for type. We found no evidence that study type moderated the early-extremes effect, $Q_M(4) = 2.45$, $p = .654$.

9.8.7.2 *Early-Extreme per Type*

Even though we found no evidence that type of study moderated the early-extremes effect, we still depict the early-extremes estimates for the separate types in Figure 9.14 below. In all five study types, the estimates of the early-extremes effect were close to zero. We did not formally test these coefficients.

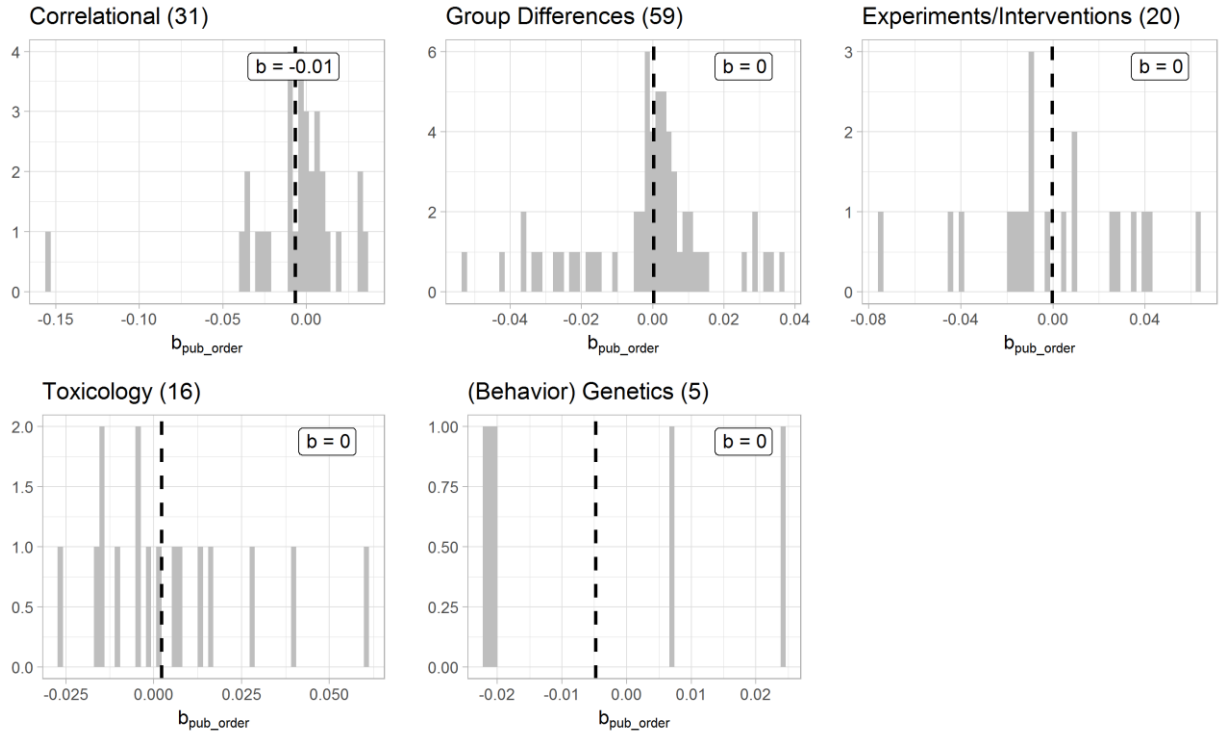


Figure 9.14

Histograms of estimated meta-regression coefficients for an early-extreme effect, split up per research type. The vertical dashed line indicates the meta-analytic weighted average of the coefficients, the estimate is also depicted in the plots. In the titles of the histograms, in parentheses, we indicated the number of meta-analyses for which we could estimate this bias.

9.8.7.3 *Early-Extreme, Controlled for SE*

As in the other analyses we also controlled for any possible influence of the standard error of the primary studies on their effect size, so we reran all 131 meta-regressions again with primary study standard error as a control variable:

$$deviation = a^j + b^j_{PubOrder} PublicationOrder_{ij} + b^j_{SE} SE_{ij} + \varepsilon_{ij}.$$

Equation 9.5

In two cases, the model did not converge. Of the 129 estimated coefficients $b^j_{PubOrder}$, four (3.1%) were significantly smaller than zero, indicating an early-extreme effect in those meta-analyses, when controlling for standard error.

We summarized the 129 $b^j_{PubOrder}$ coefficients using a random effects meta-analysis. We still found no overall evidence for an early-extreme effect, $b^j_{PubOrder}$ = 0.004, SE = 0.003, Z = 1.30, p = .194, 95% CI = [-0.002; 0.011]. Heterogeneity due to variance in true effects was very high, I^2 = 97.6%, τ^2 = 0.001 (SE = 0.000), $Q(128)$ = 39537.55, p < .001. This heterogeneity could not be explained by type of research, $Q_M(4)$ = 6.77, p = .149.

9.8.7.4 *Early-Extremes Effect: 1/Order*

Instead of using publication order as predictor, we also estimated the early-extremes effect with 1/(publication order). In six of the 131 meta-analyses (4.6%) we found evidence for an early extremes effect, if we take 1/(publication order) as a predictor. We found a stronger early-extremes effect than when taking "publication order" as a predictor, $b^j_{1/(publication\ order)}$= 0.009, SE = 0.013, Z = 0.726, p = .468, 95% CI = [-0.015; 0.033], but the effect was still small. Again, the heterogeneity due to variance in the true effects was high, I^2 = 75.4%, τ^2 = 0.008 (SE = 0.003), $Q(130)$ = 312.53, p < .001. We ran the random effects meta-analysis on $b^j_{1/(publication\ order)}$ again, and included a moderating effect for type. We found no evidence that type was a moderator, $Q_M(4)$ = 4.85, p = .303, so we did not continue estimating the decline effect for the different types of meta-analyses separately.

9.8.8 Citation bias

9.8.8.1 *Heterogeneity*

In estimating overall citation bias, we found moderate heterogeneity due to variance in true effects, I^2= 58.3%, τ^2 = 0.00 (SE = 0.00), $Q(125)$ = 167.14, p = .007. This heterogeneity could not be explained by type of research, $Q_M(4)$ = 1.80, p = .773.

9.8.8.2 *Citation Bias per Type*

Even though type of study did not seem to moderate the effects of citation bias, we still depict the citation bias estimates for the separate types in Figure 9.15 below. In all five

study types, the estimates of citation bias were close to zero. We did not formally test these coefficients.

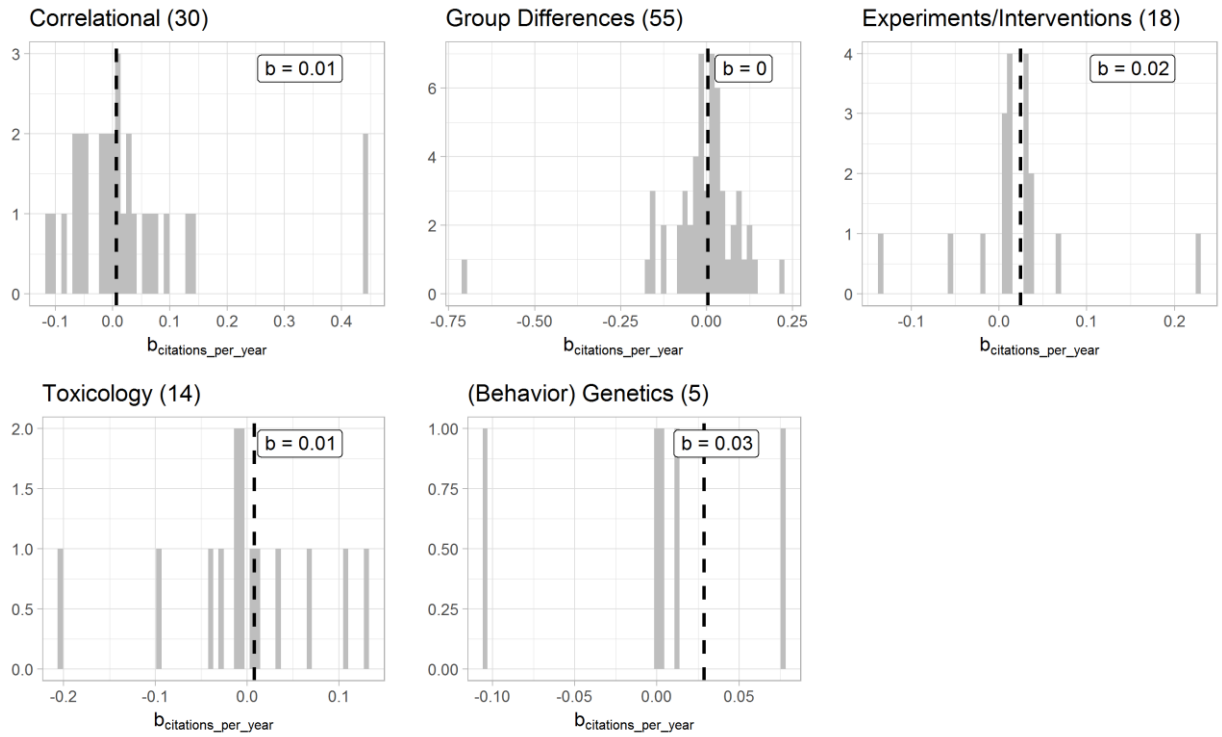


Figure 9.15

Histograms of estimated meta-regression coefficients for citation bias, split up per research type. The vertical dashed line indicates the meta-analytic weighted average of the coefficients, the estimate is also depicted in the plots. In the titles of the histograms, in parentheses, we indicated the number of meta-analyses for which we could estimate this bias.

9.8.8.3 ***Control Variables***

Even though we did not find overall evidence for citation bias, we ran several additional analyses with theoretically relevant control variables.

9.8.8.3.1 Citation + SE

Jannot et al. (2013) found that citation bias decreased when sample size was added as a control. We therefore tested the relation between citations and effect size again, and added SE as a control variable:

$$Fisher's\ Z_{ij} = a^j + b^j_{CitPerYear} CitationsPerYear_{ij} + b^j_{SE} SE_{ij} + \varepsilon_{ij}$$

Equation 9.6

where $b^j_{CitPerYear}$ indicated the number of citations per year of study *i* in meta-analysis *j*. A positive coefficient would indicate that studies with larger effects are cited more often, controlled for standard error. In 4 of the 123 meta-analyses (3.3%) in which we could estimate

this interaction, we found a significant relation between effect size and citations per year, controlled for standard error.

We summarized all 123 coefficients with a random-effects meta-analysis, and we found no overall relation between effect size and citations per year, controlled for standard error, $b^j_{CitPerYear}$ = 0.002, SE = 0.001, Z = 1.55, p = .121, 95% CI = [-0.001; 0.004]. Heterogeneity was moderate to high, I^2= 69.4%, τ^2 = 0.00 (SE = 0.00), Q(122) = 173.34, p = .002. This heterogeneity could not be explained by type of research, Q_M(4) = 2.15, p = .071.

9.8.8.3.2 Citation * SE

We were also interested in whether researchers would cite studies more or less often if the study contained an effect size likely to be overestimated. If science is truly self-correcting, small studies with effects that are likely to be overestimated should have less impact. Therefore, we investigated whether there was a positive interaction effect between SE and citations on effect size (the relation between citations and effect size becomes stronger if SE is large), by testing the following regression model:

$$Fisher's\ Z_{ij} = a^j + b^j_{CitPerYear} CitationsPerYear_{ij} + b^j_{SE} SE_{ij} + b^j_{Cit*SE} CitationsPerYear_{ij} * SE_{ij} + \varepsilon_{ij},$$

Equation 9.7

where b^j_{Cit*SE} indicated an interaction effect between the number of citations per year and the standard error of a study on its primary effect. A positive coefficient would indicate that studies that show a small study effect are cited more often.

In 6 of the 114 meta-analyses (5.3%) in which we could estimate this interaction, we found a significant interaction between standard error and citations per year on effect size. In four of these cases, the interaction was positive, indicating that researchers *more* often cited overestimated effects, and in two cases it was negative, indicating that researchers *less* often cited overestimated effects.

We summarized all 114 coefficients with a random-effects meta-analysis, and we found no overall interaction effect between SE and citations, b^j_{Cit*SE} = -0.012, SE = 0.026, Z = -0.447, p = .655, 95% CI = [-0.063; 0.040]. Heterogeneity due to true variance in effects was low, I^2= 8.23%, τ^2 = 0.005 (SE = 0.011), Q(113) = 120.58, p = .295, so we did not run this analysis again split up per type.

9.8.8.4 *Impact Factor Bias*

It has been suggested that impact factor of a journal could be a mediating variable in the effect of effect size on citations (Jannot et al., 2013). To investigate this, we ran several additional analyses. We coded the impact factor of the journal in which the primary study was published (coded in March 2015). This information was extracted from Web of Knowledge.

9.8.8.4.1 Impact Factor

First, we investigated whether studies with larger effect sizes generally get published in higher impact journals? To investigate this, we estimated the following meta-regression for each of the meta-analyses:

$$Fisher's\ Z_{ij} = a^j + b^j_{IF} ImpactFactor_{ij} + \varepsilon_{ij},$$

Equation 9.8

where a positive value for b^j_{IF} indicated a positive relation between effect size and impact factor, which would be evidence for "impact bias". In 10 cases, the model could not be fit. In 4 of the 121 remaining meta-analyses, we found evidence for impact bias. Unfortunately, we were unable to summarize all 121 b^j_{IF} coefficients due to convergence problems (the ratio largest to smallest sampling variance was extremely large).

9.8.8.4.2 Impact Factor + SE

We also investigated whether there was an effect of impact factor on effect size, controlled for precision, by estimating the following meta-regression:

$$Fisher's\ Z_{ij} = a^j + b^j_{IF} ImpactFactor_{ij} + b^j_{SE} SE_{ij} + \varepsilon_{ij},$$

Equation 9.9

where a positive value for b^j_{IF} indicated a positive relation between effect size and impact factor, controlled for precision of the study. In 16 cases, the model could not be fit. In 5 meta-analyses (4.3%), we found a positive effect of impact factor on effect size, controlled for standard error. Again, we were not able to summarize all 116 regression coefficients due to convergence problems (the ratio largest to smallest sampling variance was extremely large).

9.8.8.4.3 Impact Factor * SE

We also investigated if high impact journals were more or less likely to contain studies that show evidence for the small study effect, than low impact journals. To that end, we estimated the following regression model:

$$Fisher's\ Z_{ij} = a^j + b^j_{IF} ImpactFactor_{ij} + b^j_{SE} SE_{ij} + b^j_{IF*SE} ImpactFactor_{ij} * SE_{ij} + \varepsilon_{ij},$$

Equation 9.10

in which a positive value for b^j_{IF*SE} indicates that studies that have evidence for a small study effect are more likely to be published in a higher impact journal.

We were able to fit this model in 107 meta-analyses, and in 9 of them (8.4%) we found a significant interaction effect between impact factor and standard error, 5 of which were positive, and 4 of which were negative. When we summarized all obtained regression coefficients, we found no overall evidence for an interaction effect between impact factor and standard error, b^j_{IF*SE} = -0.046, SE = 0.073, Z = -0.623, p = .533, 95% CI = [-0.188; 0.098]. Heterogeneity was low to moderate, I^2= 42.1%, τ^2 = 0.13 (SE = 0.08), $Q(106)$ = 135.51, p = .028. This heterogeneity could not be explained by type of research, $Q_M(4)$ = 3.50, p = .478.

9.8.8.4.4 Impact Factor + Citations

If impact factor mediates the relation between citations and effect size, then we expect that the relation between citation and effect size decreases when controlling for impact factor. We tested this notion with the following regression:

$$Fisher's\ Z_{ij} = a^{j} + b^{j}_{CitPerYear} CitationsPerYear_{ij} + b^{j}_{IF} ImpactFactor_{ij} + \varepsilon_{ij},$$

Equation 9.11

where a positive value for $b^{j}_{CitPerYear}$ indicates evidence for citation bias, controlled for impact factor. We were able to fit this model in 115 meta-analyses, and in 9 of them (7.8%) we found a positive effect of citations on effect size, controlled for impact factor. Again, we were not able to summarize all 115 regression coefficients due to convergence problems (the ratio largest to smallest sampling variance was extremely large).

9.8.8.5 *Conclusion Impact Effects*

We were interested in seeing if metrics related to impact were related to effect size, and possibly to overestimated effects. We found no consistent evidence that larger effect sizes were cited more often, or were published in higher impact journals. Furthermore, we found no evidence that studies that may contain overestimated effects were cited more or less often, or published in higher or lower impact journals.

Chapter 10

Discussion Part II

In Part II of this dissertation, we focused on bias in meta-analysis. In many scientific fields there is evidence that published effect sizes are overestimated, which leads to biased estimates in meta-analyses (Button et al., 2013; Ioannidis, 2008). Here, our goal was to get a deeper understanding of how publication bias and other types of bias affect effect size estimates in general, and in the field of intelligence research in particular.

Among methodologists, there is increasing consensus that publication bias leads to an excess of false positive findings in the literature (Francis, 2012a; Ioannidis & Trikalinos, 2007). However, it has often been suggested that these false positives in the literature will eventually be corrected via replication (Crocker & Cooper, 2011; Diekmann, 2011; Murayama et al., 2013). Indeed, self-correction is one of the main characteristics of science. This line of thought is also reflected in our survey results in Chapter 7. Here, psychology students, social scientists, and quantitative psychologists almost unanimously showed the same intuition: if your goal is to accurately estimate an effect, you should always include as many published (replication) studies as possible in your estimation. However, we showed that in the current publication system, this intuition is false in many circumstances.

In Chapter 7, we found that under many circumstances replication studies may actually worsen the accuracy effect size estimates. Given the recent focus in the psychological literature on the merits of replication (see, e.g., Pashler & Wagenmakers, 2012), this is a counterintuitive finding. Ironically, this counterintuitive finding is caused by one of the very phenomena that replication is trying to correct: overestimated effects because of publication bias. Many replication studies are often not explicitly identified as such (Makel et al., 2012), and it is likely that they are affected by publication bias in the same way as original studies. Because of this, both original studies and their replications are likely to contain overestimated effects, and this problem worsens when the studies become smaller and publication bias worsens. This leads to a situation in which it may be better to discard small, underpowered (replication) studies completely and only focus on large, more precise ones (Kraemer, Gardner, Brooks, & Yesavage, 1998; Stanley, Jarrell, & Doucouliagos, 2010; but see also Borm, den Heijer, & Zielhuis, 2009; IntHout, Ioannidis, & Borm, 2016). We found that discarding small studies goes against psychologists' intuitions. In a survey, we asked psychology students, social scientists, and quantitative psychologists to choose which combination of small and large published studies would render the most accurate effect size. We found that, regardless of the level of statistics training, the respondents almost unanimously chose the scenarios with the most studies, even if they were small. Previous research already showed that people tend to overestimate the informational value of small studies (Bakker et al., 2016; Tversky & Kahneman, 1971), but in combination with publication bias the informational value decreases further and may even be negative.

Chapter 7 illustrates the severity of the problems publication bias can cause. We therefore think it is important to try and identify how much meta-analyses are affected by

publication bias. It might even be possible to identify characteristics of primary studies that mark them as "high-risk" to contain overestimated effects. This is wat we set out to do in Chapters 8 and 9, in which we analyzed large sets of meta-analyses to detect overall patterns of bias.

In Chapter 8, we reanalyzed data from a study that investigated 82 meta-analyses for patterns of bias (Fanelli & Ioannidis, 2013). In our reanalysis, we replicated Fanelli and Ioannidis' finding of a small study effect: smaller studies in meta-analyses tended to find larger effect sizes. This is a potential sign of publication bias. In these data, we failed to replicate Fanelli and Ioannidis' finding that US studies show stronger overestimation than non-US studies (but see Fanelli & Ioannidis, 2014). This finding casts doubt on the robustness of the US effect.

In Chapter 9, we analyzed 131 meta-analyses about intelligence, covering 2,439 studies and over 20 million participants, to see if the field of intelligence research is likely to be affected by different types of bias. First, we found that the power in this field is generally low, albeit not as low as has been documented in other psychological fields. We estimated that the median power across intelligence research is 48.8%, ranging from 10.5% in behavior genetics studies to 51.9% in research on group differences in intelligence. Less than a third of all studies (28.0%) reached the recommended power of 80% or more.

Even though the power in intelligence research is much lower than the recommended 80%, it is still higher than in neuroscience (8-31%; Button et al., 2013), psychology (between 12% and 44%; Szucs & Ioannidis, 2017; Stanley et al., 2017), behavioral ecology and animal research (13–16% for a small effect and 40–47% for a medium effect; Jennions & Moller, 2003), and economics (18%; Ioannidis et al., 2017). One potential reason for at least some of these discrepancies, is that the median sample size in intelligence research (N = 60) was higher than in cognitive neuroscience and psychology (N = 20 - 40; Marszalek et al., 2011; Szucs & Ioannidis, 2017; Wetzels et al., 2011). Another reason could be that the typical effect size in intelligence (*r* = .26) is slightly higher than the average effect size across the social sciences (r = .21; Richard, Bond, & Stokes-Zoota, 2003). Finally, measures in intelligence research typically have relatively high reliability, which could also explain the higher power compared to other fields (Hunt, 2010; Mackintosh, 2011; Plomin et al., 2016; Ritchie, 2015). However, even though the power in intelligence seems higher than in other fields, it is generally still much lower than the recommended 80%.

The finding that power in intelligence research is generally low is worrying. First, low power increases the risk of a false negative, but also the probability that a significant finding is a false positive (Button et al., 2013; Ioannidis, 2005). Furthermore, effects are estimated with low precision in poorly powered studies, and can be strongly under- and overestimated. In a small, underpowered study, for an effect to reach statistical significance, it has to be very large. That means that if only significant studies are published, the inflation of published

effects increases (Button et al., 2013; Kraemer et al., 1998; Nuijten et al., 2015). To make matters worse, researchers are incentivized to report significant findings, and they can strategically use researcher degrees of freedom in their analyses to "push" a non-significant finding towards significance (John et al., 2012; Simmons et al., 2011). Low power increases not only the chance that researchers fail to find a significant effect but also the likelihood that they will use these researcher degrees of freedom. This will lead to higher false positive rates and overestimation of genuine effects (Bakker et al., 2012; van Aert, Wicherts, & van Assen, 2016).

The low power in intelligence research may worsen the effect of any biases that inflate effect sizes. We investigated whether there were patterns in effect sizes that may indicate potential biases. Specifically, we focused on the small study effect, US effect, decline effect, early-extremes effect, and citation bias. In line with results from Chapter 8, we found evidence for a small study effect across the intelligence literature; smaller studies generally found larger effects, which could potentially indicate publication bias. We did not find consistent evidence that the small study effect was worse for studies from the US than for non-US studies. This finding is not in line with the notion that overestimation is worse in US studies (Doucouliagos et al., 2005; Fanelli & Ioannidis, 2013; Munafò et al., 2008). Both the results in Chapter 8 and 9, and in Fanelli et al. (2017) show that the US effect is not robust for different analytical choices. Overall, we conclude that evidence for a US effect is weak and inconsistent.

Furthermore, we did not find evidence that effect sizes within a meta-analysis decreased in size over time, or that earlier studies showed more extreme opposing effects, which is evidence against a decline effect and early-extremes effect. We also did not find evidence for citation bias: larger effects were not systematically cited more often than smaller effects. These findings suggest that intelligence research is less susceptible to these biases than other scientific fields (Fanelli et al., 2017).

An important limitation of the study in Chapter 9 is that we analyzed bias-related patterns across all meta-analyses. This means that even though we found an overall small study effect, it does not mean that every meta-analysis in intelligence research shows this pattern. Conversely, the failure to find overall evidence for a US effect, decline effect, early-extreme effect, or citation bias across all meta-analyses, does not mean that none of the meta-analyses showed these biases. Furthermore, it is important to note that the presence of a small study effect does not necessarily have to signify bias. A small study effect can also occur when smaller studies typically investigate larger true effects. This happens when researchers run a priori power analyses that show that they only need a small sample to detect the expected effect, or when researchers learn through experience how big their samples should be to find the effect of interest. However, there is evidence that researchers often do not base sample size decisions on formal power analyses (Bakker et al., 2016; Tressoldi & Giofre, 2015; Vankov et al., 2014). Furthermore, it is possible that researchers deliberately run several

studies with small samples instead of one large study, to increase the probability of finding at least one significant finding (Bakker et al., 2012).

10.1 Solutions

In Part II of this dissertation we confirmed previous findings that publication bias can have a severe effect on meta-analytic effect size estimates. What is more, Chapters 8 and 9 illustrate that this is not only a theoretical problem; we corroborated a long line of evidence that publication bias and overestimated effects seem to be widespread in psychology (see, e.g., Button et al., 2013; Fanelli et al., 2017; Fanelli & Ioannidis, 2013; Niemeyer et al., 2012, 2013). This is a worrying finding, because it decreases the validity of the published literature. We need to look for solutions that allow us to determine which effects are likely to be overestimated and correct for these overestimations. We also need to think about ways to prevent these problems in future literature.

We found that the effects of bias worsen when power is low. Therefore, one potential solution against overestimated effects, is to only evaluate studies with large sample size (and therefore high precision) when estimating an effect size (Bakker et al., 2012; Kraemer et al., 1998; Stanley et al., 2010). Relatedly, psychologists should also try to achieve higher power in the studies they conduct, whether it is a replication or not (Asendorpf et al., 2013; Brandt et al., 2014). A problem here, is that power analyses to determine sample size are often based on previously published effect sizes. These effect sizes are likely overestimated, which means that the sample size will be underestimated. A solution is to correct the observed effect sizes (Anderson, Kelley, & Maxwell, 2017; Etz & Vandekerckhove, 2016; van Aert & van Assen, 2017a, 2017b; van Assen et al., 2015; Vevea & Hedges, 1995), calculate lower bound power (Anderson et al., 2017; Perugini et al., 2014), or base a power analysis on the smallest effect size of interest (Ellis, 2010).

Another way to decrease the prevalence of overestimated effects, is to eliminate publication bias altogether (van Assen, van Aert, et al., 2014a). It is widely recognized that publication bias is harmful, and many people have suggested potential solutions. For instance, some journals specifically state to evaluate submissions only based on theory and methodology, rather than on results (e.g., *PLOS ONE*). Relatedly, it has been suggested that the decision to accept or reject a paper should only be based on the Introduction and Methods sections of a paper (De Groot, 1956/2014; Smulders, 2013; Walster & Cleary, 1970). These solutions should avoid publication bias in editorial decisions. However, there is evidence that publication bias also arises because of the authors themselves: authors are less likely to submit a paper if they did not find significant results (Cooper et al., 1997; Dickersin et al., 1987; Franco et al., 2014; Shadish, Doherty, & Montgomery, 1989).

One promising implementation of ignoring results when evaluating manuscripts is registered reports (Chambers, 2013). Here, authors submit a study design and analysis plan

(pre-registration) to a journal, *before* the study is conducted. This plan is reviewed and improved if needed. If the preregistered study meets theoretical and methodological standards, authors can receive an "in principle acceptance". This means that if they conduct the study as described in the plan, the study will be published regardless of the results. Here, it is important that the registration and the final paper are compared closely to see if the authors indeed adhered to the original plan, and if they did not, if they reported any deviations transparently (Chan & Altman, 2005). Another major advantage of registered reports, is that they prevent the use of researcher degrees of freedom. First, the reason to make use of such practices is taken away when the editorial decision does not depend on the results. Second, the analysis plan is preregistered, which means that any deviation from that plan can explicitly be pointed out as explorative.

A way to deal with publication bias in literature that is already published, it is to analyze meta-analyses for evidence of publication bias (Rothstein et al., 2005). Many different procedures exist (see, e.g., Begg & Mazumdar, 1994; Egger et al., 1997; Guan & Vandekerckhove, 2015; Ioannidis & Trikalinos, 2007), and it is recommended to run several of them as a robustness analysis (Banks et al., 2012). However, the power of these methods is generally low, so the absence of significant evidence for publication bias does not necessarily mean that there is no publication bias (see also Chapter 7). There are also procedures aimed at estimating effect sizes that are either robust against publication bias or correct for it (Copas, 2013; Duval & Tweedie, 2000a, 2000b; Hedges, 1984; Hedges & Vevea, 1996, 2005; Simonsohn et al., 2014; Stanley & Doucouliagos, 2014; van Assen et al., 2015; Vevea & Hedges, 1995; Vevea & Woods, 2005). Note that these methods are not without criticism. They often have strong assumptions, they do not always perform well in the presence of heterogeneity in true effects, and they cannot deal with the effects of researcher degrees of freedom. We still need more research to improve corrections for overestimated effects due to publication bias.

To conclude, overestimated effects seem widespread in psychology, and publication bias seems to be a major cause. Solving this problem probably requires a complex combination of factors, but several steps in the right direction can already be taken. When evaluating existing research, we recommend performing meta-analyses in combination with tests for publication bias. The limitations of such publication bias tests should be taken into account, and it is advisable to perform several sensitivity analyses, and to interpret the results of meta-analyses with care. For future research, we recommend increasing power by increasing sample size and/or reliability. We also recommend preregistration, and particularly endorse registered reports, to avoid both publication bias and prevent researcher degrees of freedom.

Chapter 11

Epilogue

In this dissertation, I focused on two problems in the psychological literature: statistical reporting inconsistencies and bias in effect size estimates. In Part I, we corroborated previous findings that reporting inconsistencies are widespread in the psychological literature, and that these inconsistencies are often in favor of reporting a significant effect. This bias towards significance is also reflected in the findings of Part II, where we concluded that psychology seems to suffer from low power in combination with publication bias, resulting in overestimated effects.

We suggested some potential solutions for the problems we highlighted in this dissertation. To detect, prevent, and correct statistical reporting inconsistencies, we developed the R package "statcheck" (Epskamp & Nuijten, 2016). Statcheck automatically extracts APA reported statistics and checks the internal consistency. We consider statcheck a useful "spellchecker" for statistics in psychology papers. We recommend the use of statcheck to check manuscripts before submissions, and to check submitted manuscripts in peer review. This recommendation is already being taken up by two major journals in the field. To solve the problem of overestimated effects, we recommend increasing power, and eliminating publication bias and researcher degrees of freedom. Especially the latter is easier said than done, but promising strategies entail two-step peer review, preregistered reports, and publication bias tests (see Chapters 7, 9, and 10 for details).

Unfortunately, reporting inconsistencies and bias in meta-analyses are only two of the factors that threaten the reliability and validity of psychological science. Many researchers have pointed out additional problems. For instance, there are strong indications that the impossibly high number of significant findings in the literature is not only caused by publication bias, but also by the strategic use of flexibility in data analysis, also referred to as "researcher degrees of freedom" (Bakker et al., 2012; Gelman & Loken, 2014; Simmons et al., 2011). Trying out several strategies, and with that running a large number of exploratory analyses, almost guarantees you to find at least one significant result. However, this may be a Type I error and likely results in an overestimated effect size. Running exploratory analyses is by no means bad practice, but it is highly misleading to present exploratory results as having been predicted from the start, which can be considered "HARKing"; hypothesizing after the results are known (Kerr, 1998; Wagenmakers et al., 2012). This problem worsens if only the statistically significant exploratory results are presented. This is a practice that a substantial proportion of psychologists admitted to (Agnoli et al., 2017; John et al., 2012; see also http://psychdisclosure.org). This practice has also been observed directly, when comparing published and unpublished results in a known population of conducted studies (Chan, Hrobjartsson, Haahr, Gotzsche, & Altman, 2004; Franco et al., 2016; O'Boyle, Banks, & Gonzalez-Mule, 2017).

11.1 Preregistration

One promising way to avoid both publication bias and researcher degrees of freedom is preregistration (Asendorpf et al., 2013; Chambers & Munafò, 2013; De Groot, 1956/2014; Nosek et al., 2015; Nosek & Bar-Anan, 2012; Nosek et al., 2012; Wagenmakers et al., 2012). When preregistering a study, researchers specify all planned aspects of their study, such as the study design, sampling plan, and analysis plan, before the data are collected. Over 80 journals are now even offering the possibility to submit registered reports (e.g., Chambers, 2013; Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Jonas & Cesario, 2016; Nosek & Lakens, 2014; for a full list of participating journals, see https://cos.io/rr/). Here, researchers submit their preregistration for peer review. If the research plan is considered theoretically and methodologically sound, the study can get an "in-principle acceptance". If the study is executed and reported according to the preregistration, it will be published, regardless of the outcomes. This is a direct prevention of publication bias.

Preregistration also removes room for undisclosed flexibility in data analysis; any deviation from the original analysis plan is now clearly identified as an exploratory analysis, and can no longer be presented as having been planned from the start. In other words, preregistration creates a clear separation between confirmatory and exploratory findings (Wagenmakers et al., 2012). Note, however, that preregistration can be done in different levels of detail (Wicherts et al., 2016), and several preregistration formats and templates have been suggested (e.g., Van 't Veer & Giner-Sorolla, 2016; for an overview, see Veldkamp, Bakker, et al., 2017). Researcher degrees of freedom can only truly be avoided, if the preregistration contains a high level of detail, and this is currently often not the case (Veldkamp, Bakker, et al., 2017). Future research is still needed to investigate formats and implementations of preregistration, but preregistration seems a promising way to avoid researcher degrees of freedom and publication bias.

11.2 Replication

Science's main self-correction mechanism is replication. Unfortunately, it seems that direct replications are seldom published (Makel et al., 2012). Indeed, researchers, editors, and reviewers seem biased against replication research and in favor of novel results (Neuliep & Crandall, 1990, 1993). Furthermore, some researchers argue against the usefulness of replications in psychology, with as a main point that psychology is too complex and too sensitive to small contextual changes to expect it to replicate (Baumeister, 2016; Dijksterhuis, 2014; Iso-Ahola, 2017; Mitchell, 2014; Stroebe & Strack, 2014). However, such statements are problematic if we want to consider psychology a falsifiable scientific field (Daniel, Yuichi, & Lindsay, 2017; Heino, Fried, & LeBel, 2017; LeBel, Berger, Campbell, & Loving, 2017; Simons, 2014).

I strongly believe that replications are essential for psychology to grow as a scientific field, and I think we need to think about ways to promote high-quality replications. As we showed in Chapter 7, replications are informative if they have high power and are not affected by publication bias and related biases caused by researcher degrees of freedom. One way to ensure both is through Registered Replication Reports (RRR; Association for Psychological Science, n.d.; examples are Alogna et al., 2014; Eerland et al., 2016; Wagenmakers et al., 2016). In an RRR, researchers first submit a preregistration of the research plan, which is then reviewed by peer reviewers and the author(s) of the original study. If the plan is approved, it is posted publicly so that other labs can follow the same protocol and conduct their own replication of that study. The main benefits of the RRRs are that the preregistration prevents researcher degrees of freedom and publication bias, and the collaborative aspect allows for larger samples and thus increases power. Moreover, the multi-lab format allows the study of whether effects are moderated by substantive or methodological factors (e.g., types of sample or specifics of the data collection).

Recently, an increasing number of multi-lab collaborations has emerged (Klein et al., 2014; Open Science Collaboration, 2015) and initiatives are forming to facilitate such collaborations (see, e.g., StudySwap at https://osf.io/9aj5g/). Furthermore, there are services such as PsychFileDrawer (http://www.psychfiledrawer.org/) and Curate Science (http://curatescience.org/) that keep track of replications and publish the results online in what you could call "real-time meta-analyses". Finally, the growing realization of the importance of high quality replications, is also reflected by the fact that The Netherlands Organization for Scientific Research (NWO) recently has set aside 1 million euros to fund replication studies. These are all very promising developments, and I hope this increased appreciation for replication research will continue to grow.

11.3 Understanding Statistics

It has been argued that the current problems surrounding publication bias and researcher degrees of freedom might arise from the widespread (mis)use of Null Hypothesis Significance Testing (NHST) and the strong focus on *p*-values smaller than .05. Performing a significance test and retaining $\alpha = .05$ seems the default analysis strategy in psychology; it has even been called a "mindless" "null ritual" (Gigerenzer, 2004; Gigerenzer, Krauss, & Vitouch, 2004). However, there are many misunderstandings surrounding the meaning of (non)significant *p*-values, and their evidential value is often overestimated (Hoekstra, Finch, Kiers, & Johnson, 2006; Tversky & Kahneman, 1971).

A wide range of solutions has been offered to solve problems surrounding the use of NHST. One recent suggestion has been to decrease the common significance level from $\alpha = .05$ to $\alpha = .005$ (Benjamin et al., 2017; but see Lakens et al., 2017). A perhaps more radical suggestion is to avoid NHST in its entirety (see, e.g., Trafimow & Marks, 2015). Alternatives to

NHST might be to focus on effect size estimation (Cumming, 2013) and/or Bayesian statistics (Wagenmakers, 2007). To my knowledge, it has not yet been investigated if a change in statistical framework will indeed decrease overestimation due to publication bias and researcher degrees of freedom, but it seems a research line that is worthwhile to further investigate.

11.4 Transparency

There have been many suggestions how we can change our research practices and the scientific system as a whole to increase the reliability of psychological science (Asendorpf et al., 2013; Munafò et al., 2017; Nosek & Bar-Anan, 2012; Nosek et al., 2012). One of the most recurring suggestions is to increase transparency and openness (Nosek et al., 2015; Poldrack et al., 2017; Vazire, 2017; Wicherts, Kievit, Bakker, & Borsboom, 2012).

This call for transparency encompasses almost all aspects of a study. First, there is a strong call for open data. Some of the greatest advantages of open data are the possibility to run secondary analyses to answer new questions, verify analyses of published work or examine the robustness of the original analyses, and compute specific effect sizes for meta-analyses (see also Wicherts, 2013). Several journals are now requiring or rewarding open data, and these policies are strongly related to increased data availability (see Chapter 4).

Ideally, when data are shared, there should be a quality check to ensure they are relevant, complete, and usable (Kidwell et al., 2016; Wilkinson et al., 2016). One way to also increase the likelihood that data remain available over time, is to publish them in online data repositories. An example of such a repository is the Open Science Framework (http://osf.io). One advantage of this platform is that it also facilitates sharing other aspects of a study, such as materials and analysis scripts. With the original materials, a study can be more easily replicated, and if analysis scripts are posted online, any flexibility in data analysis can be openly discussed. Another way to ensure quality of published data, is to publish them in the peer-reviewed *Journal of Open Psychology Data* (Wicherts, 2013).

There have also been calls for increased transparency in the publication system. For instance, some have argued that peer reviews should be openly available to insure the accountability of reviewers (Nosek & Bar-Anan, 2012; Wicherts et al., 2012), and several journals are experimenting with different forms of open peer review (see, e.g., *Collabra: Psychology*, *F1000Research*, and *PeerJ*). Related, researchers can post their manuscripts online on preprint servers (e.g., http://psyarxiv.org). Preprints allow more readers to comment on articles than is possible in the more closed peer review system exercised by most journals. Making manuscripts available online (either through preprint servers, or open access publishing) also allow more readers to assess and comment on articles after publication. Platforms such as PubPeer (https://pubpeer.com/) and increasing number of journals allow post-publication review (see, e.g., *PLOS ONE* and *Meta-Psychology*). This form of review can

strengthen the self-corrective mechanism of science particularly when combined with open research practices such as sharing of pre-registrations, data, computer scripts, and research materials.

Overall, there seems to be growing enthusiasm towards increasing transparency in research, and there is increasing talk of an "Open Science Movement" (Spellman, Gilbert, & Corker, 2017). A long list of journals has now signed the Transparency and Openness Promotion (TOP) guidelines to indicate their support for open science (see https://cos.io/our-services/top-guidelines/ for the full list). Furthermore, the Society for the Improvement of Psychological Science (SIPS) was formed specifically to "brings together scholars working to improve methods and practices in psychological science" (http://improvingpsych.org/mission/), and the past two SIPS meetings have already resulted in tangible steps forward (e.g., StudySwap and PsyArXiv) and offers a platform for scholars of all levels of seniority to share their experiences and ideas, to come up with solutions to improve the field.

11.5 Meta-Science

Psychological science is suffering from several large-scale problems that affect its validity and reproducibility. There has been an increasing number of suggestions how to solve these problems. It is important to decide which problems require most attention, and what the most viable solutions are. To do that, we need empirical research: meta-science (Ioannidis et al., 2015).

The idea underlying the field of meta-science is that we should approach problems in psychology in the same way as we approach substantive psychological questions: by using empirical methods. This way, we can provide deeper insights in the severity and nature of different types of bias, and form evidence-based solutions for academic institutions, professional organizations, funding agencies, researchers, and journal publishers. This research contributes to raising awareness among researchers about problems in contemporary science, which hopefully increases the adoption of more rigorous research practices. We already see several top-down initiatives to improve the quality of science. Two great examples are the replication grant of NWO, and the increasing number of journals that are implementing strategies to increase reproducibility. By investing in meta-scientific research, we can help implementing empirically tested solutions to improve psychological science.

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). doi:10.1371/journal.pone.0172792

Alogna, V., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., . . . Brown, C. (2014). Registered replication report: Schooler and engstler-schooler (1990). *Perspectives on Psychological Science, 9*(5), 556-578.

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. A. (2011). Public availability of published research data in high-impact journals. *PLoS One, 6*(9), e24357. doi:10.1371/journal.pone.0024357

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 1: sensitivity and specificity. *British Medical Journal, 308*, 1552. doi:10.1136/bmj.308.6943.1552

American Psychological Association. (1983). *Publication Manual of the American Psychological Association. Third Edition*. Washington, DC: American Psychological Association.

American Psychological Association. (2010). *Publication Manual of the American Psychological Association. Sixth Edition*. Washington, DC: American Psychological Association.

Anagnostou, P., Capocasa, M., Milia, N., Sanna, E., Battaggia, C., Luzi, D., & Bisol, G. D. (2015). When Data Sharing Gets Close to 100%: What Human Paleogenetics Can Teach the Open Science Movement. *PLoS One, 10*(3). doi:10.1371/journal.pone.0121409

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological science, 28*(11), 1547-1562. doi:10.1177/0956797617723724

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

Association for Psychological Science. (n.d.). *Registered Replication Reports*. Retrieved from http://www.psychologicalscience.org/index.php/replication.

Aylward, E., Walker, E., & Bettes, B. (1984). Intelligence in schizophrenia: meta-analysis of the research. *Schizophrenia Bulletin, 10*(3), 430-459.

Baker, M. (2015). Smart software spots statistical errors in psychology papers: One in eight articles contain data-reporting mistakes that affect their conclusions. *Nature News*. doi:http://www.nature.com/news/smart-software-spots-statistical-errors-in-psychology-papers-1.18657

Baker, M. (2016a). 1,500 scientists lift the lid on reproducibility. *Nature News, 533*(7604), 452.

Baker, M. (2016b). Stat-checking software stirs up psychology. *Nature, 540*, 151–152. doi:0.1038/540151a

Bakermans-Kranenburg, M. J., van IJzendoorn, M. H., & Juffer, F. (2008). Earlier is better: A meta-analysis of 70 years of intervention improving cognitive development in institutionalized children. *Monographs of the Society for Research in Child Development, 73*(3), 279-293. doi:10.1111/j.1540-5834.2008.00498.x

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological science, 27*(8), 1069-1077.

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. doi:10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*(3), 666-678. doi:10.3758/s13428-011-0089-5

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

Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policy making. *Educational Evaluation and Policy Analysis, 34*(3), 259-277. doi:10.3102/0162373712446144

Baratloo, A., Hosseini, M., Negida, A., & El Ashal, G. (2015). Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. *Emergency, 3*(2), 48-49.

Barber, L. K. (2017). Meticulous manuscripts, messy results: Working together for robust science reporting. *Stress and Health, 33*(2), 89-91. doi:10.1002/smi.2756

Baron, J. (2011). Acknowledgements and report for the year 2010. *Judgment and Decision Making, 6*(2), 1-3.

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology, 66*, 153-158. doi:10.1016/j.jesp.2016.02.003

Beaujean, A. A. (2005). Heritability of cognitive abilities as measured by mental chronometric tasks: A meta-analysis. *Intelligence, 33*(2), 187-201. doi:10.1016/j.intell.2004.08.001

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*(4), 1088-1101. doi:10.2307/2533446

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Camerer, C. (2017). Redefine statistical significance. *Nature Human Behaviour*, 1.

Berle, D., & Starcevic, V. (2007). Inconsistencies between reported test statistics and p-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research, 16*(4), 202-207. doi:10.1002/mpr.225

Binet, A., & Simon, T. (1905). New methods for the diagnosis of the intellectual level of subnormals. . *L'annee Psychologique, 12*, 191-244.

Bloom, T., Ganley, E., & Winker, M. (2014). Data access for the open access literature: PLOS's data policy. *PLoS biology, 12*(2), e1001797. doi:10.1371/journal.pbio.1001797

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2005a). Fixed-Effect Versus Random-Effects Models. In M. Borenstein, L. V. Hedges, J. P. T. Higgins, & H. R. Rothstein (Eds.), *Introduction to Meta-Analysis* (pp. 77-86). New York: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2005b). Random-Effects Model. In M. Borenstein, L. V. Hedges, J. P. T. Higgins, & H. R. Rothstein (Eds.), *Introduction to Meta-Analysis* (pp. 69-76). New York: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons, Ltd.

Borm, G. F., den Heijer, M., & Zielhuis, G. A. (2009). Publication bias was not a good reason to discourage trials with low power. *Journal of Clinical Epidemiology, 62*(1), 47-53. doi:10.1016/j.jclinepi.2008.02.017

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

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 1-12. doi:10.1038/nrn3475

Caperos, J. M., & Pardo, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema, 25*(3), 408-414. doi:10.7334/psicothema2012.207

Carlisle, J. C., Dowling, K. C., Siegel, D. M., & Alexeeff, G. V. (2009). A blood lead benchmark for assessing risks from childhood lead exposure. *Journal of Environmental Science and Health Part a-Toxic/Hazardous Substances & Environmental Engineering, 44*(12), 1200-1208. doi:10.1080/10934520903139829

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*: Cambridge University Press.

Ceci, S. J. (1988). Scientists Attitudes toward Data Sharing. *Science Technology & Human Values, 13*(1-2), 45-52.

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *Obstetrics & Gynecology, 114*(6), 1341-1345.

Chamberlain, S., Boettiger, C., & Ram, K. (2014). rplos: Interface to PLoS Journals search API. R package version 0.4.0. In. http://CRAN.R-project.org/package=rplos.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*. doi:10.1016/j.cortex.2012.12.016

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

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of" playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience, 1*(1), 4-17.

Chambers, C. D., & Munafò, M. R. (2013). *Trust in science would be improved by study pre-registration.* . Retrieved from http://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration.

Champely, S. (2017). pwr: Basic Functions for Power Analysis. R package version 1.2-1. Retrieved from https://CRAN.R-project.org/package=pwr.

Chan, A. W., & Altman, D. G. (2005). Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *British Medical Journal, 330*(7494), 753-756.

Chan, A. W., Hrobjartsson, A., Haahr, M. T., Gotzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials - Comparison of Protocols to published articles. *Jama-Journal of the American Medical Association, 291*(20), 2457-2465.

Christensen-Szalanski, J. J., & Beach, L. R. (1984). The citation bias: Fad and fashion in the judgment and decision literature. *American Psychologist, 39*(1), 75.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology, 65*(3), 145.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*(12), 997-1003.

Cohn, L. D., & Westenberg, P. M. (2004). Intelligence and maturity: Meta-analytic evidence for the incremental and discriminant validity of Loevinger's measure of ego development. *Journal of Personality and Social Psychology, 86*(5), 760-772. doi:10.1037/0022-3514.86.5.760

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2*(4), 447-452. doi:10.1037/1082-989X.2.4.447

Copas, J. B. (2013). A likelihood-based sensitivity analysis for publication bias in meta-analysis. *Applied Statistics, 62*(1), 47-66. doi:10.1111/j.1467-9876.2012.01049.x

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates - A note on meta-analysis bias. *Professional Psychology-Research and Practice, 17*(2), 136-137.

Crocker, J., & Cooper, M. L. (2011). Addressing scientific fraud. *Science, 334*, 1182. doi:10.1126/science.1216775

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

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., . . . Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological science, 18*(3), 230-232. doi:10.1111/j.1467-9280.2007.01881.x

Daniel, J. S., Yuichi, S., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science, 12*(6), 1123-1128. doi:10.1177/1745691617708630

Davies, G., Armstrong, N., Bis, J. C., Bressler, J., Chouraki, V., Giddaluru, S., . . . Deary, I. J. (2015). Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53 949). *Molecular Psychiatry, 20*, 183. doi:10.1038/mp.2014.188

De Groot, A. D. (1956/2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas]. *Acta psychologica, 148*, 188-194. doi:10.1016/j.actpsy.2014.02.001

De Winter, J., & Happee, R. (2013). Why selective publication of statistically significant results can be effective. *PLoS One, 8*(6), e66463. doi:10.1371/journal.pone.0066463

Deary, I. J., Whalley, L. J., Lemmon, H., Crawford, J., & Starr, J. M. (2000). The stability of individual differences in mental ability from childhood to old age: follow-up of the 1932 Scottish Mental Survey. *Intelligence, 28*(1), 49-55.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 11-33). New York: Wiley.

Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., & Smith, H. (1987). Publication bias and clinical trials. *Controlled Clinical Trials, 8*(4), 343-353.

Diekmann, A. (2011). Are most published research findings false? *Jahrbücher für Nationalökonomie und Statistik, 231*(5+6), 628-635. doi:10.1515/jbnst-2011-5-606

Dijksterhuis, A. (2014). Welcome Back Theory! *Perspectives on Psychological Science, 9*(1), 72-75. doi:10.1177/1745691613513472

Doucouliagos, H., Laroche, P., & Stanley, T. D. (2005). Publication bias in union-productivity research? *Relations Industrielles/Industrial Relations, 60*(2), 320-347.

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89-98. doi:10.1080/01621459.2000.10473905

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455-463. doi:10.1111/j.0006-341X.2000.00455.x

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J., Aucoin, P., . . . Carlucci, M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11*(1), 158-171.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*(7109), 629-634. doi:10.1136/bmj.315.7109.629

Eich, E. (2014). Business not as usual. *Psychological science, 25*(1), 3-6. doi:10.1177/0956797613512465

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*: Cambridge University Press.

Epskamp, S., & Nuijten, M. B. (2014). statcheck: Extract statistics from articles and recompute p values. R package version 1.0.0. http://CRAN.R-project.org/package=statcheck.

Epskamp, S., & Nuijten, M. B. (2015). statcheck: Extract statistics from articles and recompute p values. R package version 1.0.1. http://CRAN.R-project.org/package=statcheck.

Epskamp, S., & Nuijten, M. B. (2016). statcheck: Extract statistics from articles and recompute p values. R package version 1.2.2. http://CRAN.R-project.org/package=statcheck.

Epstein, W. M. (1990). Confirmational response bias among social-work journals. *Science Technology & Human Values, 15*(1), 9-38.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS One, 11*(2), e0149794.

Falkingham, M., Abdelhamid, A., Curtis, P., Fairweather-Tait, S., Dye, L., & Hooper, L. (2010). The effects of oral iron supplementation on cognition in older children and adults: a systematic review and meta-analysis. *Nutrition Journal, 9*. doi:10.1186/1475-2891-9-4

Fanelli, D. (2010). "Positive" Results Increase Down the Hierarchy of the Sciences. *PLoS One, 5*(3), e10068. doi:10.1371/journal.pone.0010068

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

Fanelli, D. (2013). Why Growing Retractions Are (Mostly) a Good Sign. *Plos Medicine, 10*(12). doi:10.1371/journal.pmed.1001563

Fanelli, D. (2014). Rise in retractions is a signal of integrity. *Nature, 509*(7498), 33-33. doi:10.1038/509033a

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1618569114

Fanelli, D., & Ioannidis, J. P. A. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences of the United States of America, 110*(37), 15031-15036. doi:10.1073/pnas.1302997110

Fanelli, D., & Ioannidis, J. P. A. (2014). Reply to Nuijten et al.: Reanalyses actually confirm that US studies overestimate effects in softer research. *Proceedings of the National Academy of Sciences, 1117*, E714-715.

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*, 120-128. doi: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. doi:10.1177/1745691612459059

Fidler, F., & Cumming, G. (2005). Teaching confidence intervals: Problems and potential solutions. *Proceedings of the 55th International Statistics Institute Session*.

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

Field, A. P. (2009). *Discovering statistics using SPSS*: Sage publications.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63*(3), 665-694. doi:10.1348/000711010X502733

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology, 108*(2), 275-297. doi:10.1037/pspi0000007

Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. *PLoS One, 9*(10), e109019. doi:10.1371/journal.pone.0109019

Francis, G. (2012a). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19*(6), 975-991. doi:10.3758/s13423-012-0322-y

Francis, G. (2012b). The same old New Look: Publication bias in a study of wishful seeing. *i-Perception, 3*(3), 176-178. doi:10.1068/i0519ic

Francis, G. (2012c). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review, 19*(2), 151-156. doi:10.3758/s13423-012-0227-9

Francis, G. (2013a). Publication bias in "Red, Rank, and Romance in Women Viewing Men" by Elliot et al. (2010). *Journal of Experimental Psychology: General, 142*, 292-296.

Francis, G. (2013b). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology, 57*(5), 153-169. doi:10.1016/j.jmp.2013.02.003

Francis, G. (2014). The Frequency of Excess Success for Articles in Psychological Science. *Psychonomic Bulletin & Review, 21*(5), 1180-1187. doi:10.3758/s13423-014-0601-x

Francis, G., Tanzman, J., & Matthews, W. J. (2014). Excess Success for Psychology Articles in the Journal Science. *PLoS One, 9*(12), e114255. doi:10.1371/journal.pone.0114255

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505. doi:10.1126/science.1255484

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in Psychology Experiments: Evidence From a Study Registry. *Social Psychological and Personality Science, 7*(1), 8-12. doi:10.1177/1948550615598377

Frank, M. C., & Saxe, R. (2012). Teaching Replication. *Perspectives on Psychological Science, 7*(6), 600-604. doi:10.1177/1745691612460686

Freund, P. A., & Kasten, N. (2012). How Smart Do You Think You Are? A Meta-Analysis on the Validity of Self-Estimates of Cognitive Ability. *Psychological Bulletin, 138*(2), 296-321. doi:10.1037/a0026556

Garcia-Berthou, E., & Alcaraz, C. (2004). Incongruence between test statistics and P values in medical papers. *Bmc Medical Research Methodology, 4*(1), 13. doi:10.1186/1471-2288-4-13

Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis - a "garden of forking paths" - explains why many statistically significant comparisons don't hold up. *American Scientist, 102*(6), 460.

Gerber, A. S., Green, D. P., & Nickerson, D. (2001). Testing for Publication Bias in Political Science. *Political Analysis, 9*(4), 385-392. doi:10.1093/oxfordjournals.pan.a004877

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587-606.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The Null Ritual. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391-408). Thousand Oakes, CA: Sage.

Giner-Sorolla, R. (2012). Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science. *Perspectives on Psychological Science, 7*(6), 562-571. doi:10.1177/1745691612457576

Giofrè, D., Cumming, G., Fresc, L., Boedker, I., & Tressoldi, P. (2017). The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *PLoS One, 12*(4), e0175583. doi:10.1371/journal.pone.0175583

Glass, G. V., Smith, M. L., & McGaw, B. (1981). *Meta-analysis in social research*: Sage Publications, Incorporated.

Gotzsche, P. C., Hrobjartsson, A., Maric, K., & Tendal, B. (2007). Data extraction errors in meta-analyses that use standardized mean differences. *Jama-Journal of the American Medical Association, 298*(4), 430-437.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1-20.

Guan, M., & Vandekerckhove, J. (2015). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review, 23*(1), 74-86.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Bruyneel, S. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546-573.

Harper, L. M., & Kim, Y. (2017). Factors affecting psychologists' adoption of an open data badge. *Proceedings of the Association for Information Science and Technology, 54*(1), 696-698.

Hartgerink, C. H. J. (2016). 688,112 statistical results: Content mining psychology articles for statistical test results. *Data, 1*(3), 14.

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

Hartgerink, C. H. J., van Assen, M. A. L. M., & Wicherts, J. M. (2017). Too Good to be False: Non-Significant Results Revisited. *Collabra: Psychology, 3*(1), 1-18. doi:10.1525/collabra.71

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*, 61-85.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21*(4), 299-332.

Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 145-174). New York: Wiley.

Hedrick, T. E. (1985). Justifications for and obstacles to data sharing. *Sharing research data*, 123-147.

Heino, M. T. J., Fried, E. I., & LeBel, E. P. (2017). Commentary: Reproducibility in Psychological Science: When Do Psychological Phenomena Exist? *Frontiers in Psychology*. doi:10.3389/fpsyg.2017.01004

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review, 13*(6), 1033-1037.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology-and its future prospects. *Educational and Psychological Measurement, 60*, 661-681. doi:10.1177/0013164400605001

Hunt, E. (2010). *Human intelligence*: Cambridge University Press.

IntHout, J., Ioannidis, J. P. A., & Borm, G. F. (2016). Obtaining evidence by a single well-powered trial or several modestly powered trials. *Statistical Methods in Medical Research, 25*(2), 538-552. doi:10.1177/0962280212461098

Ioannidis, J. P. A. (1998). Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *Jama-Journal of the American Medical Association, 279*(4), 281-286.

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

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640-648. doi:10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A. (2011). Excess Significance Bias in the Literature on Brain Volume Abnormalities. *Archives of General Psychiatry, 68*(8), 773-780.

Ioannidis, J. P. A. (2013). Clarifications on the application and interpretation of the test for excess significance and its extensions. *Journal of Mathematical Psychology, 57*(5), 184-187.

Ioannidis, J. P. A. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly, 94*(3), 485-514.

Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLoS biology, 13*(10), e1002264-e1002264. doi:10.1371/journal.pbio.1002264

Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A., & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature genetics, 29*(3), 306-309.

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal, 127*(605).

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology, 58*(6), 543-549. doi:10.1016/j.jclinepi.2004.10.019

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4*(3), 245-253. doi:10.1177/1740774507079441

Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology, 96*, 505-524. doi:10.1348/000712605x53542

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

Jannot, A.-S., Agoritsas, T., Gayet-Ageron, A., & Perneger, T. V. (2013). Citation bias favoring statistically significant studies was present in medical research. *Journal of Clinical Epidemiology, 66*(3), 296-301.

Jennions, M. D., & Moller, A. P. (2002). Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill' method. *Biological Reviews, 77*(2), 211-222. doi:10.1017/s1464793101005875

Jennions, M. D., & Moller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology, 14*(3), 438-445.

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

Jonas, K. J., & Cesario, J. (2016). How can preregistration contribute to research in our field? In: Taylor & Francis.

Kavvoura, F. K., McQueen, M. B., Khoury, M. J., Tanzi, R. E., Bertram, L., & Ioannidis, J. P. A. (2008). Evaluation of the Potential Excess of Statistically Significant Findings in Published Genetic Association Studies: Application to Alzheimer's Disease. *American Journal of Epidemiology, 168*(8), 855-865. doi:10.1093/aje/kwn206

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196-217.

Kidwell, M. C., Lazarevic, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., . . . Nosek, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLoS biology, 14*(5), e1002456. doi:10.1371/journal.pbio.1002456

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, J., Reginald B., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A “Many Labs” Replication Project. *Social Psychology, 45*(3), 142-152. doi: 10.1027/1864-9335/a000178

Kraemer, H. C., Gardner, C., Brooks, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3*, 23-31.

Krawczyk, M., & Reuben, E. (2012). (Un)Available upon Request: Field Experiment on Researchers' Willingness to Share Supplementary Materials. *Accountability in*

Research: Policies and Quality Assurance, 19, 175-186. doi:10.1080/08989621.2012.678688

Krueger, J. (2001). Null hypothesis significance testing - On the survival of a flawed method. *American Psychologist, 56*(1), 16-26.

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

Lakens, D. (2015). What p-hacking really looks like: A comment on Masicampo & Lalande (2012). *Quarterly Journal of Experimental Psychology*.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., . . . Zwaan, R. A. (2017). Justify your alpha: A response to "Redefine statistical significance". *Retrieved from psyarxiv. com/9s3y6*.

Langan, D., Higgins, J., & Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Research Synthesis Methods, 8*(2), 181-198.

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.org: Grassroots Support for Reforming Reporting Standards in Psychology. *Perspectives on Psychological Science, 8*(4), 424-432. doi:10.1177/1745691613491437

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

Lester, B. M., LaGasse, L. L., & Seifer, R. (1998). Drug abuse - Cocaine exposure and children: The meaning of subtle effects. *Science, 282*(5389), 633-634. doi:10.1126/science.282.5389.633

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. doi:10.1080/03637750903074685

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28*(4), 612-625. doi:10.1111/j.1468-2958.2002.tb00828.x

Lexchin, J., Bero, L. A., Djulbegovic, B., & Clark, O. (2003). Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *Bmj, 326*(7400), 1167-1170.

Lindsay, D. S. (2017). Sharing Data and Materials in Psychological Science. *Psychological science, 28*(6), 699-702. doi:10.1177/0956797617704015

Longo, D. L., & Drazen, J. M. (2016). Data sharing. *The New England Journal of Medicine, 374*, 276-277. doi:10.1056/NEJMe1516564

Mackintosh, N. J. (2011). *IQ and human intelligence*: Oxford University Press.

Maddock, J. E., & Rossi, J. S. (2001). Statistical power of articles published in three health-psychology related journals. *Health Psychology, 20*(1), 76.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*, 161-175.

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

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample Size in Psychological Research over the Past 30 Years. *Perceptual and Motor Skills, 112*(2), 331-348. doi:10.2466/03.11.pms.112.2.331-348

Mathes, T., Klaßen, P., & Pieper, D. (2017). Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. *Bmc Medical Research Methodology, 17*(1), 152.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147-163. doi:10.1037/1082-989x.9.2.147

McAuley, L., Pham, B., Tugwell, P., & Moher, D. (2000). Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet, 356*(9237), 1228-1231.

McDaniel, M. A. (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence, 33*(4), 337-346. doi:10.1016/j.intell.2004.11.005

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. In: Elsevier.

Mitchell, J. (2014). *On the evidentiary emptiness of failed replications*. Retrieved from http://jasonmitchell.fas.harvard.edu/Papers/Mitchell_failed_science_2014.pdf.

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *Bmc Medical Research Methodology, 9*. doi:10.1186/1471-2288-9-2

Morey, R. D. (2013). The consistency test does not–and cannot–deliver what is advertised: A comment on Francis (2013). *Journal of Mathematical Psychology, 57*(5), 180-183.

Morris, P. E., & Fritz, C. O. (2017). Meeting the challenge of the Psychonomic Society's 2012 Guidelines on Statistical Issues: Some success and some room for improvement. *Psychonomic Bulletin & Review*, 1-7.

Munafò, M. R., Attwood, A. S., & Flint, J. (2008). Letter to the Editor: Bias in genetic association studies: effects of research location and resources. *Psychological Medicine, 38*(8), 1213-1214.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., . . . Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021.

Murayama, K., Pekrun, R., & Fiedler, K. (2013). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review*, 1088868313496330.

Neisser, U., Boodoo, G., Bouchard Jr, T. J., Boykin, A. W., Brody, N., Ceci, S. J., . . . Sternberg, R. J. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77.

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality, 5*(4), 85-90.

Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior and Personality, 8*(6), 21-29.

Newcombe, R. G. (1987). Towards a reduction in publication bias. *British Medical Journal, 295*(6599), 656-659.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Niemeyer, H., Musch, J., & Pietrowsky, R. (2012). Publication bias in meta-analyses of the efficacy of psychotherapeutic interventions for schizophrenia. *Schizophrenia Research, 138*(2-3), 103-112. doi:10.1016/j.schres.2012.03.023

Niemeyer, H., Musch, J., & Pietrowsky, R. (2013). Publication Bias in Meta-Analyses of the Efficacy of Psychotherapeutic Interventions for Depression. *Journal of Consulting and Clinical Psychology, 81*(1), 58-74. doi:10.1037/a0031152

Nord, C. L., Valton, V., Wood, J., & Roiser, J. P. (2017). Power-up: A Reanalysis of 'Power Failure' in Neuroscience Using Mixture Modeling. *The Journal of Neuroscience, 37*(34), 8051-8061. doi:10.1523/jneurosci.3592-16.2017

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425. doi:10.1126/science.aab2374

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific Utopia: I. Opening Scientific Communication. *Psychological Inquiry, 23*(3), 217-243. doi:10.1080/1047840x.2012.692215

Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: Making sense of replications. *eLife, 6*, e23383. doi:10.7554/eLife.23383

Nosek, B. A., & Lakens, D. (2014). Registered reports. In: Hogrefe Publishing.

Nosek, B. A., Spies, J., & Motyl, M. (2012). Scientific Utopia: II - Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science, 7*, 615-631. doi:10.1177/1745691612459058

Nuijten, M. B. (2017). Share Analysis Plans and Results. *Nature, 551*, 559.

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

Nuijten, M. B., Van Assen, M. A. L. M., Van Aert, R. C. M., & Wicherts, J. M. (2014). Standard analyses fail to show that US studies overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences, 111*(7), E712-E713.

Nuijten, M. B., Van Assen, M. A. L. M., Veldkamp, C. L. S., & Wicherts, J. M. (2015). The replication paradox: combining studies can decrease accuracy of effect size estimates. *Review of General Psychology, 19*(2), 172-182.

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mule, E. (2017). The Chrysalis Effect: How Ugly Initial Results Metamorphosize Into Beautiful Articles. *Journal of Management, 43*(2), 376-399. doi:10.1177/0149206314527133

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657-660. doi:10.1177/1745691612462588

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

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

Pashler, H., & Wagenmakers, E. J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science A Crisis of Confidence? *Perspectives on Psychological Science, 7*(6), 528-530.

Pereira, T. V., & Ioannidis, J. P. A. (2011). Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of Clinical Epidemiology, 64*(10), 1060-1069.

Perugini, M., Galucci, M., & Constantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspectives on Psychological Science, 9*(3), 319-332.

Petrocelli, J., Clarkson, J., Whitmire, M., & Moon, P. (2012). When ab ≠ c – c': Published errors in the reports of single-mediator models: Published errors in the reports of single-mediator models. *Behavior Research Methods*, 1-7. doi:10.3758/s13428-012-0262-5

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS One, 2*(3), e308. doi:10.1371/journal.pone.0000308

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2016). Top 10 Replicated Findings From Behavioral Genetics. *Perspectives on Psychological Science, 11*(1), 3-23. doi:10.1177/1745691615617439

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2015). Estimating the Difference Between Published and Unpublished Effect Sizes: A Meta-Review. *Review of Educational Research*, 1-30.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., . . . Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18*(2), 115-126.

R Core Team. (2014). R: A Language and Environment for Statistical Computing. http://www.R-project.org/.

Renkewitz, F., Fuchs, H. M., & Fiedler, S. (2011). Is there evidence of publication biases in JDM research? *Judgment and Decision Making, 6*(8), 870-881.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331-363.

Ritchie, S. (2015). *Intelligence: All that matters*: Hodder & Stoughton.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59-82.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*, 646-656.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis. Prevention, assessment, and adjustments*. New York: Wiley.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57*, 416-428.

Sakaluk, J., Williams, A., & Biernat, M. (2014). Analytic Review as a Solution to the Misreporting of Statistical Results in Psychological Science. *Perspectives on Psychological Science, 9*(6), 652-660. doi:10.1177/1745691614549257

Schimmack, U. (2012). The Ironic Effect of Significant Results on the Credibility of Multiple-Study Articles. *Psychological Methods, 17*(4), 551-566. doi:10.1037/a0029487

Schmidt, T. (2016). *Sources of false positives and false negatives in the STATCHECK algorithm: Reply to Nuijten et al. (2016)*. Retrieved from https://arxiv.org/abs/1610.01010.

Schmidt, T. (2017). *Statcheck does not work: All the numbers. Reply to Nuijten et al. (2017)*. https://psyarxiv.com/hr6qy/.

Schönbrodt, F. D. (2015). p-checker: One-for-all p-value analyzer. Retrieved from http://shinyapps.org/apps/p-checker/.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470*(7335), 437-437.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*(2), 309.

Shadish, W. R., Doherty, M., & Montgomery, L. M. (1989). How many studies are in the file drawer - An estimate from the family marital psychotherapy literature. *Clinical Psychology Review, 9*(5), 589-603.

Shrout, P. E., & Rodgers, J. L. (2017). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology, 69*(1).

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*, 1359 –1366. doi:10.1177/0956797611417632

Simons, D. J. (2014). The Value of Direct Replication. *Perspectives on Psychological Science, 9*(1), 76-80. doi:10.1177/1745691613514755

Simonshon, U. (2013). It really just does not follow, comments on Francis (2013). *Journal of Mathematical Psychology, 57*(5), 174-176.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological science, 24*(10), 1875-1888. doi:10.1177/0956797613480366

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547.

Singh Chawla, D. (2017). Controversial software is proving surprisingly accurate at spotting errors in psychology papers. *Science*.

Smulders, Y. M. (2013). A two-step manuscript submission process can reduce publication bias. *Journal of Clinical Epidemiology, 66*(9), 946-947. doi:10.1016/j.jclinepi.2013.03.023

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., . . . Harvey, I. (2010). Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess, 14*(8), 1-193.

Spellman, B., Gilbert, E. A., & Corker, K. S. (2017). Open Science: What, Why, and How. *PsyArXiv*.

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. In: SAGE Publications Sage CA: Los Angeles, CA.

Spitz, H. H. (1986). *The raising of intelligence: A selected history of attempts to raise retarded intelligence*: Routledge.

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2017). *What meta-analyses reveal about the replicability of psychological research*. http://www.deakin.edu.au/__data/assets/pdf_file/0007/1198456/WhatMeta-AnalysesReveal_WP.pdf.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*(1), 60-78.

Stanley, T. D., Jarrell, S. B., & Doucouliagos, H. (2010). Could it be better to discard 90% of the data? A statistical paradox. *The American Statistician, 64*(1), 70-77.

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*, 30-34. doi:10.2307/2282137

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited - The effect of the outcome of statistical tests on the decision to publish and vice-versa. *American Statistician, 49*(1), 108-112. doi:10.2307/2684823

Sterling, T. D., & Weinkam, J. J. (1990). Sharing Scientific-Data. *Communications of the Acm, 33*(8), 112-119. doi:10.1145/79173.79182

Stern, J. M., & Simes, R. J. (1997). Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *British Medical Journal, 315*(7109), 640-645.

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 75-98). New York: Wiley.

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99-110). New York: Wiley.

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology, 53*(11), 1119-1129. doi:10.1016/S0895-4356(00)00242-0

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., . . . Higgins, J. P. I. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal, 342*. doi:10.1136/bmj.d4002

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*(5), 401-426. doi:https://doi.org/10.1016/j.intell.2006.09.004

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

Sutton, A. J., Duval, S., Tweedie, R., Abrams, K. R., & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal, 320*(7249), 1574-1577.

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS biology, 15*(3). doi:10.1371/journal.pbio.2000797

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22*(13), 2113-2126. doi:10.1002/sim.1461

The Royal Society. (2017). The Royal Society's research integrity statement.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1-2. doi:10.1080/01973533.2015.1012991

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

Trikalinos, T. A., & Ioannidis, J. P. A. (2005). Assessing the Evolution of Effect Sizes over Time. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 241-259). New York: Wiley.

Tsilidis, K. K., Papatheodorou, S. I., Evangelou, E., & Ioannidis, J. P. A. (2012). Evaluation of Excess Statistical Significance in Meta-analyses of 98 Biomarker Associations with Cancer Risk. *Journal of the National Cancer Institute, 104*(24), 1867-1878. doi:10.1093/jnci/djs437

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.

Van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology - A discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2-12. doi:http://dx.doi.org/10.1016/j.jesp.2016.03.004

van Aert, R. C. M., & van Assen, M. A. L. M. (2017a). Bayesian evaluation of effect size after replicating an original study. *PLoS One, 12*(4). doi:10.1371/journal.pone.0175302

van Aert, R. C. M., & van Assen, M. A. L. M. (2017b). Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication.

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11*(5), 713-729.

van Assen, M. A. L. M., van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014a). Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS One, 9*(1). doi:10.1371/journal.pone.0084896

Van Assen, M. A. L. M., Van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014b). Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS One, 9*(1), e84896.

Van Assen, M. A. L. M., Van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014c). Why we need to publish all studies. Retrieved from http://www.plosone.org/annotation/listThread.action?root=78405

Van Assen, M. A. L. M., Van Aert, R. C. M., & Wicherts, J. M. (2014). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*.

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20*(3), 293-309. doi:10.1037/met0000025

van Dalen, H. P., & Henkens, K. (2012). Intended and Unintended Consequences of a Publish-or-Perish Culture: A Worldwide Survey. *Journal of the American Society for Information Science and Technology, 63*(7), 1282-1293. doi:10.1002/asi.22636

Van Der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review, 113*(4), 842.

Vandekerckhove, J., Guan, M., & Styrcula, S. A. (2013). The consistency test may be too weak to be useful: Its systematic application would not improve effect size estimation in meta-analyses. *Journal of Mathematical Psychology, 57*(5), 170-173.

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

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra, 1*(1), 1-5. doi:10.1525/collabra.13

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology, 3*(1). doi:10.1525/collabra.74

Veldkamp, C. L. S., Bakker, M., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Soderberg, C. K., . . . Wicherts, J. M. (2017). *Restriction of opportunistic use of researcher degrees of freedom in pre-registrations on the Open Science Framework*. Preprint retrieved from https://psyarxiv.com/g8cjq.

Veldkamp, C. L. S., Hartgerink, C. H. J., Van Assen, M. A. L. M., & Wicherts, J. M. (2017). *Shared responsibility for statistical analyses and statistical Reporting errors in psychology articles published in PLOS ONE (2003 – 2016)*. Retrieved from https://psyarxiv.com/g8cjq.

Veldkamp, C. L. S., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A. L. M., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS One, 9*(12), e114876. doi:10.1371/journal.pone.0114876

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., . . . Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods, 7*(1), 55-79.

Vevea, J. L., Clements, N. C., & Hedges, L. V. (1993). Assessing the effects of selection bias on validity data for the General Aptitude-Test Battery. *Journal of Applied Psychology, 78*(6), 981-987.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*(3), 419-435.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods, 10*(4), 428-443. doi:10.1037/1082-989x.10.4.428

Viechtbauer, W. (2010). The metafor package: A meta-analysis package for R (Version 1.3-0). Retrieved from http://cran.r-project.org/web/packages/metafor/index.html

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., . . . Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current biology, 24*(1), 94-97.

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

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*, 779-804. doi:10.3758/BF03194105

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R., . . . Blouin-Hudon, E.-M. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11*(6), 917-928.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., Maas, H. L. J. v. d., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632-638. doi:10.1177/1745691612463078

Walster, G., & Cleary, T. (1970). A proposal for a new editorial policy in the social sciences. *American Statistician, 24*, 16-19.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*(3), 291-298.

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature, 480*, 7. doi:10.1038/480007a

Wicherts, J. M. (2013). Science revolves around the data. *Journal of Open Psychology Data, 1*(1), e1. doi:10.5334/jopd.e1

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence*. doi:10.1016/j.intell.2012.01.004

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS One, 6*(11), e26828. doi:10.1371/journal.pone.0026828

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*, 726-728. doi:10.1037/0003-066X.61.7.726

Wicherts, J. M., Kievit, R. A., Bakker, M., & Borsboom, D. (2012). Letting the daylight in: reviewing the reviewers and other ways to maximize transparency in science. *Frontiers in Computational Neuroscience, 6*, 20. doi:10.3389/fncom.2012.00020

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R. C., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*, 160018. doi:10.1038/sdata.2016.18

Zhang, J.-P., Burdick, K. E., Lencz, T., & Malhotra, A. K. (2010). Meta-Analysis of Genetic Variation in DTNBP1 and General Cognitive Ability. *Biological Psychiatry, 68*(12), 1126-1133. doi:10.1016/j.biopsych.2010.09.016

Summary

Psychology is facing a “replication crisis”. Many psychological findings could not be replicated in novel samples, which lead to the growing concern that many published findings are overly optimistic or even false. In this dissertation, we investigated potential indicators of problems in the published psychological literature.

Most conclusions in psychology are based on statistics, so it is important that statistical results are reported correctly. In Part I of this dissertation, we looked at inconsistencies in the statistical results in published psychology papers. To facilitate our research, we developed the free tool *statcheck*; a “spellchecker” for statistics (Chapters 2 and 3). Using *statcheck*, we found that roughly half of the articles published in psychology contained at least one inconsistency. Moreover, in one in eight papers we found at least one gross inconsistency that may have affected the statistical conclusion (Chapter 2). Against our expectations, we did not find evidence that articles that shared their raw data had a lower probability of inconsistencies (Chapter 4). To prevent statistical reporting inconsistencies, I advise editors of scientific journals to use *statcheck* to check submissions for any potential errors (Chapter 5).

Statistical reporting inconsistencies are only one aspect of the problems currently affecting psychological science. One other major problem that I investigated in Part II of this dissertation is *publication bias*: studies that find statistically significant effects have a higher chance of being published than studies that do not. When only the “success stories” are published, we get a biased view of the effects in the scientific literature. Counterintuitively for most researchers, this problem is not solved when you combine information from multiple, comparable studies. On the contrary, in the presence of publication bias, the overestimation of effects can even become *worse* if you combine studies (Chapter 7). Indeed, we analyzed studies from the social sciences in general (Chapter 8) and from intelligence research (Chapter 9) and found strong evidence that published effects are overestimated. What is more, in intelligence research the sample sizes are systematically too small. In our analyses we could not identify any specific study characteristics that are related to a stronger overestimation of effects.

In this dissertation we found evidence for a high prevalence of statistical reporting inconsistencies and overestimated effects in psychological science. These are worrying findings, and it is important to think about concrete solutions to improve the quality of psychological research. One of the solutions is *preregistration*: researchers publish their entire research plan online, before they start collecting data. This way, studies that did not find any effects can not disappear into the file drawer unnoticed, and publication bias is countered. Furthermore, we should encourage replication research, to ensure that overestimated effects are corrected in the long run. Finally, researchers should more often share raw data and research materials, so that any errors can more easily be detected and corrected.

In the end it all revolves around the question: how can we improve the quality of psychological research? To select the best strategies to do that, we need research on research:

meta-research. If we use scientific methods to study how we can improve science, I predict a great future for psychology.

Nederlandse Samenvatting

De psychologie zit in een crisis. Het blijkt dat een steeds groter aantal “bewezen” psychologische effecten uit eerder wetenschappelijk onderzoek niet meer gevonden worden als dat onderzoek herhaald wordt. Hierdoor maken steeds meer onderzoekers zich zorgen dat psychologische effecten in de wetenschappelijke literatuur overschat zijn, of zelfs überhaupt niet bestaan. In dit proefschrift hebben we onderzocht of we in gepubliceerde psychologische artikelen aanwijzingen konden vinden voor overschatte effecten en andere problemen.

De meeste conclusies in de psychologie zijn gebaseerd op statistiek, dus het is belangrijk dat de statistische resultaten goed gerapporteerd worden. In het eerste deel van dit proefschrift hebben we daarom onderzoek gedaan naar statistische inconsistenties in gepubliceerde artikelen. Om dit onderzoek sneller en makkelijker uit te kunnen voeren, hebben we het programma *statcheck* ontwikkeld; een “spellingschecker” voor statistiek, als het ware (Hoofdstuk 2 en 3). Met behulp van *statcheck* vonden we dat ongeveer de helft van de gepubliceerde psychologische artikelen ten minste een inconsistent statistisch resultaat bevat. In een op de acht artikelen vonden we zelfs grote inconsistenties die de statistische conclusie zouden kunnen veranderen (Hoofdstuk 2). Tegen onze verwachting in, vonden we geen bewijs dat artikelen die ook hun ruwe data deelden, minder kans hadden om statistische inconsistenties te bevatten (Hoofdstuk 4). Om statistische inconsistenties te voorkomen adviseer ik redacteurs van wetenschappelijke tijdschriften onder andere om *statcheck* te gebruiken om ingestuurde artikelen te controleren op mogelijke fouten (Hoofdstuk 5).

Statistische inconsistenties zijn niet het enige probleem in de psychologie. Een ander groot probleem dat we onderzochten in het tweede deel van dit proefschrift is *publicatiebias*: artikelen die bewijs voor het bestaan van een effect vinden, hebben een grotere kans om gepubliceerd te worden dan artikelen die geen bewijs voor een effect vinden. Wanneer alleen “succesverhalen” worden gepubliceerd, levert dit een enorme vertekening op van de effecten in de wetenschappelijke literatuur. Tegen de intuïties van de meeste wetenschappers in, wordt dit probleem niet opgelost als je de informatie uit meerdere, vergelijkbare studies combineert. Integendeel, als er publicatiebias is, kan de vertekening zelfs *erger* worden als informatie uit verschillende studies wordt gecombineerd (Hoofdstuk 7). We vinden sterk bewijs dat effecten overschat zijn in een analyse van studies uit de sociale wetenschappen in het algemeen (Hoofdstuk 8), en in een grote set studies naar intelligentie (Hoofdstuk 9). In het intelligentieonderzoek blijkt bovendien dat de steekproeven systematisch te klein zijn. In onze analyses hebben we geen studiekenmerken kunnen identificeren die samenhangen met effecten die sterker overschat zijn.

In dit proefschrift vinden we bewijs voor een hoge prevalentie van statistische inconsistenties en overschatte effecten in de psychologie. Dit zijn problematische bevindingen, en het is belangrijk om na te denken over concrete oplossingen om de kwaliteit van psychologisch onderzoek te verhogen. Een van de oplossingen is *preregistratie*: onderzoekers publiceren hun hele onderzoeksplan online, voordat ze data gaan verzamelen.

Op deze manier kunnen onderzoeken die geen bewijs vinden voor effecten niet ongezien verdwijnen in de archiefkasten, en wordt publicatiebias tegengegaan. Daarnaast is het belangrijk om herhaalonderzoek (*replicaties*) aan te moedigen, zodat eventuele overschatte effecten uiteindelijk gecorrigeerd worden. Verder zouden onderzoekers ruwe data en onderzoeksmaterialen vaker moeten delen, zodat eventuele fouten makkelijker aan het licht kunnen komen en verbeterd kunnen worden.

Uiteindelijk draait het om de vraag: hoe kunnen we de kwaliteit van psychologisch onderzoek verbeteren? Om hiervoor de beste strategieën te selecteren, moeten we onderzoek doen naar onderzoek: *meta-onderzoek*. Als we wetenschappelijke methodes gebruiken om te bestuderen hoe we de wetenschap kunnen verbeteren, zie ik de toekomst van de psychologie rooskleurig tegemoet.

Dankwoord

De afgelopen vijf jaar heb ik hard aan dit proefschrift gewerkt. Het eindresultaat had niet tot stand kunnen komen zonder de hulp van een grote groep mensen om mij heen. Ik wil deze ruimte graag gebruiken om hen te bedanken.

Op de eerste plaats wil ik mijn promotoren bedanken. Jelte en Marcel, wat hebben jullie mij de afgelopen jaren geweldig begeleid, ondersteund, en gemotiveerd. Door jullie verschillen vulden jullie elkaar perfect aan, maar waar jullie duidelijk overeenkwamen was jullie inzet voor jullie AiO's. Tijdens mijn hele project, en zeker de laatste paar weken, waren jullie altijd razendsnel met inhoudelijke en opbouwende feedback. Zonder jullie was mijn project niet zo soepel verlopen.

Jelte, ik heb ontzettend veel geleerd van jouw strategische inzicht, je schrijftips, en je ambitie. Al in mijn eerste jaar psychologie aan de UvA wist je je enthousiasme over methodenleer op me over te brengen, en dat is eigenlijk nooit opgehouden. Er is meer dan één moment geweest waarop ik twijfelde aan mijn projecten, maar na elke vergadering die we hadden, ging ik weer vol goede moed en nieuwe ideeën aan de slag.

Marcel, jij wist me met beide benen op de grond te houden ("we zijn toch allemaal losers") en met je oog voor detail heb je mijn standaarden een stuk hoger weten te leggen. Jouw eerste zorg was altijd het welzijn van je AiO's, en dat uitte je door meerdere keren per week even je hoofd om de deur te steken om te vragen of we niet te hard werkten en of we nog wel gelukkig waren.

I would also like to thank my committee for taking the time to read my dissertation and to travel to Tilburg. Chris Chambers, Eric-Jan Wagenmakers, Rolf Zwaan, and Marjan Bakker, a lot of your work has served as an inspiration for this dissertation and it is an honor to have you in my committee.

Niels, ruim achtentwintig jaar geleden heb je mijn geboortekaartje getekend en nu de omslag van mijn proefschrift. Als dat niet symbolisch is! Bedankt voor dit kunstwerkje, ik ben er heel erg blij mee.

A massive thank you also goes out to all my colleagues at MTO. Somehow you managed to create an atmosphere of relaxed cooperation rather than competition, which has been very valuable to me. Special thanks go to Erwin, Eva, Sara, Mattis, Robbie, Inga, Chris, Niek, Jules, Hilde, Lianne, and Florian, with whom I shared many complaints over a cup of coffee, weird lunch conversations, and long afternoons in the Esplanade. I would also like to thank the MTO band "The Significant Others"; we were awesome and we should go on tour. De afgelopen jaren bestonden niet alleen uit onderzoek, maar ook uit onderwijs. Wilco, Paulette, Elise, Marjan, Jelte, Luc, en onze fantastische student-assistenten: bedankt voor jullie inzichten en de fijne samenwerking. Marieke, Liesbeth, Rianne, en Anne-Marie, bedankt voor jullie onmisbare ondersteuning!

I would like to thank the Meta-Research Center (a.k.a. our lab group's biweekly cookie eating contest) for providing a supportive environment for scientific (and not so scientific)

discussions. I loved having this group of smart, critical, methodological terrorists around. Many of you were also co-authors on chapters in this dissertation; Robbie, Chris, Hilde, Coosje, but also Elise, Jeroen, Linda, and Sofie, thank you for your time and effort!

Ik wil ook graag mijn oude roomies bedanken. Robert (Nestor), Coosje, en Paulette, wat heb ik ontzettend veel aan jullie gehad de afgelopen jaren. Zelfs op de dagen dat ik er even geen zin meer in had, waren jullie de reden dat ik altijd met plezier naar m'n werk ging. De clandestiene Nespresso, de weekendverhalen, de mannenslinger... Ik kijk met enorm veel genegenheid terug op mijn tijd met jullie en ik hoop dat we elkaar nog heel lang blijven zien.

During my time in Tilburg I also went to quite a number of (inter)national conferences where I met amazing people. Thank you SIPS, COS, BITSS, METRICS, IOPS, and all my "Twitter colleagues". You have been and still are an inspiration! Sean, thank you so much for turning the statcheck R package into a way more user-friendly app.

Het is alweer een tijdje geleden, maar mijn studietijd aan de UvA bij PML heeft zeker bijgedragen aan dit proefschrift. Graag bedank ik mijn oude docenten voor hun lessen, maar ook voor het creëren van een sfeer waarin studenten zich onderdeel voelde van de afdeling. Ik wil ook graag mijn medestudenten bedanken: Charlotte, Claudia, Mattis, Marie, Anja, Sjoerd, Paul, Alexander, bedankt voor alle uren samen blokken in de methodologiewinkel en voor de biertjes bij Kriterion. Sacha, dankjewel dat je me al die jaren terug hebt geleerd hoe ik een R package moest schrijven en dat je me hebt betrokken bij het statcheck project. Ik denk niet dat een van ons het succes van statcheck had kunnen voorzien! Lieve Janneke, dankjewel voor je directheid, openheid, creativiteit, en gezelligheid. De wereld is een geweldig mens kwijtgeraakt.

Moving from Amsterdam to Tilburg was a big step, and I have to admit that I was a bit reluctant about it. Who would have thought it would turn out so great? Dear Lizzie, Gaby, Chrissie, Willem, Maaike, Fieke, Bastian, Nina, and Tünde, you guys have made my time in Tilburg truly amazing. Thank you for all the (theme) parties, game nights, trips, and other weird events we organized. For a group of doctors (to be), the level of conversation was always comfortably low. Even though our girls' nights became more grown-up (from blue-cheese dip and deep-fried brie to four of us independently bringing buckets of cherry tomatoes, what's wrong with us), the serious topics are luckily still alternated with terrible jokes and loads of prosecco. I love you all!

Special thanks go out to my fantastic paranymphs. Dear Byron, you were my first friend here in Tilburg. And since everybody knows you (seriously, how do you do that??), you introduced me to most of the friends I still have today. Thank you for always being the life of the party, my karaoke buddy, and a truly great person. Never change! Lieve Paulette, mijn kantoormaatje van het eerste (nou ja, tweede) uur. In de loop der jaren hebben we al heel wat kantoorbubbels, bankhangwijntjes, en studiococktails soldaat gemaakt. Ik heb veel

geleerd van je rust en je bescheidenheid. Dankjewel voor je luisterend oor en je goede adviezen. Ik hoop dat we samen nog menig rondje om de bieb mogen lopen!

Ik heb ook veel te danken aan mijn trouwe vrienden van vroeger. Lieve Iris, dankjewel voor je jaren trouwe vriendschap. Als ik weer eens aan het stressen was, was het altijd een soort mini-vakantie als ik bij jou, Dennis, en jullie pluizige kinderen mocht komen eten, kletsen, en Friends-marathonnen. Lieve Malou, sinds de entrée van Jaap Walvis in ons honoursklasje was het duidelijk: deze vriendschap kon niet meer stuk. We deelden onze liefde voor lekker eten, goede wijn, en dure spullen, waardoor we onze studententijd behoorlijk extravagant ingevuld hebben. Dankjewel dat ik deel uit mag maken van je prachtige gezin. Lieve Floor, onze vriendschap begon zestien (!) jaar geleden in de brugklas van het Boni. We zijn in de tussentijd behoorlijk verschillende kanten op gegaan en toch houdt onze vriendschap stand. Jij bent mijn lijntje naar Amsterdam en het Amsterdamse leven. Bedankt voor je reminders dat het leven te kort is om altijd maar te werken.

Lieve mama, dankjewel voor de goede gesprekken, de wijntjes op de Homeruslaan, de dagjes naar de sauna. Samen met papa heb je ervoor gezorgd dat we in een warm nest konden opgroeien met onvoorwaardelijke steun, wat onze plannen ook waren. Van jou heb ik geleerd om dingen van de positieve kant te bekijken (“Het is een avontuur!”), om bij alles een lied te associëren, en om kaartjes te sturen (al vraag ik me af of ik ooit zo attent word als jij).

Maretje, Rettie, Mezus, ik moet vaak terugdenken aan dat moment dat je mij zag zitten achter al m’n studieboeken en zei: daar ga ik niet aan beginnen hoor, ik ben toch niet gek! Ik ben blij dat je me ondanks mijn gekte toch altijd hebt gesupport. En ik ben ook blij (en mama ook, denk ik) dat tenminste één van ons nog praktisch en ruimtelijk inzicht heeft. Van jou zou ik er wel twee willen hebben. Ik ben er trots op jouw zus (Jezus) te mogen zijn.

Cees en Floris, jullie weten altijd de nodige rust mee te brengen (dat wordt ook zeer gewaardeerd door Goofie, Milo, en Neko). Floris, dankjewel voor je zorgzaamheid en de taxiritjes bij nood. Cees, bedankt voor je interesse in mijn onderzoek en je enthousiasme voor de wetenschap.

Dan mijn lieve schoonfamilie: Hanneke, Sander, Floortje, en Sterre. Dankjulliewel dat jullie me zo in je hart hebben gesloten. Het voelt altijd als thuiskomen als ik bij jullie over de vloer ben. Bedankt voor de gastronomische hoogstandjes, de potjes Carcassonne, jullie enthousiasme, jullie support, en jullie trots.

Lieve Paul, ik kan niet in woorden uitdrukken hoe blij ik ben dat ik jou ben tegengekomen. Met je engelengeduld en je onvoorwaardelijke support heb je me door de zwaarste weken van mijn PhD gesleept (je was vast blij toen ik de Skype chat van ons werk ontdekte en ik je de godganse dag lastig kon vallen met R vragen). Dankjewel voor je liefde, je humor (“Pakaa!”), je Bourgondische houding, je zakelijk inzicht, en de oneindige stroom dierenplaatjes. Ik kan niet wachten om met jou een volgende stap te zetten. Wij horen bij elkaar.

Tot slot: lieve papa, dankjewel voor je rust, je enthousiasme, je trots. Van jou heb ik geleerd hoe ik een lamp op moet hangen, hoe ik andijviestamppot moet maken, en hoeveel een vrije vrijdag waard kan zijn. Wat had ik graag gehad dat je hierbij had kunnen zijn. Ik mis je ontzettend. Dit proefschrift draag ik op aan jou.