THE HUMAN FALLIBILITY OF SCIENTISTS

Dealing with error and bias
in academic research

Coosje Lisabet Sterre Veldkamp

Colophon

Coosje Lisabet Sterre Veldkamp
The human fallibility of scientists,
206 pages.

PhD thesis, Tilburg University, Tilburg, the Netherlands (2017)

Cover painting: Carolien Veldkamp
Graphic design cover and inside: Rachel van Esschoten, DivingDuck Design (www.divingduckdesing.nl)
Printed by: Gildeprint Drukkerijen, Enschede (www.gildeprint.nl)

ISBN/EAN: 9789462337527

THE HUMAN FALLIBILITY OF SCIENTISTS

Dealing with error and bias in academic research

Coosje Lisabet Sterre Veldkamp

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 8 november 2017 om 14:00 uur
door Coosje Lisabet Sterre Veldkamp
geboren te Hilversum.

Promotiecommissie

Promotores: Prof. dr. J. M. Wicherts
Prof. dr. M. A. L. M. van Assen

Overige leden: Prof. dr. L. M. Bouter
Prof dr. E. M. Wagenmakers
Prof. dr. K. Sijtsma
Prof. dr. F. Agnoli
Dr. S. Vazire
Dr. R. Hoekstra

CONTENTS

Appendices

Addendum

CHAPTER 1

Introduction

THE HUMAN FALLIBILITY OF SCIENTISTS

Just like any other professional endeavor involving human beings, science is prone to human fallibility. Obvious and extreme examples of fallibility, such as the tendency to commit scientific fraud, have received considerable attention (e.g. Bouter, 2015; Buyse et al., 1999; Carlisle, 2012; Diekmann, 2007; Kornfeld, 2012; Marusic, Wager, Utrobicic, Rothstein, & Sambunjak, 2015; Mosimann, Dahlberg, Davidian, & Krueger, 2002; Mosimann, Wiseman, & Edelman, 1995; Simonsohn, 2013; Tijdink et al., 2016; Tijdink, Verbeke, & Smulders, 2014). However, the kind of frailties to which all scientists fall prey, such as proneness to error, confirmation bias, hindsight bias, and motivated reasoning have largely been ignored. While a small number of scholars have been pointing to the hazards of errors and bias in science for over 75 years (Bacon, 1621/2000; Feist, 1998; Mahoney, 1976, 1979; Merton, 1942; Mitroff, 1974; Tversky & Kahneman, 1971; Watson, 1938), empirical research on the effects of human fallibility in science and on how to reduce these effects has been relatively scarce.

The reason for this dearth of empirical research may lie in a lack of acknowledgement of the fallibility of scientists. According to Mahoney, the scientist is “viewed as the paragon of reason and objectivity, an impartial genius whose visionary insights are matched only by his quiet humility” (Mahoney, 1976, p. 3). He argued that not only lay people have this image, but also that “the scientist tends to paint himself generously in hues of objectivity, humility, and rationality”, and that “the average scientist tends to be complacently confident about his rationality and his expertise, his objectivity and his insight”(Mahoney, 1976, p. 4). Although Mahoney did not provide a lot of empirical evidence himself to support these claims, he avidly called for studies of the psychology of the scientist.

Recently, problems with the reliability and reproducibility of research results in various fields have led to widespread debate (e.g. Baker, 2016; Begley & Ioannidis, 2015; Chang & Li, 2015; Ioannidis, 2005b, 2007; Open Science Collaboration, 2015) concerning many of the problems in scientific research that Mahoney pointed out. A field of research addressing these problems in science has been emerging quickly, and has been dubbed ‘meta-research’ (Ioannidis, Fanelli, Dunne, & Goodman, 2015; Poldrack et al., 2016). In this young, still rather fragmented field, scientists from different scientific backgrounds strive for improvements in the way we perform, communicate, verify, evaluate, and reward research (Ioannidis et al., 2015; Munafò et al., 2017). In psychology, which has been said to suffer from a so-called ‘reproducibility crisis’ (Maxwell, Lau, & Howard, 2015; Pashler & Harris, 2012; Pashler & Wagenmakers, 2012; Spellman, 2015) several scholars have been pointing out that the way psychologists conduct their research and analyze their data is often problematic and in need of improvement (e.g. Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; Asendorpf et al., 2013; Bakker, van Dijk, & Wicherts, 2012; Bakker & Wicherts, 2011; Cumming, 2014; Eich, 2014; Funder et al., 2014; John, Loewenstein, & Prelec, 2012; LeBel, Borsboom, Giner-Sorolla, Hasselman, Peters, Ratliff, & Tucker Smith, 2013; Lindsay, 2015; Morey et al., 2016; Nosek et al., 2015; Nosek & Bar-Anan, 2012; Nosek, Spies, & Motyl, 2012; Simmons, Nelson, & Simonsohn, 2011; Simonsohn, Nelson, & Simmons, 2014b; Spellman, 2015; Vazire, 2015, 2017; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Wicherts, 2011, 2013; Wicherts & Bakker, 2012). Proposed solutions include systemic changes, such as study pre-registration, open peer review, and stricter requirements for documenting, archiving and sharing data (Morey et al., 2016; Nosek et al., 2015; Nosek & Bar-Anan, 2012; Nosek et al., 2012; Wicherts & Bakker, 2012).

As a psychologist trained in social and developmental psychology, I joined the meta-research group at Tilburg University that focuses on potential solutions for error and bias in psychological science. Before I started examining potential solutions myself, I aimed to answer a more fundamental question: are scientists likely to acknowledge the need for such solutions? I addressed this question by examining to what extent scientists recognize their own fallibility (Chapter 2). Then I focused on psychological science, and examined potential solutions to reduce the probability of error (Chapters 3 and 4) and bias (Chapters 5 and 6) in the use of the most widely employed statistical framework in psychology, null hypothesis significance testing (NHST).

In Chapter 2, we investigated recognition of the human fallibility of scientists by examining lay people's and scientists' belief in the 'storybook image' of the scientist; the image that a scientist is a person who embodies the virtues of objectivity, rationality, intelligence, open-mindedness, integrity, and communality (Mahoney, 1976, 1979). We examined this in four studies. Studies 1 and 2 tested whether highly-educated lay people and scientists believed the storybook characteristics of the scientist to apply more strongly to scientists than to other highly-educated people. Studies 3 and 4 zoomed in on whether scientists attributed higher levels of the storybook characteristics to scientists of their own social group (i.e. scientists of the same academic level or gender) than to other scientists.

Chapters 3 and 4 concern errors in psychological science. In the studies reported in these chapters, we examined a particular type of error that scientists can fall prey to: errors in the reporting of statistical results. We replicated the alarmingly high prevalence of such errors found in earlier studies (Bakker & Wicherts, 2011; Bakker & Wicherts, 2014a; Berle & Starcevic, 2007; Caperos & Pardo, 2013; Garcia-Berthou & Alcaraz, 2004; Wicherts, Bakker, & Molenaar, 2011) using an automated procedure called 'statcheck' (Epskamp & Nuijten, 2013, 2015). This software package is able to quickly retrieve and check statistical results that are reported according to the publication manual of the American Psychological Association (American Psychological Association, 2010) and can be applied to large samples of articles. Moreover, we evaluated a potential solution to reduce such errors: the so called 'co-pilot model of statistical analysis' (Wicherts, 2011). This model entails a simple code of conduct prescribing that statistical analyses are always conducted independently by at least two persons (typically co-authors). This would stipulate double-checks of the analyses and the reported results, open discussions on analytic decisions, and improved data documentation, that facilitates later replication of the analytical results by (independent) peers. The co-pilot model of statistical analysis was based on how the field of aviation deals with the hazards of human error, where the co-pilot's double checking of the pilot's every move significantly reduces the risk of airplane crashes (Beaty, 2004; Wiegman & Shappell, 2003).

In Chapter 3, we studied the potential effectiveness of the co-pilot model by examining the relationship between the reporting errors that statcheck found in a sample of 697 articles published in six flagship psychology journals, and whether the co-pilot model was employed in these articles. Specifically, by means of an online survey among authors, we documented which authors were involved in various aspects of the data analysis, and whether the data was shared among co-authors. Our goal was to see whether the use of collaborative co-piloting practices was associated with a lower prevalence of reporting errors in the articles. In light of our relatively small sample size and potential drawbacks of our survey methodology, such as memory effects (the survey pertained to articles published a year earlier), response bias, and socially desirable responding, we conducted a second study. In this study, we examined a much larger set of articles and employed a different method to measure co-piloting.

In Chapter 4, we scanned the full population of psychology articles ever published in the multidisciplinary Open Access journal PLOS ONE (14,946) for statistical reporting errors, using statcheck. To measure whether co-piloting occurred in these articles, we made use of the mandatory author contribution statements made in all of these articles. From these author contribution sections and other meta-data on the articles, we automatically retrieved how many authors were listed on the article, how many authors were responsible for the analyses, and whether the first author was responsible for the analyses. Employing the author contribution statements eliminated the limitations of the use of a survey in the previous study and enabled us to obtain co-piloting data of many more articles than in the previous study to determine whether the use of co-piloting was associated with a lower prevalence of reporting errors in the articles.

Chapters 5 and 6 concern biases in psychological science. In these chapters, we focus on a particular type of bias that emerges because of the many choices researchers face in formulating their hypotheses, in designing their studies, in collecting their data, in analyzing their data, and in reporting their results. Psychological studies involve numerous choices that are often arbitrary from a substantive or methodological point of view. A key issue with these choices is that researchers might use these so-called researcher degrees of freedom strategically in order to obtain statistically significant results (Bakker et al., 2012; Simmons et al., 2011). Opportunistic use of researcher degrees is commonly known as '*p*-hacking' (Gelman & Loken, 2013; John et al., 2012; Simmons, Nelson, & Simonsohn, 2013; Simonsohn, Nelson, & Simmons, 2014a) and is problematic for two main reasons. First, *p*-hacking greatly increases the chances of finding a false positive result (DeCoster, Sparks, Sparks, Sparks, & Sparks, 2015; Ioannidis, 2005b; Simmons et al., 2011). Second, it may inflate effect sizes (Bakker et al., 2012; Ioannidis, 2008; Simonsohn et al., 2014a; van Aert, Wicherts, & van Assen, 2016). Hence, together with publication bias (or the failure to publish non-significant results), the opportunistic use of researcher degrees of freedom might play a central role in the publication of research findings that later prove to be difficult to replicate in new samples (Asendorpf et al., 2013).

In Chapter 5, we present and discuss an overview of researcher degrees of freedom that psychological researchers have in formulating their hypotheses, designing their experiments, collecting their data, analyzing their data, and reporting of their results. For each of these phases separately, we describe how various choices can be used opportunistically. With the list of researcher degrees of freedom presented in Chapter 5 we aim to raise awareness of the risk of bias implicit in many psychological studies, to provide a practical checklist to assess the potential for bias in such studies, and to provide a tool to be used in research methods education. In addition, the list served as a basis for the study presented in Chapter 6, where we examined the effectiveness of a potential solution to restrict opportunistic use of researcher degrees of freedom: study pre-registration.

Pre-registration has received the most attention as a solution to counteract the opportunistic use of researcher degrees of freedom and its elevated chances of finding false positive results and possibly inflated effect size estimates (Chambers, 2013; Chambers & Munafo, 2013; de Groot, 1956/2014; van Aert et al., 2016; Wagenmakers et al., 2012). Pre-registration requires the researcher to stipulate in advance the research hypothesis, data collection plan, data analyses, and what will be reported in the paper. Different forms of pre-registration are currently emerging in psychology, mainly varying in terms of the level of detail with respect to the research plan they require researchers to provide. The differences between pre-registration formats suggest that a statement that a particular study was pre-registered may not be indicative of how well researcher degrees of freedom were restricted in that study. We argue in Chapter 5 that in order to be effective in restricting opportunistic use of researcher degrees of freedom, pre-registrations need to be sufficiently *specific*, *precise*, and *exhaustive*. That is, the ideal preregistration should provide a detailed description of all steps that will be taken from hypothesis to the final report (it should be specific). Moreover, each described step should allow only one interpretation or implementation (it should be precise). Finally, a preregistration should exclude the possibility that other steps may also be taken (it should be exhaustive).

In Chapter 6, we examined two types of pre-registrations that are currently hosted by the Center for Open Science on the Open Science Framework: 'Standard Pre-Data Collection Registrations' and 'Prereg Challenge Registrations'. The Standard Pre-Data Collection Registrations format simply asks for a summary of the research plan and asks researchers to indicate whether they have already collected or looked at the data before composing the pre-registration. The Prereg Challenge format, on the other hand, requires authors to fill out a specific form consisting of 26 sections asking for details about many different aspects of the study plan. In our study, we evaluated to what extent random samples of each of these two types of pre-registrations restricted opportunistic use of the researcher degrees of freedom presented in Chapter 5, with the goals to assess the quality of current pre-registrations, to learn on which aspects these pre-registrations currently fall short of countering bias, and to provide recommendations to improve pre-registrations in the future. To evaluate the pre-registrations, we developed a scoring protocol. This protocol can also be used in future studies of pre-registrations, or serve as a guide for reviewers assessing pre-registrations.

Finally, in Chapter 7, I reflect on the findings in this dissertation and offer suggestions for future research. The ambition of this dissertation is to raise awareness of the role of human fallibility in science, and to advance the development of solutions that help scientists deal with their fallibility. With a focus on vexing issues in the use of null hypothesis significance testing in psychological science, we attempt to commence this ambition by contributing to the strengthening of psychological science and increasing the trustworthiness of psychological research.

CHAPTER 2

Who believes in the storybook image of the scientist?

This chapter is published as Veldkamp, C. L. S., Hartgerink, C. H. J., van Assen, M. A. L. M., & Wicherts, J. M. (2017). Who believes in the storybook image of the scientist? *Accountability in research, 24*(3), 127-151.

ABSTRACT

Do lay people and scientists themselves recognize that scientists are human and therefore prone to human fallibilities such as error, bias, and even dishonesty? In a series of three experimental studies and one correlational study (total N = 3,278) we found that the 'storybook image of the scientist' is pervasive: American lay people and scientists from over 60 countries attributed considerably more objectivity, rationality, open-mindedness, intelligence, integrity, and communality to scientists than other highly-educated people. Moreover, scientists perceived even larger differences than lay people did. Some groups of scientists also differentiated between different categories of scientists: established scientists attributed higher levels of the scientific traits to established scientists than to early-career scientists and PhD students, and higher levels to PhD students than to early-career scientists. Female scientists attributed considerably higher levels of the scientific traits to female scientists than to male scientists. A strong belief in the storybook image and the (human) tendency to attribute higher levels of desirable traits to people in one's own group than to people in other groups may decrease scientists' willingness to adopt recently proposed practices to reduce error, bias and dishonesty in science.

"Scientists are human, and so sometimes do not behave as they should as scientists."

An anonymous science Nobel Prize Laureate in our sample, 2014

The storybook image of the scientist is an image of a person who embodies the virtues of objectivity, rationality, intelligence, open-mindedness, integrity, and communality (Mahoney, 1976, 1979). However, to avoid placing unreasonable expectations on scientists, it is important to recognize that they are prone to human frailties, such as error, bias, and dishonesty (Feist, 1998; Mahoney, 1976; Merton, 1942; Mitroff, 1974; Nuzzo, 2015; Watson, 1938). Acknowledging scientists' fallibility can help us to develop policies, procedures, and educational programs that promote responsible research practices (Shamoo & Resnik, 2015).

According to Mahoney, the scientist is "viewed as the paragon of reason and objectivity, an impartial genius whose visionary insights are matched only by his quiet humility"(Mahoney, 1976, p. 3). With respect to scientists' self-image, he claimed that "although somewhat more restrained in his self-portrait, the scientist tends to paint himself generously in hues of objectivity, humility, and rationality", and that "the average scientist tends to be complacently confident about his rationality and his expertise, his objectivity and his insight"(Mahoney, 1976, p. 4). However, Mahoney never supported these claims with empirical evidence. Others had demonstrated that scientists are indeed prone to human biases (Mitroff, 1974; Rosenthal, 1966) and Mahoney himself showed that the reasoning skills of scientists were not significantly different from those of nonscientists (Mahoney & DeMonbreun, 1977), but actual belief in the storybook image of the scientist itself has never been examined. Hence, it remains unclear to what degree lay people and scientists recognize that scientists are only human.

Some early data suggest that the belief in the storybook image of the scientist may be strong among lay people. In a seminal study (Mead & Metraux, 1957), the analysis of a nationwide-sample of essays written by American high school students exposed the stereotypical image of the scientist: in terms of appearance, the scientist was depicted as "a man who wears a white coat and works in a laboratory. He is elderly or middle-aged and wears glasses. He is small, sometimes small and stout, or tall and thin. He may be bald. He may wear a beard, may be unshaven and unkempt. He may be stooped and tired" (Mead & Metraux, 1957, pp. 386-387). In terms of traits, the scientist was depicted as "a very intelligent man – a genius or almost a genius. He has long years of expensive training – in high school, college, or technical school, or perhaps even beyond – during which he studied very hard. He is interested in his work and takes it seriously. He is careful, patient, devoted, courageous, open-minded. He knows his subject. He records his experiments carefully, does not jump to conclusions, and stands up for his ideas even when attacked […]" (Mead & Metraux, 1957, p. 387). A similar, male image was found in later studies (e.g. Basalla, 1976; Beardslee & O'dowd, 1961). The stereotypical image in terms of appearance consistently returned in studies using the now classic 'Draw a Scientist Test' (Beardslee & O'dowd, 1961, p. 998; Chambers, 1983; Fort & Varney, 1989; Newton & Newton, 1992; ó Maoldomhnaigh & Hunt, 1988). More recently, European and American surveys have demonstrated that lay people have a stable and strong confidence both in science (Gauchat, 2012; Smith & Son, 2013) and in scientists (Ipsos MORI, 2014; Smith & Son, 2013). For example, the scientific community was found to be the second most trusted institution in the US (Smith & Son, 2013), and in the UK, the general public believed that scientists meet the expectations of honesty, ethical behavior, and open-mindedness (Ipsos MORI, 2014).

As far as we know, no empirical work has addressed scientists' views of the scientist. Although preliminary results from Robert Pennock's 'Scientific Virtues Project' (cited in "Character traits: Scientific virtue," 2016) indicate that scientists consider honesty, curiosity, perseverance, and objectivity to be the most important virtues of a scientist, these results do not reveal whether scientists believe that the typical scientist actually *exhibits* these virtues. A number of studies on scientists' perceptions of research behavior suggest that scientists may not believe that the typical scientist lives up to the stereotypical image of the scientist. First, a large study among NIH-funded scientists (Anderson, Martinson, & De Vries, 2007) found that scientists considered the behavior of their typical colleague to be more in line with *unscientific* norms such as secrecy, particularism, self-interestedness and dogmatism than with the traditional scientific norms of communality, universalism, disinterestedness, and organized skepticism (Merton, 1942; Mitroff, 1974). Second, a meta-analysis including studies from various fields of science showed that over 14% of scientists claimed that they had witnessed serious misconduct by their peers, and that up to 72% of scientists reported to have witnessed questionable research practices (Fanelli, 2009). Third, publication pressure and competition in science are perceived as high (Tijdink et al., 2014; Tijdink, Vergouwen, & Smulders, 2013), while scientists have expressed concerns that competition "contributes to strategic game-playing in science, a decline in free and open sharing of information and methods, sabotage of others' ability to use one's work, interference with peer-review processes, deformation of relationships, and careless or questionable research conduct" (Anderson, Horn, et al., 2007). Based on these reports, one would expect scientists' belief in the storybook image of the scientist to be low compared to lay people's belief.

On the other hand, there is also reason to hypothesize that scientists do believe in the storybook image: scientists may be prone to the well-established human tendencies of in-group bias and stereotyping (Tajfel & Turner, 1986; Turner, Hogg, Oakes, Reicher, & Wetherell, 1987). In-group bias might lead them to evaluate scientists more positively than non-scientists, or their own group of scientists more positively than other groups of scientists and non-scientists, while stereotyping might lead scientists to believe that some scientists (e.g. elderly and/or male scientists) fit the storybook better than other scientists.

In this paper, we will address potential in-group bias and stereotyping among scientists by examining two versions of social grouping that are particularly relevant in science: the scientist's career level and his or her gender. Status differences of established scientists, early-career scientists and PhD students may be perceived as reflecting the degree to which different scientists fit the storybook image, while in-group biases may lead scientists to attribute more of the storybook characteristics to scientists of their own professional level. For instance, due to the stereotypical image of a scientists being an elderly male (Mead & Metraux, 1957), established scientists might be viewed overall as fitting the storybook image of the scientist better than early-career scientists. Yet, in-group bias might lead early-career scientists to regard themselves as fitting the storybook image of the scientist better than established scientists. It is relevant to study these views among scientists because differences in how researchers view their typical colleague and their own group could play a role in the adoption of recent efforts in science aimed at dealing with human fallibilities. For instance, if established scientists view early-career scientists as being more prone to biases in their work, these established scientists might believe that programs aimed at improving responsible conduct of research should be targeted at early-career scientists, while early-career scientists themselves might feel otherwise.

Similarly, while gender inequality in science is still a widely debated topic (Miller, Eagly, & Linn, 2014; Shen, 2013; Sugimoto, 2013; Williams & Ceci, 2015), male scientists may be believed to fit the storybook image better than female scientists because of the common stereotype of the scientist being male (Chambers, 1983; Hassard, 1990; Mead & Metraux, 1957). However, at the same time in-group biases may lead scientists to attribute more of the storybook characteristics to scientists of their own gender. Knowing how male and female scientists view applicability of the storybook image of the scientist to male and female scientists could contribute to the debate on the nature and origins of gender disparities in science (Ceci & Williams, 2011; Cress & Hart, 2009; Shen, 2013; Sugimoto, 2013; West, Jacquet, King, Correll, & Bergstrom, 2013).

We investigate lay people's and scientists' belief in the storybook image of the scientist in four studies. Studies 1 and 2 aimed to test whether highly-educated lay people and scientists believe the storybook characteristics of the scientist to apply more strongly to scientists than to other highly-educated people. In Study

1, we used an experimental between-subjects design to compare the perception of the typical scientist to the perception of the overall group of other highly-educated people who are not scientists, whereas in Study 2, we used a mixed design with random ordering to compare scientists with nine specific other professions that require a high level of education, like medical doctors or lawyers. We expected that both scientists and non-scientists with a high level of education would attribute higher levels of objectivity, rationality, open-mindedness, intelligence, cooperativeness, and integrity to people with the profession of scientist than to people with one of the other nine professions.

Studies 3 and 4 only involved scientist respondents and zoomed in on potential effects of in-group biases and stereotypes related to academic levels and gender. In Study 3, we used an experimental between-subjects design to study whether scientists overall believe that scientists of higher professional levels fit the storybook image of the scientist better than scientists of lower professional levels, as the ‘elderly’ stereotype prescribes. We also studied whether scientists at different career stages differ in this belief, because in-group biases might lead them to attribute more of the storybook characteristics to their own professional level.

In Study 4, we used a similar experimental between-subjects design to test the hypothesis that scientists believe that male scientists fit the storybook image of the scientist better than female scientists, as expected on the basis of the predominantly male stereotype of the scientist. Moreover, Study 4 addresses the question whether male and female scientists are prone to in-group biases leading them to believe that the storybook characteristics apply more strongly to scientists of their own gender.

STUDY 1

Method

Participants

Three groups of participants participated in Study 1, constituting the variable Respondent Group. These groups are specified below.

Scientists

To obtain a representative sample of scientists, we extracted e-mail addresses of corresponding authors from scientific articles published in 2014 that were listed in the Web of Science database (Thomson Reuters, 2014). We sent out batches of e-mail invitations until we reached our desired sample sizes (see power calculations in our study pre-registration through https://osf.io/z3xt6/). Our e-mailed invitations to participate in our study yielded 1,088 fully completed responses from across the globe, of which 343 were from the United States. The response rate was 10.6% (see Table S1 in the supplementary materials in Appendix A). In order to compare results of scientists with results of American highly-educated lay people (see below), only responses from American scientists were used in our statistical analyses. After a priori determined outlier removal (see study pre-registration through https://osf.io/z3xt6/), we were able to use the responses of 331 American scientists (34% female). Their mean age was 49 years (SD = 11.4, range = 26 – 77).

Highly-educated lay people

Survey software and data collection company Qualtrics provided us with 315 fully completed responses of a representative sample of highly-educated non-scientists. These respondents were members of the Qualtrics' paid research panel, and were selected on the following criteria: American citizen, aged over 18, and having obtained a Bachelor's degree, a Master's degree, or a Professional degree, but not a PhD. Response rates could not be computed for this sample, as Qualtrics advertises ongoing surveys to all its eligible panel members and terminates data collection when the required sample size is reached. However, Qualtrics indicates that their response rate for online surveys generally approaches 8%. After a priori determined outlier removal we were able to use the responses of 312 respondents (46% female). Their mean age was 49.2 years (SD = 13.8, range = 23 – 84).

Nobel Prize laureates

To our sample of scientists and highly-educated lay people we added a sample of scientists who might be viewed as the 'paragon of the ideal scientist': Nobel Prize laureates in the science categories. As we anticipated that the size of this additional sample would be too small to include in the statistical analyses, we decided in advance that the data of this extra sample would be used descriptively in the graphical representation of the data but not in the statistical analyses. We conducted an online search for the e-mail addresses of all Nobel Prize laureates in the science fields to date as listed on the Official Web Site of the Nobel Prize (Nobelprize.org, 2014). Our emailed invitations yielded 34 fully completed responses from science Nobel Prize laureates (100% male). The response rate in this sample was 19.0%. The mean age was 75.3 (SD = 12.7, range = 45 – 93).

Materials and procedure

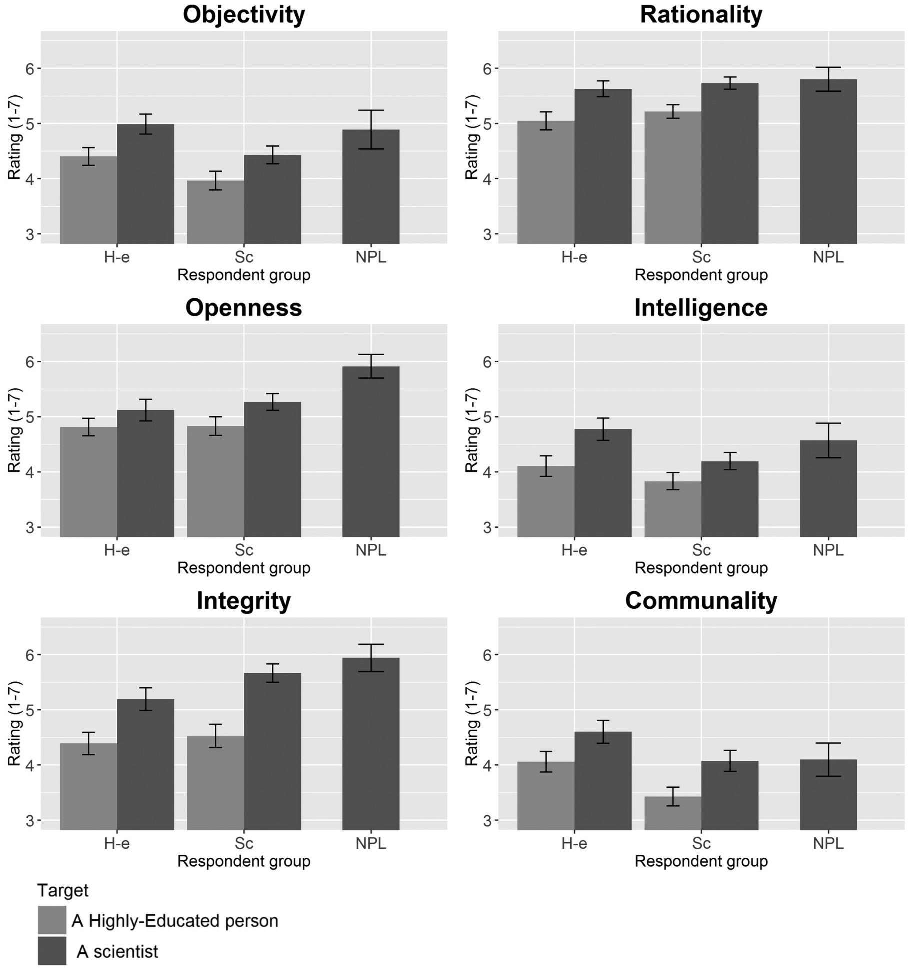
We programmed our between-subjects experimental design into an electronic questionnaire using Qualtrics software, Version March 2014 (Qualtrics, 2014). The program randomly assigned the scientist respondents and the highly-educated respondents to one of two conditions (Targets): either to a condition in which the questions pertained to the 'typical scientist' (Target 'Scientist', defined as "a person who is trained in a science and whose job involves doing scientific research or solving scientific problems"), or to a condition in which the statements pertained to the 'typical highly-educated person' (Target 'Highly-educated person', defined as "a person who obtained a Bachelor's Degree or a Master's Degree or a Professional Degree and whose job requires this high level of education"). Participating Nobel Prize laureates were always assigned to the condition in which the Target was 'Scientist'. By using a between-subjects design, we explicitly ensured that respondents did not compare the Target 'Scientist' to the Target 'Highly-educated person', but rated their Target on its own merits.

Respondents were asked to indicate on a seven-point Likert scale to what extent they agreed or disagreed with 18 statements about the objectivity, rationality, open-mindedness, intelligence, integrity, and communality (cooperativeness) of their Target (either a scientist or a highly-educated person (depending on the experimental condition to which they had been assigned). The statements were presented in randomized order. Each set of three statements constituted a small but internally consistent scale: Objectivity ($\alpha = 0.73$), Rationality ($\alpha = 0.76$), Open-mindedness ($\alpha = 0.77$), Intelligence ($\alpha = 0.73$), Integrity ($\alpha = 0.87$), and Communality ($\alpha = 0.79$). The statements were based on the 'testable hypotheses about scientists' postulated by Mahoney in his evaluative review of the psychology of the scientist [7] and can be found in the 'Materials' section of the supplementary materials in Appendix A and on our Open Science Framework page (https://osf.io/756ea/). The instructions preceding the statements emphasized that respondents should base their answers on *how true* they believed each statement to be, rather than on how true they believed the statement *should* be. Finally, all respondents were asked to answer a number of demographic questions, and were given the opportunity to answer an open question asking whether they had any comments or thoughts they wished to share.

Results

The results of Study 1 are presented in Figure 2.1. In line with our expectations, there was a main effect of Target for each of the characteristics: respondents who were assigned to the Target 'Scientist' ascribed more objectivity (Cohen's $d = 0.47$, 95% CI = [0.31, 0.63]), rationality ($d = 0.63$ [0.48; 0.79]), open-mindedness ($d = 0.35$ [0.19; 0.50]), intelligence ($d = 0.44$, 95% CI = [0.29, 0.60]), integrity ($d = 0.77$, 95% CI = [0.61, 0.93]), and communality ($d = 0.48$, 95% CI = [0.32, 0.63]) to their Target than respondents who were assigned to the Target 'Highly-educated person'. The absence of any interaction effects indicated that there was no evidence that the effects of Target were different in size in the respondent groups.

Figure 2.1 *Attributions of Objectivity, Rationality, Open-mindedness, Intelligence, Integrity, and Communality to the Targets 'Highly-educated person' and 'Scientist', by Respondent Group.*



Note: H-e = Highly-educated respondent group; Sc = Scientist respondent group; NPL = Nobel Prize laureates respondent group. Nobel Prize laureates were always assigned to the Target 'Scientist'. The error bars represent 95 % confidence intervals.

In addition, there were main effects of Respondent Group: scientists on average tended to be *less* generous than lay people in their attributions of objectivity (*d* = 0.45, 95% CI = [0.29, 0.60]), intelligence (*d* = 0.36, 95% CI = [0.21, 0.52]), and communality (*d* = 0.47, 95% CI = [0.31,-0.62]), but a little *more* generous in their attributions of rationality (*d* = 0.16, 95% CI = [0.00, 0.31]) and integrity (*d* = 0.23, 95% CI = [0.07, 0.38]). As can be seen in Figure 2.1, Nobel Prize laureates tended to attribute relatively high levels of the storybook characteristics to their Target 'Scientists'. In all studies, we conducted separate analyses for each of the six storybook characteristics and employed an alpha of 0.008333 (0.05/6) for the interaction effects or main effects. We used an alpha of 0.05 for subsequent analyses of simple effects. Detailed descriptive results for each subsample and all statistical test results can be found in supplementary Tables S1-S4 in Appendix A.

Discussion of Study 1

Study 1 confirmed our hypothesis that lay people perceive scientists as considerably more objective, rational, open-minded, honest, intelligent, and cooperative than other highly-educated people. We also found scientists' belief in the storybook image to be similar to lay people's belief. Comparable patterns were found among scientists from Europe (N = 304) and Asia (N = 117, see Figure S1 in the supplementary materials in Appendix A), indicating that the results may generalize to scientists outside the USA. Nobel laureates' ratings of the Target 'Scientist' were generally similar to, albeit somewhat higher than other scientists' ratings of the Target 'Scientist'.

One potential drawback of the design of Study 1 was that the scale may have been used differently in the two conditions; because the concept 'a highly-educated person' refers to a more heterogeneous category than the concept 'a scientist', respondents may have given more neutral scores in the 'highly-educated' condition than in the 'scientist' condition. In Study 2, we addressed this issue by examining whether similar results would be obtained when explicit comparisons were made between the profession of scientist and other specific professions that require a high level of education.

STUDY 2

Method

Participants

Two groups of participants participated in Study 2, constituting the variable Respondent Group. Sample sizes were smaller than in Study 1 because Study 2 employed a mixed design in which all respondents rated all targets (in a randomized order).

Scientists

We recruited a group of scientist respondents in the same manner as in Study 1. After excluding the 281 non-American responses, our method to recruit participants yielded 123 complete responses. The response rate was 11.0% (see Table S5 in the supplementary materials in Appendix A). After a priori determined outlier removal we were able to use the responses of 111 American scientists (20% female). Their mean age was 49.9 years (SD = 12.4, range = 27 – 85).

Highly-educated lay people

Qualtrics provided us with 81 fully completed responses from a representative sample of highly-educated American people. These respondents were members of the Qualtrics' paid research panel, and they were selected on the following criteria: American citizen, aged over 18, and having obtained a Bachelor's degree, a Master's degree, or a Professional degree, but not a PhD. Response rates could not be computed for this sample, as Qualtrics advertises ongoing surveys to all its eligible panel members and terminates data collection when the required sample size is reached. However, Qualtrics indicates that their response rate for online surveys generally approaches 8%. After a priori determined outlier removal we were able to use 75 of their responses (47% female). The mean age in this group was 46.3 years (SD = 14.7, range = 22 – 83).

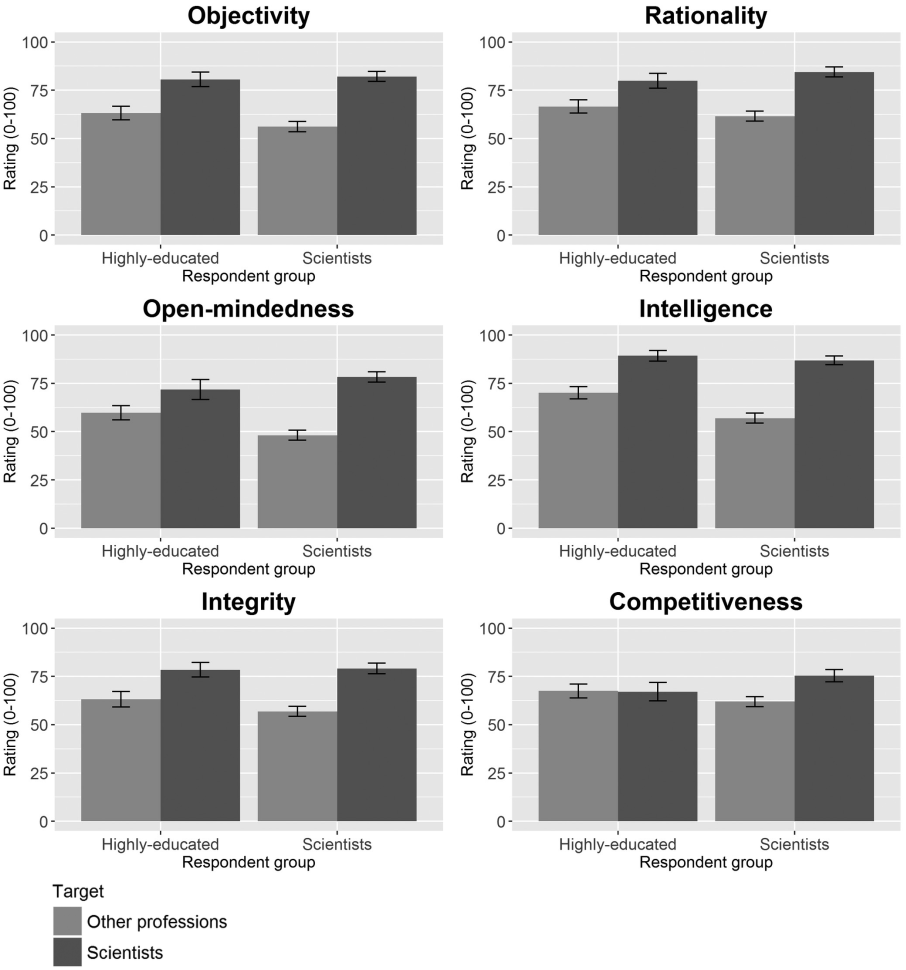
Materials and procedure

We programmed a mixed between-subjects / within-subjects design into an electronic questionnaire using Qualtrics software, Version March 2014 (Qualtrics, 2014). This time, respondents were not randomly assigned to one of two conditions, but all respondents were asked how much each of the six characteristics of the ideal scientist (objectivity, rationality, open-mindedness, integrity, intelligence and communality) applied to ten different professions requiring a high-level education. For each of the features, respondents indicated on slider bars ranging from 0 to 100 how much they believed it applied to the typical person with the profession of lawyer, politician, journalist, medical doctor, accountant, army-lieutenant, banker, judge, detective, and scientist. Respondents were explicitly instructed to indicate how much they believed each feature *really applied* to the typical person within this profession rather than how much the feature *should apply* to the typical professional in each category. We used Mahoney's (1979) antonym 'competitiveness' instead of 'communality' because we were concerned that the term 'communality' might be unclear for respondents. The characteristics were presented in random order, and within the characteristics, the professions were also presented in random order. Finally, just as in Study 1, all respondents were asked to answer a number of demographic questions and were given the opportunity to answer an open question asking whether they had any comments or thoughts they wished to share.

Results

Results of Study 2 are presented in Figure 2.2. Because we were specifically interested in the overall differences in perception between the profession of the scientist and other professions that require a high level of education, we pooled the ratings of the non-scientist professions and compared these to the ratings of the scientist profession. The means of the ten different professions separately are

Figure 2.2 *Attributions of Objectivity, Rationality, Open-mindedness, Intelligence, Integrity, and Communality to people with highly-educated professions and people with the profession of scientist, by Respondent Group.*



Note: The error bars represent 95 % confidence intervals.

presented in Figure S2 in the supplementary materials in Appendix A and indicate that the patterns were similar across professions, justifying the pooling of their means.

Similar to Study 1, respondents attributed more objectivity, rationality, open-mindedness, intelligence, integrity, and competitiveness to scientists than to other types of professionals. However, this time, interactions qualified the effects. Follow-up analyses of the effect of Target in each Respondent Group (scientists and highly-educated lay people) indicated that scientists perceived greater differences between scientists and other types of professionals than lay people did. The effect sizes of the difference in attributions to scientists and to the other types of professionals were much larger in the scientist respondent group (objectivity: d = 1.76, 95% CI = [1.57, 1.94], rationality: d = 1.50, 95% CI = [1.31, 1.69], open-mindedness: d = 1.71, 95% CI = [1.52, 1.90], intelligence: d = 1.88, 95% CI = [1.69, 2.07], integrity: d = 1.51, 95% CI = [1.32, 1.69], and competitiveness: d = 0.75, 95% CI = [0.56, 0.93]) than in the lay people respondent group (objectivity: d = 1.02, 95% CI = [0.79, 1.25], rationality: d = 0.79, 95% CI = [0.56, 1.02], open-mindedness: d = 0.63, 95% CI = [0.40, 0.86], intelligence: d = 1.44, 95% CI = [1.21, 1.67], integrity: d = 0.87, 95% CI = [0.64, 1.10], and competitiveness: d = -0.03, 95% CI = [-0.26, 0.20]). Detailed descriptive results and statistical test results can be found in supplementary Tables S5-S8 in Appendix A.

Discussion of Study 2

Study 2 again confirmed the hypothesis that scientists are perceived as considerably more objective, more rational, more open-minded, more honest, and more intelligent than other highly-educated professionals. Study 2 did not confirm that scientists are perceived as more communal than other highly-educated professionals. Our choice of measuring perceived ‘communality’ (a potentially unclear term) through its opposite ‘competitiveness’ might explain the difference with Study 1, where scientists were perceived as more communal than other highly-educated people: respondents may not have perceived competitiveness as an antonym of communality.

Comparing specific professions ruled out the potential alternative explanation for the results of Study 1: that the highly-educated Target was referring to a more heterogeneous category than the scientist Target and therefore elicited more neutral responses. Again, similar patterns were found among European (n = 67) and Asian scientists (n = 20, see Figure S3 in the supplementary materials in Appendix A), indicating that these results may generalize beyond American scientists. While in Study 1 there was no evidence that the effect of Target was larger in one respondent group than in the other respondent group, in Study 2 we did find that the effect of Target was larger in the Scientist respondent group: scientists perceived much larger differences between people with the profession of scientist and people with other highly-educated professions than highly-educated lay respondents did.

Although our studies are not equipped to test whether any of these perceived differences between professions in attributed traits reflect actual differences in these traits, our finding that scientists rate scientists higher on the storybook traits than lay people do may be explained by in-group biases among scientists. In-group biases, or tendencies to rate one's own group more favorably, are not expected to play any role among the heterogeneous sample of lay respondents (not specifically sampled to be in any of the nine remaining professions), but might have enhanced ratings of scientists among the scientists. In-group biases among scientists are further investigated in Studies 3 and 4.

STUDY 3

Method

Participants

We recruited an international sample of scientists in the same manner as in Studies 1 and 2. This time our method to recruit participants yielded 1,656 complete responses from scientists who fulfilled our inclusion criteria for PhD student, early-career scientist (defined as having obtained a PhD 10 years ago or less, and not having obtained tenure), or established scientist (defined as having obtained a PhD more than 10 years ago and having obtained tenure). The response rate was 10.6% (see Table S9 in the supplementary materials in Appendix A). Because the sample of PhD students turned out much too small compared to the size required by our sample size calculations (see supplementary materials in Appendix A), we decided not to use their responses in our analyses. Because in this study we did not compare results with lay people from the US, we included responding scientists from across the globe. After removal of the PhD students and a priori determined removal of outliers we were able to use the responses of 515 early-career scientists from 55 countries (32% female) and 903 established scientists from 63 countries (22% female) in our analysis. The mean age of the early-career scientists was 35.2 years (SD = 5.8, range = 26 – 94), the mean age of the established scientists was 51.9 years (SD = 9.2, range = 35 – 90). The data of the PhD students are retained in the publicly available data file on the Open Science Framework (see https://osf.io/756ea/).

Materials and procedure

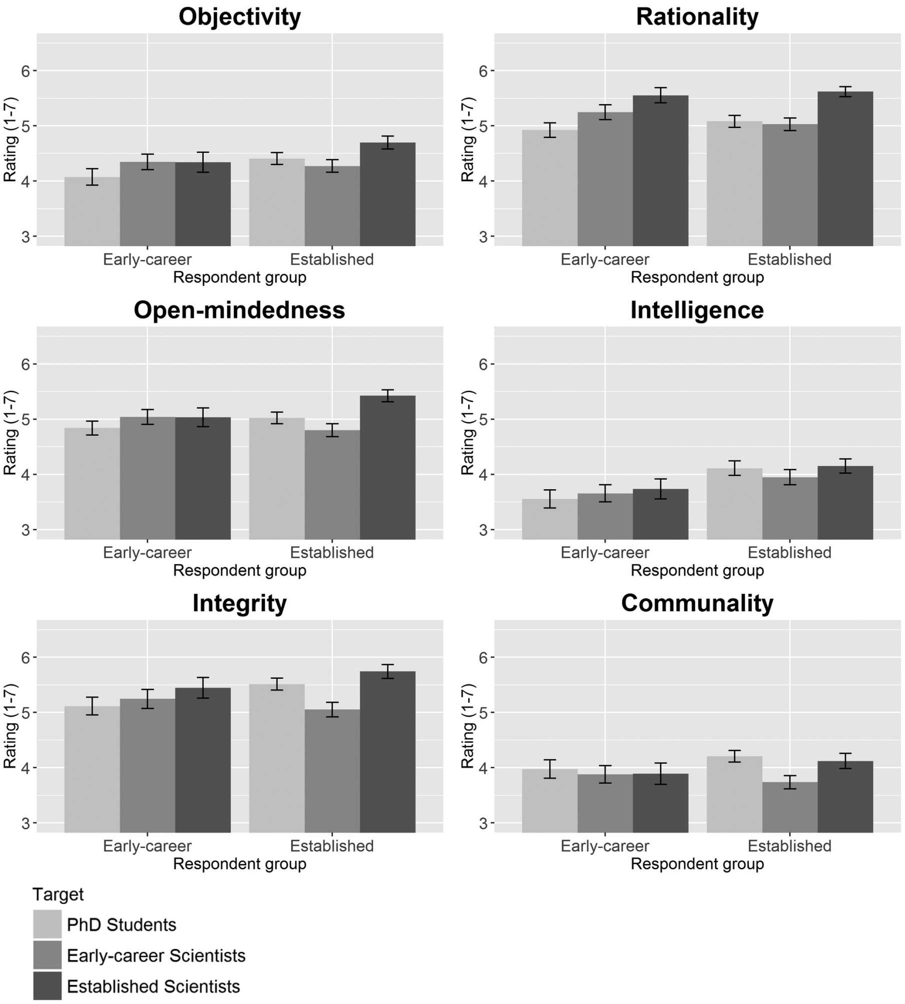
As in Study 1, we programmed a between-subjects experimental design into an electronic questionnaire using Qualtrics software, Version March 2014 (Qualtrics, 2014). The program randomly assigned respondents to one of three conditions; either to a condition in which the statements pertained to an established scientist (Target 'Established scientist'), to a condition in which the statements pertained to an early-career scientist (Target 'Early-career scientist'), or to a condition in which the statements pertained a PhD student (Target 'PhD student'). The sets of statements again constituted sufficiently consistent scales: Objectivity (α = 0.63), Rationality (α = 0.74), Open-mindedness (α = 0.67), Intelligence (α = 0.70), Integrity (α = 0.82), and Communality (α = 0.63). As in the other studies, the instructions preceding the statements emphasized that respondents should base their answers on *how true* they believed each statement to be, rather than on how true they believed the statement *should* be. The 18 statements were presented in randomized order. Finally, all respondents were asked to answer a number of demographic questions, and they were given the opportunity to answer an open question asking whether they had any comments or thoughts they wished to share.

Results

Results of Study 3 are presented in Figure 2.3. In line with the notion of in-group biases, interactions were statistically significant for all features except intelligence and communality, indicating that effects of Target were different in the two analyzed respondent groups. Subsequent analyses of the effects in the separate respondent groups of early-career scientist respondents and established scientist respondents indicated that established scientists who were assigned to the Target 'Established scientist' attributed considerably more objectivity (*d* = 0.41, 95% CI = [0.25, 0.57]), rationality (*d* = 0.64, 95% CI = [0.48, 0.81]), open-mindedness (*d* = 0.62, 95% CI = [0.46, 0.79]), and integrity (*d* = 0.61, 95% CI = [0.45, 0.77]) to their Target than established scientists who were assigned to the Target 'Early-career scientist'. Established scientists who were assigned to the Target 'Established scientist' also attributed more objectivity (*d* = 0.30, 95% CI = [0.13, 0.45]), rationality (*d* = 0.36, 95% CI = [0.15; 0.58]), open-mindedness (*d* = 0.42, 95% CI = [0.26, 0.58]), and integrity (*d* = 0.22, 95% CI = [0.06, 0.38]) to their Target than established scientists who were assigned to the Target 'PhD student'. Interestingly, established scientists who were assigned to the Target 'Early-career scientist' attributed *less* open-mindedness (*d* = -0.23, 95% CI = [-0.49,-0.07]) and integrity (*d* =-0.44, 95% CI = [-0.60,-0.27]) to their Target than established scientists who were assigned to the Target 'PhD student'.

The effects were smaller among early-career scientists; early-career scientists who were assigned to the Target 'Early-career scientist' only attributed more objectivity (*d* = 0.28, 95% CI = [0.07, 0.50]) and rationality (*d* = 0.60, 95% CI =

Figure 2.3 *Attributions of Objectivity, Rationality, Open-mindedness, Intelligence, Integrity, and Communality to the Targets 'Established scientists', 'Early-career scientists' and 'PhD student' by Respondent Group.*



Note: The error bars represent 95 % confidence intervals.

[0.44, 0.76]) to their Target than early-career scientists who were assigned to the Target 'PhD student', and early-career scientists who were assigned to the Target 'Established scientist' only attributed more rationality (*d* = 0.34, 95% CI = [0.12, 0.55]) to their Target than early-career scientists who were assigned to the Target 'Early-career scientist'. Detailed descriptive results and statistical test results can be found in Tables S9-S12 in Appendix A.

Discussion of Study 3

Study 3 partially confirmed our hypothesis that scientists, just like other human beings, are prone to in-group bias. Although stereotypes may play a role here as well, the in-group effect appears to be stronger among established scientists than among early-career scientists. This may be explained by research showing that high status group members have been found to be more prone to in-group bias than low status group members (Bettencourt, Charlton, Dorr, & Hume, 2001). In-group biases have also been found to be stronger among people who identify more strongly with their group (Tajfel & Turner, 1986; Turner et al., 1987), which might apply more to established scientists than to early-career scientists because they have been a scientist for a larger part of their lives.

The difference in in-group bias between early-career scientists and established scientists may also be partly explained by belief in the stereotypical image of the scientist as an old and wise person: if both early-career scientists and established scientists believe that established scientists fit the storybook image better, this would enhance the apparent in-group bias among established scientist, but not among early-career scientists. However, as the early-career scientists only agreed to some extent that established scientists fit the storybook image better than early-career scientists, the effect of the stereotypical image of the scientists cannot be fully responsible for the stronger in-group effect among established scientists. In addition, the stereotypical image of the older scientist cannot explain either why established scientists believe that in some respects, PhD students fit the storybook image of the scientist better than early-career scientists. In Study 4, we tested whether in-group biases among scientists generalize to another highly relevant form of social grouping in science: in-group bias in terms of gender.

STUDY 4

Method

Participants

We recruited an international sample of scientists in the same manner as in the first three studies. This time method to recruit participants yielded 1,003 complete responses (response rate 12.0%, see Table S13 in the supplementary materials in Appendix A). After a priori outlier removal we were able to use the responses of 711 male scientists from 63 countries (mean age = 45.1, SD = 11.9, range = 25 – 86) and 286 female scientists from 46 countries (mean age = 41.8, SD = 10.3, range = 24 – 73).

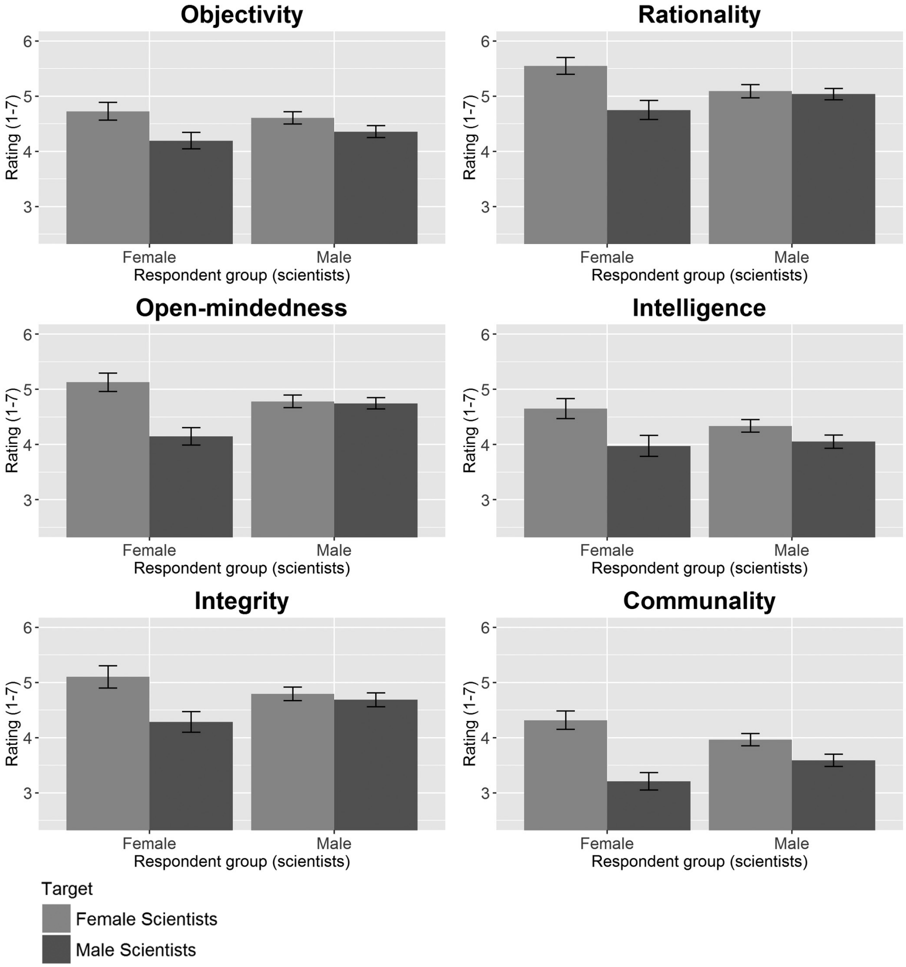
Materials and procedure

As in Studies 1 and 3, we programmed a between-subjects experimental design into an electronic questionnaire using Qualtrics software, Version March 2014 (Qualtrics, 2014). The program randomly assigned respondents to one of two conditions; either to a condition in which the statements pertained to a female scientist (Target 'Female scientist'), or to a condition in which the statements pertained to a male scientist (Target 'Male scientist'). The sets of statements constituted sufficiently consistent scales: Objectivity (α = 0.58), Rationality (α = 0.78), Open-mindedness (α = 0.67), Intelligence (α = 0.62), Integrity (α = 0.79), and Communality (α = 0.58). As in the other studies, the instructions preceding the statements emphasized that responders should base their answers on *how true* they believed each statement to be, rather than on how true they believed the statement *should* be. The 18 statements were presented in randomized order. Finally, all respondents were asked to answer a number of demographic questions and were given the opportunity to answer an open question asking whether they had any comments or thoughts they wished to share.

Results

The results of Study 4 are presented in Figure 2.4. Interactions were significant for all features except objectivity and intelligence, indicating that the effect of Target was different for male and female respondents. Subsequent analyses of the effects for male and female respondents separately indicated that female scientists who were assigned to the condition 'Female scientist' attributed more rationality (d = 0.82, 95% CI = [0.57, 1.06]), more open-mindedness (d = 0.99, 95% CI = [0.75, 1.24]), more integrity (d = 0.69, 95% CI = [0.45, 0.93]), and much more communality (d = 1.13, 95% CI = [0.88, 1.38]) to their Target than female scientists who were assigned to the Target 'Male scientist'. Male scientists who were assigned to the Target 'Female scientist' attributed only somewhat more communality (d = 0.35 [0.20; 0.50]) to their Target than male scientists who were assigned to the Target 'Male scientist'. We thus found support for in-group bias among female scientists, but not for in-group bias among male scientists. Furthermore, we found no evidence for the stereotypical notion that male scientists are believed to fit the storybook image of the scientist better than female scientists. If anything, overall, higher levels of the storybook characteristics were attributed to female scientists than to male scientists. Detailed descriptive results and statistical test results can be found in Tables S13-S16 in Appendix A.

Figure 2.4 *Attributions of Objectivity, Rationality, Open-mindedness, Intelligence, Integrity, and Communality to female scientists and to male scientists, by Respondent Group.*



Note: The error bars represent 95 % confidence intervals.

Discussion of Study 4

Although there are no empirical data on actual gender differences in scientific traits or behavior (except for a study showing that relatively more male scientists than female scientists get caught for scientific misconduct; Fang, Bennett, & Casadevall, 2013), Study 4 showed that female scientists are generally believed to exhibit higher levels of the scientific traits than male scientists. This contrasts with lay people's stereotypical image of the scientist being male. At the same time, we found interactions between the respondent groups and the targets that could be explained in part by in-group biases among both male and female scientists. While women perceived a larger difference between female and male scientists than men did, we cannot rule out that in-group bias led male scientists to rate female scientists lower on the scientific traits than women themselves did.

The finding that women tended to perceive larger differences between male and female scientists in terms of scientific traits might be explained by the fact that in most countries, universities are still male dominated (Shen, 2013). As minority group members, women may be more aware of inequalities and make an effort to have their in-group evaluated positively (Tajfel, 1981). In addition, minority group members tend to identify more strongly with their in-group than majority group members, and stronger group identification is associated with stronger in-group bias (Tajfel & Turner, 1986; Turner et al., 1987). Strikingly, research on intragroup and intergroup perception among male and female academics in a natural setting yielded results very similar to ours: in evaluations of qualities of male and female scientists in an environment where female scientists were clearly a minority, female scientists demonstrated clear in-group favoritism, while male scientists did not (Brown & Smith, 1989).

Even though respondents were intentionally randomly assigned to rate either male or female scientists to prevent them from explicitly comparing the two groups, in this particularly study the implicit comparison was of course obvious. As academic environments are considered rather liberal and progressive, social desirability may have played a significant role in respondents' answers. E-mails we received from male participants in particular indicated that the study topic was quite sensitive.

While this study was designed to test scientists' in-group bias and stereotyping, the unexpected results warrant further investigation of gender differences in scientists' perceptions of colleagues, of the sensitivity of the topic, and of actual gender differences in the scientific traits. The results also advocate taking gender into account in future studies comparing lay people's and scientists' perceptions of scientists.

GENERAL DISCUSSION

Our results indicate strong belief among both lay people and scientists in the storybook image of the scientist as someone who is relatively objective, rational, open-minded, intelligent, honest, and communal. However, while the stereotypical image predicts that older, male scientists would be believed to fit the storybook image best, our results suggest that scientists believe that older, female scientists fit the image best. In addition, our research suggests that scientists are not immune to the human tendency to believe that members of one's own social group are less fallible than members of other groups.

The extent to which our results generalize outside our samples may be limited by selection bias among scientist respondents. The method we used to recruit scientists yielded a high number of respondents, but the overall response rate was low (around 11%). However, our experimental designs in which participants were randomly assigned to different conditions should largely cancel out the potential effects of selection bias occurring through the possibility that scientists who were more interested in the topic of our study were more likely to agree to participate than scientists who were less interested in the topic. With respect to the generalizability of our samples of highly-educated Americans, we cannot exclude the possibility that although the survey panel provider Qualtrics assures representativeness of the American (highly-educated) population, people who sign up to be paid survey panel members may differ in a number of aspects from people who do not sign up to be paid survey panel members.

Our findings are particularly interesting in the context of current discussions on policy and practices aimed at reducing adverse effects of human fallibility in science. In recent years, mounting retractions due to scientific misconduct and error (Zimmer) and increasing doubts about the reproducibility of findings in many scientific fields (Ioannidis, 2005b, 2012; Open Science Collaboration, 2015) have evoked numerous proposals for methods to help us stop 'fooling ourselves' (Nuzzo, 2015): new ways to reduce error, bias, and dishonesty in science. Examples include initiatives that promote transparency in the research process, publication and peer review (Nosek et al., 2015; Nosek & Bar-Anan, 2012), pre-registration of hypotheses and data analysis plans (Chambers & Munafo, 2013; de Groot, 1956/2014; Nosek & Lakens, 2015; Nosek et al., 2012; Wagenmakers et al., 2012), collaboration on statistical analysis (Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, 2014; Wicherts, 2011), blind data analysis (MacCoun & Perlmutter, 2015), reforms in incentive structures (Chambers, 2015; Nosek et al., 2012), training in research integrity (Steneck, 2013), and modifications of reward systems (Ioannidis, 2014). However, the question that arises from our results is then: are scientists willing to adopt these practices if they believe that the typical scientist is mostly immune to human fallibility? Do they deem these initiatives necessary? And if they do deem them necessary, do they deem them necessary for themselves, or only for other (groups of) scientists?

We found that scientists may be prone to in-group bias. Here, social grouping was only made salient in terms of professional level and gender, but in real academic settings, social grouping can occur at more levels and in different ways. Scientists may categorize themselves as members of a research group, a faculty department, a faculty, an institution, a scientific field, a certain paradigm, and so on. If scientists are indeed prone to in-group biases, they may recognize that scientists are human, but still believe that scientists outside their group are more fallible than scientists within their group, and that new research policies aimed to counter human fallibilities need not focus to scientists like themselves.

The remarkable finding that established scientists believe that early-career scientists fit the storybook image of the scientist less well than PhD students may be related to a perceived relationship between publication pressure and use of questionable research practices (QRPs) or academic misbehavior. Early- and mid-career scientists have expressed concerns that competition and publication pressures negatively affect how science is done (Anderson, Horn, et al., 2007), and academic age has been found to be negatively correlated with experienced publication pressure (Tijdink et al., 2013). This may lead established scientists to believe that early-career scientists are more likely to engage in QRPs (and thus fit the storybook image less well) than PhD students and established scientists, but studies comparing self-admitted usage of QRPs and misbehavior between scientists of different career-stages have yielded mixed results. Some studies found that younger scientists are more likely to admit to undesirable scientific behavior (Anderson, Martinson, et al., 2007; Tijdink et al., 2014), while other studies found that older scientists are more likely to admit to this kind of behavior (Martinson, Anderson, Crain, & De Vries, 2006; Martinson, Anderson, & de Vries, 2005). Another explanation might be sought in the idea that Ph.D. students represent potential rather than practice, making it easier to imagine them as matching the ideal.

Just like any other professional endeavor involving human beings, science is prone to human error and bias. As long as we lack objective data on higher levels of objectivity, rationality, open-mindedness, intelligence, integrity or communality among scientists, the scientific community would benefit from acknowledging the human fallibility of scientists by encouraging or even implementing measures that reduce the effect of human factors. Not only scientists themselves, but science policy makers, science funders, academic institutes, and scientific publishers should all actively strive together for a 'scientific utopia' (Nosek & Bar-Anan, 2012; Nosek et al., 2012): a transparent, reproducible science system in which there is room for correction of error. Institutes like the Center of Open Science (https://cos.io/) are working hard to create user-friendly platforms such as the Open Science Framework (https://osf.io/) that enable scientists to manage their entire research cycle practicing transparency, open collaboration, proper documenting, archiving and sharing of research materials, data, and analysis scripts, and to benefit in other ways from open science (McKiernan et al., 2016). Peer-reviewed study pre-registration, as offered and encouraged by the Center for Open Science's Pre-registration Challenge (see https://cos.io/prereg/), reduces 'researcher degrees of freedom' (Simmons et al., 2011) and helps scientists to avoid falling prey to human biases such as confirmation bias and hindsight bias. It's time to step off our pedestal, accept our humanness and collaborate to create an open research culture that acknowledges, but at the same time addresses, our fallibility.

CHAPTER 3

Statistical reporting errors and collaboration on statistical analyses in psychological science

This chapter is published as 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

ABSTRACT

Statistical analysis is error prone. A best practice for researchers using statistics would therefore be to share data among co-authors, allowing double-checking of executed tasks just as co-pilots do in aviation. To document the extent to which this 'co-piloting' currently occurs in psychology, we surveyed the authors of 697 articles published in six top psychology journals and asked them whether they had collaborated on four aspects of analyzing data and reporting results, and whether the described data had been shared between the authors. We acquired responses for 49.6% of the articles and found that co-piloting on statistical analysis and reporting results is quite uncommon among psychologists, while data sharing among co-authors seems reasonably but not completely standard. We then used an automated procedure to study the prevalence of statistical reporting errors in the articles in our sample and examined the relationship between reporting errors and co-piloting. Overall, 63% of the articles contained at least one *p*-value that was inconsistent with the reported test statistic and the accompanying degrees of freedom, and 20% of the articles contained at least one *p*-value that was inconsistent to such a degree that it may have affected decisions about statistical significance. Overall, the probability that a given *p*-value was inconsistent was over 10%. Co-piloting was not found to be associated with reporting errors.

Most conclusions in psychological research (and related fields) are based on the results of null hypothesis significance testing (NHST) (Cohen, 1994; Hubbard & Ryan, 2000; Krueger, 2001; Levine, Weber, Hullet, Park, & Lindsey, 2008; Nickerson, 2000; Sterling, Rosenbaum, & Weinkam, 1995). Although the use and interpretation of this method have been criticized (e.g. Cumming, 2014; Gigerenzer & Edwards, 2003; Wagenmakers et al., 2011), it continues to be the main method of statistical inference in psychological research (Bakker & Wicherts, 2011; Wetzels et al., 2011). Not only for the readers of the psychological literature to be able to interpret and assess the validity of research results, but also for the credibility of the field, it is thus crucial that NHST results are correctly reported. Recent results however suggest that reported results from *t*, *F*, and χ^2 tests in the scientific literature are characterized by a great deal of errors (Bakker & Wicherts, 2011; Berle & Starcevic, 2007; Caperos & Pardo, 2013; Garcia-Berthou & Alcaraz, 2004; Wicherts et al., 2011). An example of such an error can be found in the following results (which, apart from the variable names, appeared in a published article): "All two-way interactions were significant: A × B, $F(1, 20) = 9.5$, $p < .006$; A × C, $F(1, 20) = 0.54$, $p < .03$; and C × B, $F(1, 20) = 6.8$, $p < .02$". Even without recalculation, the experienced user of NHST may notice that the second of these *p*-values is inconsistent with the reported F-statistic and the accompanying degrees of freedom. The *p*-value that corresponds to this F-statistic and these degrees of freedom equals .47.

Bakker and Wicherts (2011) found that 50% of the articles reporting the results of NHST tests in the psychological literature contained at least one such inconsistent *p*-value, and that 18% of the statistical results was incorrectly reported. Similar yet slightly lower error rates have been found in the medical literature (Berle & Starcevic, 2007; Garcia-Berthou & Alcaraz, 2004) and in recent replications (Bakker & Wicherts, 2014a; Caperos & Pardo, 2013; Leggett, Thomas, Loetscher, & Nicholls, 2013). Bakker and Wicherts (2011) discuss different reasons why these inconsistent *p*-values may appear. For example, the output for a three-way Analysis of Covariance (ANCOVA) in the current version of the popular package SPSS contains no less than 79 numbers, many of which are redundant and therefore easily incorrectly retrieved. When several analyses are conducted, results are readily mixed up and typographic errors occur easily. Other reasons for statistical errors may be misunderstanding of data analysis in general (Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993) or misunderstanding of NHST (Nickerson, 2000) in particular.

In many areas where human errors are common and potentially consequential, systems have been implemented to help reduce the likelihood of these errors (Reason, 1990). An example of such a system is co-piloting in aviation: double-checking the pilot's every move significantly reduces the risk of human errors leading to airplane crashes (Beaty, 2004; Wiegman & Shappell, 2003). Another example is pair-programming in Agile Software Engineering, which is found to help reduce errors in programming code (Lindvall et al., 2004). Wicherts (2011) suggested that scientists should learn from aviation and other fields that deal with human error, and proposed a method to reduce errors in the reporting of statistical results: the co-pilot model of statistical analysis. This model involves a simple code of conduct prescribing that statistical analyses are always conducted independently by at least two persons (typically co-authors). This would stipulate double-checks of the analyses and the reported results, open discussions on analytic decisions, and improved data documentation that facilitates later replication of the analytical results by (independent) peers.

Contrary to common practice in medical sciences where statisticians usually conduct the statistical analyses, psychological researchers typically conduct their statistical analyses themselves. Although multiple authors on papers have become the *de facto* norm in psychology (Cronin, Shaw, & La Barre, 2003; Mendenhall & Higbee, 1982; Over, 1982), it is currently unknown how many authors are generally involved in (double-checking) the analyses and reporting of the statistical results. Co-piloting in statistical analysis may concern the independent re-execution of the analyses (e.g., reproducing the results of a test in SPSS), verifying the sample size details, scrutinizing the statistical results in the manuscript, and sharing the data among co-authors before and after publication. In this study, we therefore defined co-piloting as having at least two people involved in conducting the statistical analyses, in writing down the sample details, in reporting the statistical results, and in checking the reported statistical results. In addition, co-piloting in our definition means that at least two people have access to the data before the manuscript is submitted, and that at least two people still have access to the data five years after publication of the article. Data sharing between at least two authors ensures shared responsibility for proper documentation and archiving of the data.

In the present study we estimated the prevalence of inconsistent *p*-values resulting from *t, F,* χ^2*, r, Z* and *Wald* tests in articles published in six flagship journals in psychology. To this end, we employed an automated procedure to document the prevalence of statistical reporting errors in 697 articles published in high-impact journals representing six main empirical psychology disciplines. Moreover, we documented the extent to which co-piloting currently occurs in psychology by asking the authors of the articles in our sample a number of questions about the first (or only) study reported in the article: we asked them to indicate whether they had collaborated on four aspects of analyzing data and reporting results, and whether the described data had been shared between the authors. Our design enabled us to fulfill a third objective: to examine the relationship between statistical reporting errors and co-piloting. As we are not aware of any other work documenting collaboration practices on statistical analyses in psychology or any other research area, we had no hypotheses regarding the extent to which co-authors currently employ the co-pilot model. We did however hypothesize that co-piloting would be associated with a reduced risk of statistical reporting errors, and thus expected the probability of a given p-value being incorrect to be lower in papers in which the statistical analyses and the reporting of the results had been co-piloted (i.e. where more than one person had been involved). This time-stamped hypothesis can be found at the Open Science Framework via http://osf.io/dkn8a.

METHODS

The Prevalence of Statistical Reporting Errors

Sample

For each psychology subfield as listed in the search engine of Thompson Reuters' 2012 Journal Citation Reports (Applied Psychology, Biological Psychology, Clinical Psychology, Developmental Psychology, Educational Psychology, Experimental Psychology, Mathematical Psychology, Multidisciplinary Psychology, Psychoanalysis, and Social Psychology), we chose the journal with the highest 5-year Impact Factor, which (1) was published in English, (2) required the publication style of the American Psychological Association (APA) (American Psychological Association, 2010), and (3) published at least 80 empirical articles in 2011. Four subfields were excluded for different reasons. Educational Psychology was excluded because high-ranking journals in Educational Psychology and Developmental Psychology largely overlapped. Mathematical Psychology was excluded because articles in this field do not usually report the results of NHST. We excluded Multidisciplinary Psychology because we did not regard this field useful to compare subfields of psychology, and we excluded Psychoanalysis because hardly any empirical studies are reported in this field. From the remaining six subfields, the following journals were included in our sample: the *Journal of Applied Psychology* (Applied Psychology), the *Journal of Consulting and Clinical Psychology* (Clinical Psychology), the *Journal of Child Psychology and Psychiatry* (Developmental Psychology), the *Journal of Cognitive Neuroscience* (Experimental Psychology), the *Journal of Personality and Social Psychology* (Social Psychology), and *Psychophysiology* (Biological Psychology). On 24 October 2012, we downloaded all 775 articles published in these journals since Jan 1st, 2012 and then read each abstract in order to determine whether an article was empirical or not. After this selection, our final sample consisted of 697 empirical articles (see Table 1).

Table 3.1 *Sample.*

Field	Journal title	5-year IF	No. of articles	Empirical
Applied Psychology	Journal of Applied Psychology (JAP)	6.850	97	78
Biological Psychology	Psychophysiology (PP)	4.049	129	127
Clinical Psychology	Journal of Consulting and Clinical Psychology (JCCP)	6.369	120	105
Developmental Psychology	Journal of Child Psychology and Psychiatry (JCPP)	6.104	114	90
Experimental Psychology	Journal of Cognitive Neuroscience (JCN)	6.268	150	147
Social Psychology	Journal of Personality and Social Psychology (JPSP)	6.901	165	150
Total			775	697

Note: 5-yr IF = five-year Impact Factor in 2011. Articles = number of articles published in 2012. Empirical = number of empirical articles published in 2012

Procedure

To assess the accuracy of the *p*-values reported in our sample of articles, we used a recently developed automated procedure called *statcheck* (Epskamp & Nuijten, 2013). Statcheck is a package in R, a free software environment for statistical computing and graphics (R Core Team, 2013), and is available through https://github.com/MicheleNuijten/statcheck. The version of statcheck that we used for this paper (0.1.0) extracts *t, F,* χ^2, *r, Z* and *Wald* statistics from articles that are reported as prescribed by the APA Publication Manual (American Psychological Association, 2010). Statcheck re-computes *p*-values in the following way: first, it converts a PDF or HTML file to plain text, and then scans the text for statistical results. Next, it re-computes *p*-values using the test statistics and the degrees of freedom. Finally, it compares the reported and recomputed *p*-value and indicates whether these are consistent or not, while taking into account the effects of rounding. In addition, it specifies whether an inconsistent *p*-value comprises a 'gross error': when the *p*-value is inconsistent to the extent that it may have affected a decision about statistical significance (in this case: when it is reported as smaller than 0.05 while the recomputed *p*-value is larger than 0.05, or vice versa). It is important to note that statcheck's error prevalence estimate may somewhat underestimate or overestimate the true error prevalence because it cannot read statistical results that are inconsistent with the APA's reporting guidelines (American Psychological Association, 2010) or statistical results that contain additional symbols representing for example effect sizes.

In total, 8,110 statistical results were retrieved from 430 of the 697 empirical articles (see Table 2). Five *p*-values that were seemingly reported as larger than 1 were excluded after determining that these had been incorrectly retrieved due to the program's inability to read *p*-values reported as '*p* times 10 to the power of'. A close inspection of the retrieved results revealed that statcheck also had difficulties reading results containing the χ^2 symbol and results in which effect sizes or other measures had been included between the p-values and the test statistics (e.g. $F(1, 46) = 8.41$, $\eta_p^2 = .16$, $p = .006$). This explains at least partly why results were retrieved from a relatively low number of articles. For each of the remaining 8,105 retrieved results, two independent coders tracked down whether the test was reported as one-sided or two-sided, and whether the results belonged to the first (or only) study reported in the article or not. Moreover, the two coders manually checked all statistical results that statcheck identified as 'gross errors' using a strict coding protocol that required the coders to verify whether these *p*-values indeed constituted an error related to statistical significance. Inter-rater reliability was high: in most cases, both coders agreed on whether the study belonged to Study 1 (Cohen's Kappa = 0.92) and on whether the results were reported as one-sided or as two-sided (Cohen's Kappa = 0.85). The inter-rater reliability for decision errors was somewhat lower (Cohen's Kappa = 0.77), because of possible disagreement on whether the result was tested as one-sided or as two-sided due to ambiguous reporting. Such ambiguity in reporting sidedness of the test highlights the importance of reporting standards, hence we suggest that one-sided tests always be described as "one-tailed", "one-sided", or "directional". Whenever two coders disagreed on the test's sidedness, a third coder was asked to independently code the final result.

In the second phase of the protocol, we manually checked the statistical results for which the *p*-value had been reported as '*p*=0.05'. We realized that this was necessary because statcheck could not determine whether a result that had been reported this way was classified as significant by the authors of the article. We therefore looked up all 105 *p*-values reported as '=0.05' in the text of the article, determined whether the result had been described as significant or not, and copied the sentence in which the result was reported in into our data file. Again, in those cases where the two coders disagreed (in 2 of the 105 cases), a third coder was asked to independently code the result. For a detailed description of the coding protocol and the flowchart we used, please refer to the supplementary materials (Appendix B: Table S1, Figure S1, and Figure S2). All manual checks were conducted before the link was made with the survey responses in order to keep the coders blind to whether or not particular analyses were co-piloted.

Table 3.2 *Number of articles from which p-values were retrieved, number of p-values retrieved per journal, and mean number of p-values retrieved per article and per journal.*

Journal	No. of articles	No. of *p*-values retrieved	Mean no. of *p*-values retrieved per article
JAP	42	340	8.10
JCCP	67	833	12.43
JCN	107	1721	16.08
JCPP	39	444	11.38
JPSP	133	4018	30.21
PP	42	749	17.83
Total	430	8105	18.86

Note: JAP = Journal of Applied Psychology; JCCP = Journal of Consulting and Clinical Psychology; JCN = Journal of Cognitive Neuroscience; JCPP = Journal of Child Psychology and Psychiatry; JPSP = Journal of Personality and Social Psychology; PP = Psychophysiology;

Co-Piloting in Psychology

Participants

We searched for the contact details of all 3,087 authors of the 697 empirical articles in our sample and obtained at least one email address for each article in our sample. In total, we managed to track down the email addresses of 2,727 authors (88.3%) and sent them an invitation to participate in our online survey in the first week of July, 2013. We sent two reminders to non-responding authors and stopped collecting data one week after sending the second reminder. This way, we aimed to obtain at least one response for most articles. In total, we received at least one response for 346 articles, amounting to an article response rate of 49.6%. Using personalized hyperlinks to the survey (containing the article title and the 'author number' indicating whether the respondent was first author, second author, etc.) we were able to establish whether more than one author of an article had responded. To make sure that no more than one response per article was used in the analyses that included survey responses, we only retained the response of the 'first responding author', i.e., the author with the lowest author number.

Procedure

The online survey was generated using Qualtrics software version 500235 (Qualtrics, 2012). We programmed the survey in such a way that each respondent was asked the same questions, but that the questions pertained to a specific article published by the individual respondent. In the invitation to the survey, we explicated ethical issues (see below) and stated that survey responses would be linked to the accuracy of the *p*-values in the article with which the survey questions were concerned. In addition, we provided the first author's email address for respondents to write to if they had further questions before deciding whether to participate.

At the beginning of the survey we encouraged respondents to have the articles near at hand by asking them to indicate how many authors were listed in the paper. Many articles reported more than one study. As different people may have contributed to different studies, the questions would have been difficult or even impossible to answer if they had pertained to all studies. Therefore, the respondents were presented with a set of six questions about the *first or only study* reported in the article asking them to specify who, as indicated by the author number (or 'other' category), was involved in: (1) conducting the statistical analyses, (2) writing down the sample details, (3) reporting the statistical results, and (4) checking the reported statistical results. The last two questions in this set pertained to data sharing and asked how many people (5) had access to the data when the manuscript was submitted, and (6) currently have access to the data. These six questions allowed us to construct six corresponding 'co-piloting' variables: if only one person was involved, the variable was coded '0' (not co-piloted), if two or more persons were involved, the variable was coded '1' (co-piloted). Finally, we asked respondents whether they wished to receive a report about the accuracy of the *p*-values reported in their article, and whether they wished to participate in a raffle in which they could win one of five $100 Amazon.com vouchers. The invitation e-mail and the survey itself can be found at the Open Science Framework via http://osf.io/ncvxg.

The Relationship between Co-Piloting and Statistical Reporting Errors

To analyze the relationship between co-piloting and the accuracy of *p*-values reported in the first or only study in the corresponding articles, we merged the data file containing the retrieved *p*-values from each article and the data file containing the survey responses. While *p*-values had been retrieved from 430 out of 697 articles, and survey responses were obtained for 346 articles, these sets of articles did not completely overlap (i.e., for some articles statistical results were retrieved but no survey response was obtained, and vice versa). In total, the data of 210 articles (48.8% of the 430 articles from which statistical results had been retrieved) could be matched. Thus, we matched each statistical result retrieved from the first (or only) study reported in these articles to the survey responses given by the respondent with the lowest author number. The statistical results of the remaining 220 articles were retained in the file to analyze the effect of non-response. A schematic overview of our sample is presented in Figure 3.1. Based on the study of Wicherts, Bakker, and Molenaar (2011) who found a relationship between will-

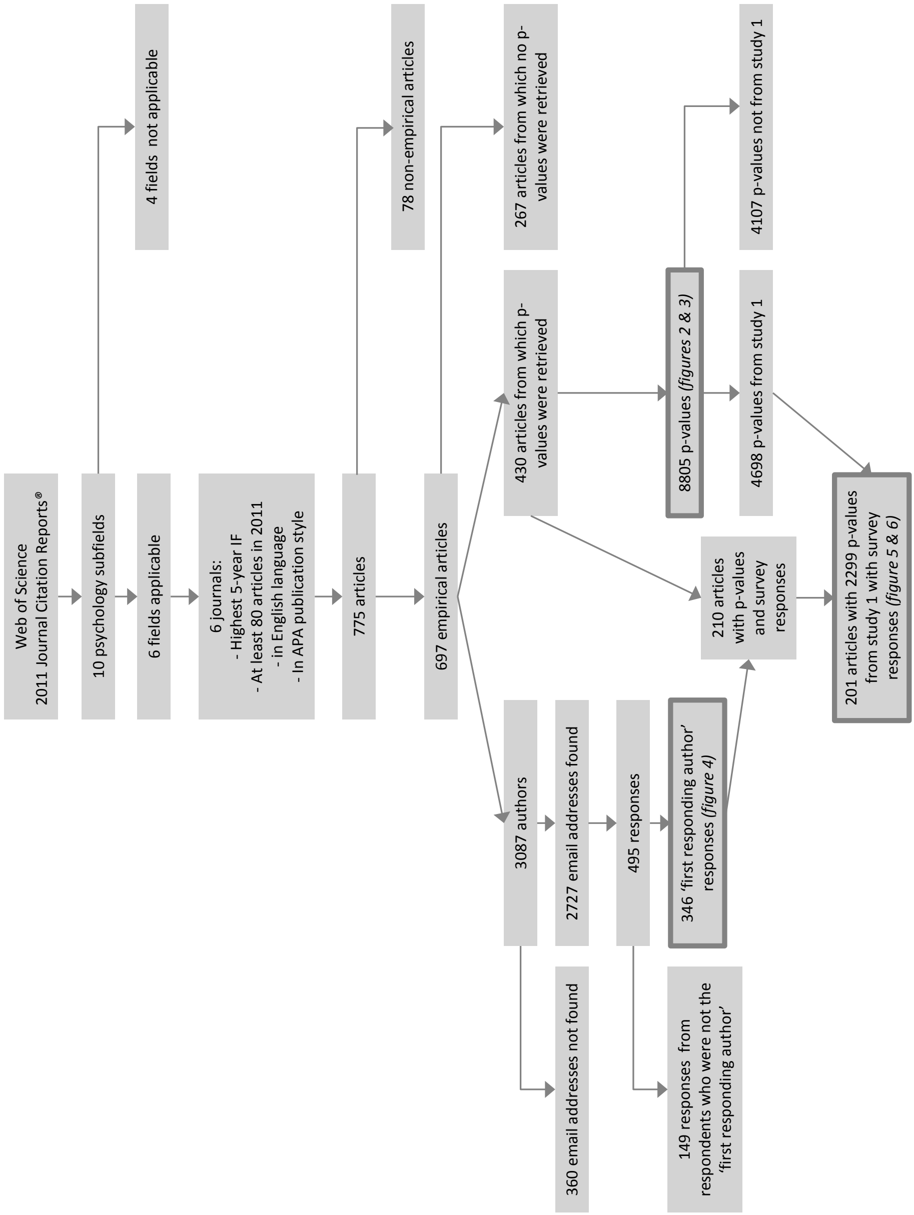

Web of Science
2011 Journal Citation Reports®
10 psychology subfields
4 fields not applicable
6 fields applicable
6 journals:
- Highest 5-year IF
- At least 80 articles in 2011
- in English language
- In APA publication style
775 articles
78 non-empirical articles
697 empirical articles
267 articles from which no p-values were retrieved
430 articles from which p-values were retrieved
8805 p-values (figures 2 & 3)
4107 p-values not from study 1
4698 p-values from study 1
210 articles with p-values and survey responses
201 articles with 2299 p-values from study 1 with survey responses (figure 5 & 6)
3087 authors
2727 email addresses found
360 email addresses not found
495 responses
346 'first responding author' responses (figure 4)
149 responses from respondents who were not the 'first responding author'

Figure 3.1 *Flow chart for composition of sample*

ingness to share research data and the prevalence of reporting errors in a sample of 48 articles, we expected to have enough power to detect a relationship between co-piloting and statistical reporting errors in our sample of 430 papers from which p-values were retrieved. With the 210 articles for which we obtained survey responses and had retrieved p-values, our sample was still more than four times as large as in Wicherts et al.'s study (2011).

Ethics Statement

This study was approved by the ethics committee of the Tilburg School of Social and Behavioral Sciences under the following conditions: (1) specific errors uncovered during this study would not be discussed in publications, presentations, writing or conversation with others, (2) the survey responses would be processed anonymously, and (3) survey respondents would receive feedback about the accuracy of the *p*-values in their article if they wished. In total, 384 respondents requested and received feedback via email. Respondents provided informed consent by ticking 'yes' at the statement 'I have read and understood the above and agree to participate' on the introductory page of the survey.

Statistical analysis

We uploaded our analytical plan regarding the confirmatory analyses at the Open Science Framework (https://osf.io/qutsy) before collecting the survey data. As our main hypotheses were tested by six analyses each, we corrected our alpha levels in these analyses for multiple testing by dividing .05 by six. All our analyses were conducted by at least two of the authors in order to reduce the probability of any errors on our own part. All scripts used to prepare the data files, to conduct the analyses and to construct the graphs, to anonymize our data, and to draw the winners of the raffle can be found on the Open Science Framework via http://osf.io/ekush.

Data availability

The first, non-merged anonymous survey data file can be viewed via http://osf.io/4bvqh. The data on the Open Science Framework are open access with no copy right issues and can be accessed by readers in the same manner as the authors. The second, merged data file contains *p*-values that can be traced back to individual articles, and can therefore not be shared without restrictions imposed by our ethics committee. The Psychology Ethics Committee of the Tilburg School of Social and Behavioral Sciences approved this study under the strict condition that we would not make these data file publicly available. We will however share these data after written agreement, and only with other researchers wishing to verify our results (see Article 8.14 of the APA ethical principles of psychologists and code of conduct (American Psychological Association, 2002)). Requests for data can be sent to the authors Coosje L. S. Veldkamp (C.L.S.Veldkamp@tilburguniversity.edu) or Jelte M. Wicherts (J.M.Wicherts@tilburguniversity.edu).

RESULTS

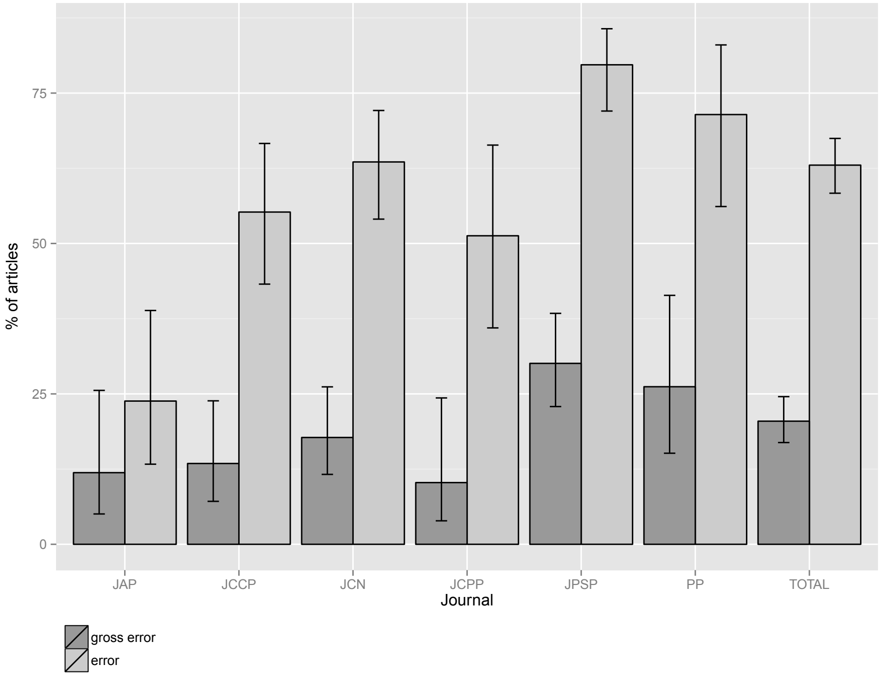
The Prevalence of Statistical Reporting Errors

Our first aim was to estimate the prevalence of statistical reporting errors in journals representing different areas of psychology using an automated procedure. First, we present the error rates at the article level: what is the probability that an article contains at least one *p*-value that comprises an error? As the dependent variable was dichotomous (the article does or does not contain at least one inconsistent *p*-value), we carried out simple logistic regression analyses (intercept only models) to estimate the probabilities and their 95% confidence intervals (CI). The results collapsed over all journals revealed that almost two out of three articles (63.0%, CI [58.4 – 67.5]) contained at least one *p*-value that comprised an error, and that one in five articles (20.5%, CI [16.9 – 24.5] contained at least one *p*-value that comprised a gross error.

We also compared the error prevalence across different journals/fields. Logistic regression analyses with journal as predictor revealed that there were differences in error rates between the journals: χ^2 (5, *N* = 430) = 49.46, *p* < .001. The probability that an article contained at least one *p*-value that comprised an error was lower in the *Journal of Applied Psychology* than in all other journals (23.8 %, CI [13.3 - 38.9], all *p*s ≤ .002 < 0.05/6) except the *Journal of Child Psychology and Psychiatry* (51.3%, CI [36.0 – 66.4], *p* = .012 > 0.05/6). At the same time, this probability was higher in *the Journal of Personality and Social Psychology* (79.7%, CI [72.0- 85.7], all *p*s ≤ .006 < 0.05/6) than in all other journals except *Psychophysiology* (71.4%, CI [56.1 – 83.0], *p* = .264 > 0.05/6). These differences may be attributable to differences between journals in the mean number of reported *p*-values per article, as a higher number of reported *p*-values entails a higher probability that an article contains an error. For example, an article in the *Journal of Personality and Social Psychology* contains more than 30 *p*-values on average, whereas the average article in the *Journal of Applied Psychology* contains only slightly more than eight *p*-values. The probability that an article contained at least one *p*-value that comprised a gross error differed also by journal: χ^2 (5, *N* = 430) = 15.46, *p* = .009, but no journal differed significantly from any other journal (all *p*s ≥ 0.012 > 0.05/6). The error probabilities for the sample as a whole and for each field separately are presented in Figure 3.2, together with their 95% confidence intervals.

Next, we present the results at the level of the individual *p*-value: i.e. what is the probability that a given *p*-value comprises an error? Because the dependent variable was again dichotomous (the *p*-value is either inconsistent or not) and because the *p*-values are nested within their articles, we carried out multilevel logistic regression analyses with article as random factor to estimate the probabilities and their 95% confidence intervals (CI). The results collapsed over all *p*-values

Figure 3.2 *The probability per journal that an article contains at least one p-value comprising an error or gross error (with 95% confidence interval).*

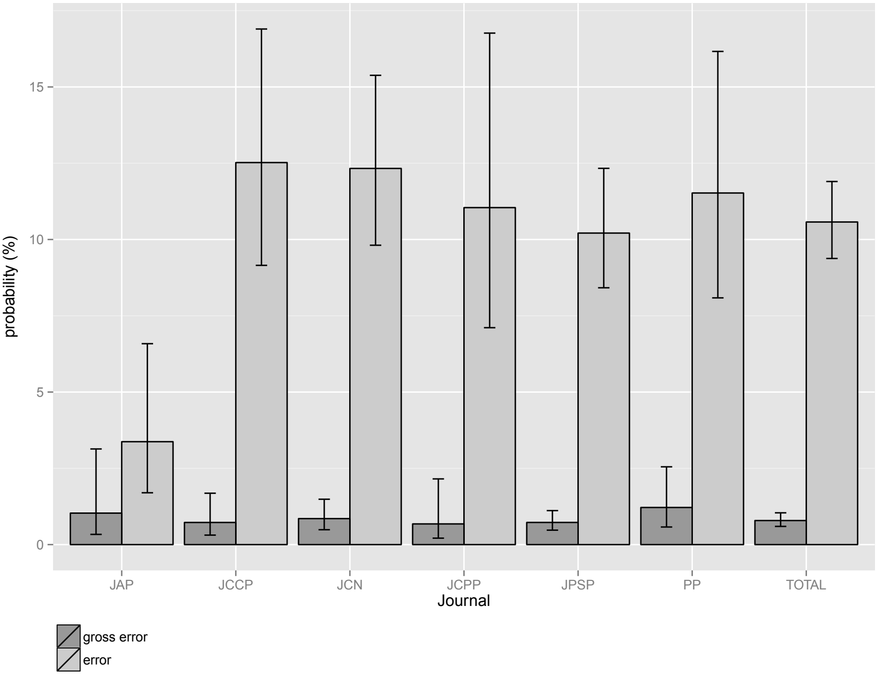


Note: JAP = Journal of Applied Psychology (*n* = 42); JCCP = Journal of Consulting and Clinical Psychology (*n* = 67); JCN = Journal of Cognitive Neuroscience (*n* = 107); JCPP = Journal of Child Psychology and Psychiatry (*n* = 39); JPSP = Journal of Personality and Social Psychology (*n* = 133); PP = Psychophysiology (*n* = 42); TOTAL = all articles together (N = 430).

showed that approximately one in ten *p*-values comprised an error (10.6%, CI [9.4 – 11.9]) and one in 125 *p*-values comprised a gross error (0.8%, CI [0.6 – 1.0]).

Running the multilevel logistic regression analyses with article as a random factor and journal as a fixed factor revealed that there were differences in the *p*-values' error probabilities between journals: χ^2 (5, *N* = 8105) = 17.53, *p* =.004. The probability that a given *p*-value comprised an error was lower in the *Journal of Applied Psychology* (3.4%, CI [1.7 – 6.6]) than in all other fields (all *p*s ≤ .004 < 0.05/6). One explanation for the lower error probability in this journal may be that its low mean number of reported *p*-values per article (8.10) renders errors more easily detectable by (co-)authors and other readers. The probability that a *p*-value comprised a gross error did not significantly differ between journals: χ^2 (5, *N* = 8105) = 1.92, *p* = .860. The error probabilities for the sample of *p*-values as a whole as well as for the *p*-values in each field separately are presented in Figure 3.3, together with their 95% confidence intervals.

Figure 3.3 *The probability per journal that a given p-value comprises an error or a gross error (with 95% confidence interval).*



Note: JAP = Journal of Applied Psychology (*n* = 340); JCCP = Journal of Consulting and Clinical Psychology (*n* = 833); JCN = Journal of Cognitive Neuroscience (*n* = 1,721); JCPP = Journal of Child Psychology and Psychiatry (*n* = 444); JPSP = Journal of Personality and Social Psychology (*n* = 4,018); PP = Psychophysiology (*n* = 749); TOTAL = all *p*-values together (N = 8,105).

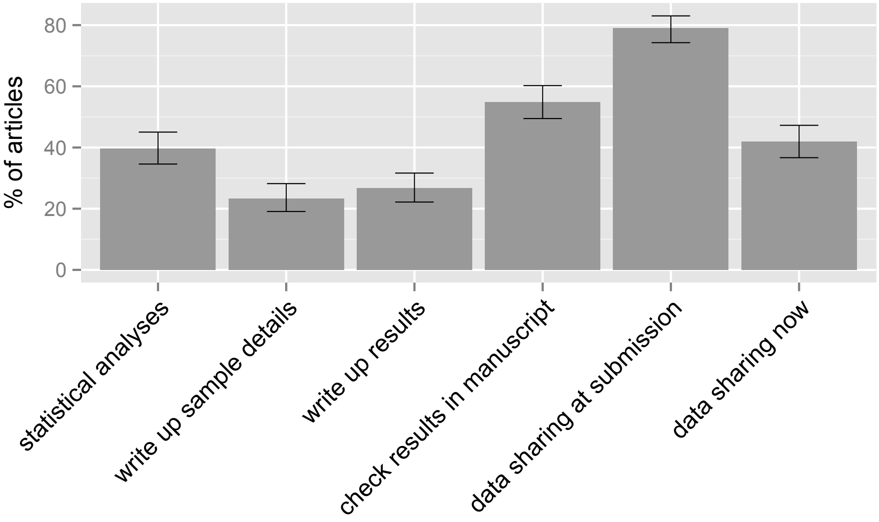
One may notice that for the *Journal of Applied Psychology (JAP)* the probability that a *p*-value comprises a gross error seems relatively high compared to the overall probability that a *p*-value in *JAP* comprises an error. However, we tested if the conditional probability of a gross error given an error was different across journals, and this does not appear to be the case: $\chi^2(5, N = 1149) = 9.09, p = .106$.

Co-Piloting in Psychology

Our second aim was to document the extent to which co-piloting currently occurs in psychological research. To answer this, we computed a co-piloting variable for each of the six co-piloting questions in the survey. Specifically, we computed for each of the processes (analyzing the data, writing down the sample details in the manuscript, writing down the statistical results in the manuscript, checking the results written down in the manuscript, sharing the data among co-authors before submission, and archiving the data after submission) how many people had been involved and coded those parts in which two or more people had been involved as 1 (co-piloted), and those parts in which only one person had been involved as 0 (not co-piloted). We ran a logistic regression analysis (intercept model only) to estimate the probabilities and 95% confidence intervals for each of the six processes (see Figure 3.4). Because journal was no significant predictor in any of the six analyses (all *p*s ≥.015 > 0.05/6), the results presented in Figure 3.4 are collapsed over journals. Note that the sample sizes for the individual analyses differed slightly due to some missing values that were excluded pairwise (see note below Figure 3.4).

As can be seen in Figure 3.4, the statistical analyses were most often conducted by one person only: co-piloting occurred in just 39.7% of the articles (CI [34.6% – 45.0%]. Similarly, in most articles, only one person wrote down the sample details and the statistical results in the manuscript (co-piloting occurred in 23.3 % (CI [19.1%- 28.2%]) and 26.6% (CI [22.2%- 31.6%]) of the articles, respectively). However, the results of the analyses as written down in the manuscript were checked by a second person slightly more often than not (54.9%, CI [49.5%- 60.2%). On the other hand, in most articles, the data had been shared with at least one other person

Figure 3.4 *The percentage of articles in which co-piloting occurred for various processes (with 95% confidence intervals).*



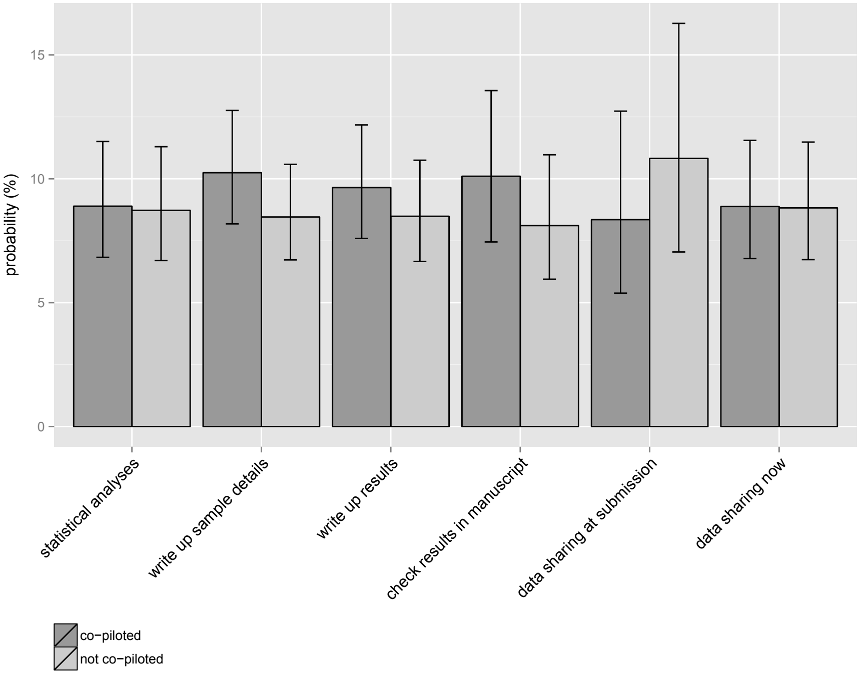
Note: statistical analyses = conducting the statistical analyses (N= 335); write up sample details = writing the sample details in the manuscript (N = 330); write up results = writing up the results in the manuscript (N = 334); check results in manuscript = checking of the results in the manuscript by someone other than the person who wrote up the results in the manuscript (N= 326); data sharing at submission = having access to the data at the moment the manuscript was submitted (N= 333); data sharing now = having access to the data at the moment the survey was being filled in (N= 332).

when the manuscript was submitted (79.0%, CI [74.3% - 83.0%], meaning that at least one other person had the opportunity to look at the data set before the article was published. Less than two years after publication however, data storage by more than one person occurred only in the minority of cases (41.9%, CI [36.7%- 47.2%].

The Relationship between Co-Piloting and Reporting Errors

Our third aim was to establish whether a relationship exists between co-piloting and the probability that a *p*-value comprised an error. To answer this question, we only took into account those *p*-values of articles of which at least one author responded to our survey. By means of a multilevel logistic regression analysis with article as random factor and journal as fixed factor we first established that in this subsample there were no differences between journals in the probabil-

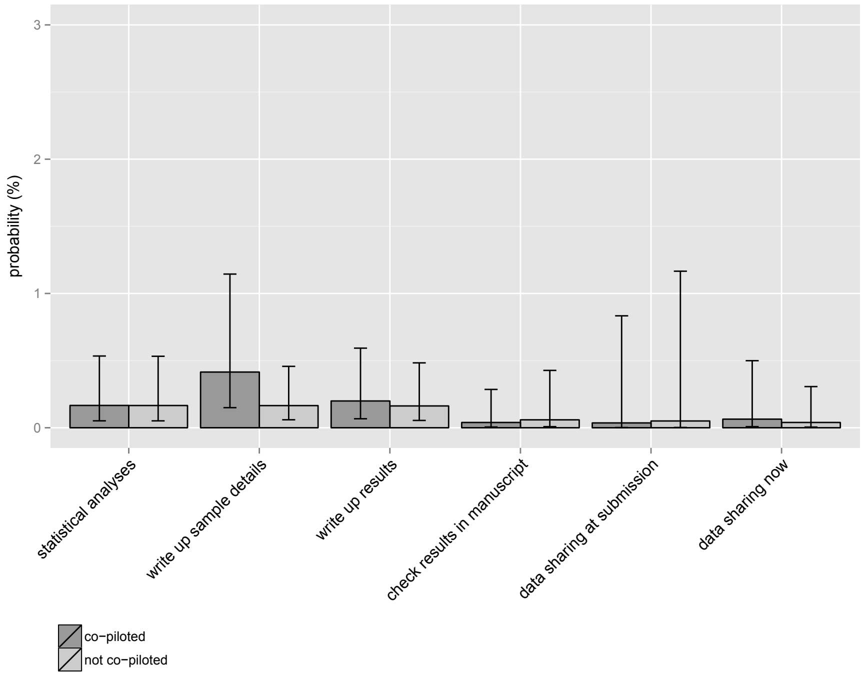
Figure 3.5 *The probability that a p-value in the first (or only) study reported comprises an error: co-piloted studies versus non-co-piloted studies.*



Note: statistical analyses = conducting the statistical analyses (N= 2,247); write up sample details = writing the sample details in the manuscript (N = 2,215); write up results = writing up the results in the manuscript (N = 2,231); check results in manuscript = checking of the results in the manuscript by someone other than the person who wrote up the results in the manuscript (N= 2,185); data sharing at submission = having access to the data at the moment the manuscript was submitted (N= 2,228); data sharing now = having access to the data at the moment the survey was being filled in (N= 2,226).

ity that a *p*-value comprised an error (χ^2 (5, *N* = 2299) = 6.35, *p* = .274), nor in the probability that a *p*-value comprised a gross error (χ^2 (5, *N* = 2299) = 4.15, *p* = .528). We then ran six different multilevel logistic regression analyses, each with article as random factor, and one of the six co-piloting variables as fixed factor. Our hypothesis that co-piloting is related to the probability that a given *p*-value was associated with a reduced error risk lacked support: we found no significant differences for any of the six processes between articles in which co-piloting had occurred and articles in which co-piloting had not occurred in the probability that a given *p*-value was inconsistent (all *p*s ≥ .283 > 0.05/6, see Figure 3.5), nor in the probability that a *p*-value was inconsistent to the extent that it may have affected a decision about statistical significance (all *p*s ≥.323 > 0.05/6, see Figure 3.6).

Figure 3.6 *The probability that a p-value in the first (or only) study reported comprises a gross error: co-piloted studies versus non-co-piloted studies.*



Note: statistical analyses = conducting the statistical analyses (N= 2,247); write up sample details = writing the sample details in the manuscript (N = 2,215); write up results = writing up the results in the manuscript (N = 2,231); check results in manuscript = checking of the results in the manuscript by someone other than the person who wrote up the results in the manuscript (N= 2,185); data sharing at submission = having access to the data at the moment the manuscript was submitted (N= 2,228); data sharing now = having access to the data at the moment the survey was being filled in (N= 2,226).

Non-Response

Finally, we studied the effect of non-response by comparing the error probabilities in articles for which we obtained survey responses to the error probabilities in articles for which we did not obtain survey responses. Responses to the survey were not significantly associated either with the probability that an article contained at least one *p*-value that comprised an error (Wald *Z* =-0.882, *p* = .378), or with the probability that an article contained at least one *p*-value that comprised a gross error (Wald *Z* =-1.308, *p* = .191). These results indicate that the probability that an article contained at least one *p*-value that comprised an error did not seem to be associated with whether the authors responded to the survey.

At the *p*-value level, there was an effect of response on the probability that a given *p*-value comprised an error (Wald *Z* =-2.194, p = .028), but there was no significant effect of response on the probability that a given *p*-value comprised a gross error (Wald *Z* =-1.819, *p* = .069). In other words, these results indicate that the probability that a *p*-value comprised an error was higher in articles published by authors who did not respond to the survey than in articles published by authors who did respond to the survey, but that no such association was found with respect to the probability that a *p*-value comprised a gross error.

DISCUSSION

We estimated the prevalence of inconsistent *p*-values in six top psychology journals by means of an automated procedure to retrieve and check errors in the reporting of statistical results, in order to replicate earlier estimates of error rates in the psychological literature (Bakker & Wicherts, 2011; Bakker & Wicherts, 2014a; Caperos & Pardo, 2013; Leggett et al., 2013). Our results show a somewhat higher probability for articles to contain at least one *p*-value that comprises an error compared to the two studies by Bakker and Wicherts (63% vs. 45% (2014a) and 50% (2011)), and a higher probability for articles to contain at least one *p*-value that comprises a gross error (20% vs. 15% (2011; 2014a)). Our error probability estimates at the article level may be somewhat higher because the top journals in our sample typically require more than one study and hence include the results of more tests than the lower-ranked journals in Bakker and Wicherts' (2011) study. Our estimate of the probability that a *p*-value comprises an error was in between the estimates of Bakker and Wicherts: 10% vs 7% (2014a) and 18% (2011). A possible explanation for the difference with the higher estimate of 18% (Bakker & Wicherts, 2011) is that in the first study by Bakker and Wicherts (2011), statistical results that were not exactly reported as prescribed by the APA manual (American Psychological Association, 2010) were counted as errors, whereas in their later study (Bakker & Wicherts, 2014a) and in our study, this type of error was not taken into account. On the other hand, our error prevalence estimates may have been somewhat inflated by excluding reported statistical results that included an effect size. If we would assume that reporting effect sizes is associated with more knowledge about statistics, authors who reported effect sizes may have made fewer mistakes in reporting. In any case, our estimates are alarmingly high: almost two out of three of the articles published in one of these flagship journals contain at least one statistical reporting error and one in every ten reported p-values is inconsistent with the reported test statistic and the accompanying degrees of freedom.

Moreover, we documented the extent to which co-piloting currently occurs in psychology. Ours is the first study that looked at how often psychology researchers work together on the analyses and reporting of results and at how often data are shared among co-authors. Although 99.1% of the articles had more than one author, not all co-authors appear to feel shared responsibility for the accuracy of the data analysis. In most articles the analyses were conducted by one person only and the results in the manuscript were checked only slightly more often than not by a co-author or someone else. We realize however that ‘checking the results in the manuscript’ may not actually constitute re-analysis or recalculation of the *p*-values and that the term ‘checking the results’ may therefore have been somewhat ambiguous. Yet data sharing among co-authors seems quite common: the results indicated that data from four out of five articles had been shared among at least two authors at the time the manuscript was submitted. This means that at least one co-author had the opportunity to have a look at the data file before submission, although this does not mean that they have actually done so, and if they did, in what way they inspected the data. On the other hand, we find it rather disconcerting that even if the data were shared before submission, the data of more than half of the published articles are currently stored by one person only. If the data are stored in a safe place (e.g., in a data repository, or in the ‘cloud’), this may not constitute an archiving problem (specifically if they are well documented). However, if the data are stored on one researcher’s hard drive, the risk of loss of the data is considerable. Recent results show that the availability of research data declines rapidly over time (Vines et al., 2013), notwithstanding that ethical guidelines (American Psychological Association, 2002) and professional standards require the archiving of data for at least five years after publication. Sharing data with co-authors requires rigorous documentation, which is likely to increase the chances that data are still available for re-analyses, verification, and further use in the future (Simonsohn, 2013; Wicherts, 2013; Wicherts & Bakker, 2012; Wicherts, Borsboom, Kats, & Molenaar, 2006). Finally, our survey results show that in the majority of articles only one author wrote down the sample details and the statistical results. However, one could argue that these two variables may not have captured the concept of co-piloting very well, as it may not have been clear to our respondents how we actually envisioned co-piloting on actual writing. Our intention was to measure whether the sample size details and the results were discussed between authors before/during the actual writing process, as we believe that some errors may occur in this phase and may be reduced by discussion of the results and the output of the analyses among co-authors. Such aspects should be subject to further study. Another interesting avenue for further research would be a more fine-grained analysis of specific roles of each author and potential differences between responding authors in their responses to co-piloting questions.

Finally, we looked at whether co-piloting on statistical analyses, reporting of results, and data sharing among co-authors was associated with a reduced risk of statistical reporting errors. Contrary to our expectations, we did not find support for this relationship. A relationship may simply not exist, but we believe that the relationship may have been obscured by confounding mechanisms. For instance, our reliance on self-report may have produced socially desirable responses (in this case, answers indicating shared responsibility). The fact that we asked respondents to indicate *which authors* were involved in each part of the processes rather than asking *how many* people were involved may have also rendered the survey more sensitive. Another factor that may have played a role is that the difficulty of the statistical analyses may have increased both the error probability of the reported statistical result and the probability that the authors collaborated on the statistical analyses, which may have offset the effect of collaboration on the error probability. Finally, the finding that the probability that a given *p*-value was inconsistent was higher in articles of which the authors did not respond to our survey may be an indication that authors who worried that some of their *p*-values might turn out inconsistent were less inclined to respond to our survey. Note, however, that the relation between responses and inconsistent *p*-value probability was weak. Finally, because we could only use those statistical results that were part of the first or only study reported in the articles and because those results could not always be matched to survey responses, our sample size (and hence our statistical power) turned out lower than we expected.

Even if co-piloting turns out not to be associated with a reduced risk of statistical reporting errors, we do believe that co-piloting helps to intercept other human errors in the use of statistics and in scientific research in general. The risk of many forms of slips and lapses, to which experts in any field are particularly prone (Reason, 1990) should diminish considerably when more than one person is involved (Beaty, 2004; Wiegman & Shappell, 2003). In addition, co-piloting may benefit science by requiring transparency: co-piloting among co-authors requires proper data documentation, data archiving, openness, and discourse about statistical and methodological decisions. Most articles concerning data sharing focus on data sharing with people outside the research group (Ceci, 1988; Ceci & Walker, 1983; Vogeli et al., 2006; Wicherts & Bakker, 2012; Wicherts et al., 2006), but we believe that sharing of data as well as of methodological and statistical decisions ought to start within the research group, i.e., among co-authors themselves. Even if full co-piloting as defined in this article is not feasible due to time or other constraints, we encourage authors to implement at least some double-checking of data files, analysis scripts, and results into their routines. For example, this double-checking could be part of regular PhD-student supervision, fostering acuity on the part of both the student and the supervisor(s). At the same time, such a practice would set an example in emphasizing the importance of meticulousness in data analysis and reporting.

There have been suggestions for journal editors to increase author accountability by requesting a description of author contributions to each stage of the research process (Balon, 2005), a policy that enhances transparency and accountability and has now been adopted by a number of journals including the *Journal of the American Medical Association, The Lancet, PLOS ONE,* and *Psychological Science*. Finally, reporting confidence intervals and effect sizes as suggested by many statisticians trying to improve the use and reporting of statistics (e.g. (Cumming, 2014) and as prescribed by the APA's updated reporting guidelines (American Psychological Association, 2010) may reduce error rates as this requires more scrutiny in interpreting results and may allow (co-)authors (and other readers) to quickly spot striking inconsistencies between reported numbers. Like all fields of science, psychological science depends on the accuracy of the results reported in its literature. Human error of all forms is a part of science, but scientists nonetheless have the responsibility to eliminate as much error as possible.

CHAPTER 4

Shared responsibility for statistical analyses and statistical Reporting errors in psychology articles published in PLOS ONE (2003 – 2016)

This manuscript will be submitted as Veldkamp, C. L. S., Hartgerink, C. H. J.,
van Assen, M. A. L. M., & Wicherts, J. M. Shared responsibility for statistical analyses and statistical reporting errors in psychology articles published in PLOS ONE (2003 – 2016).

ABSTRACT

While it has become clear that many of the statistical test results reported in the psychological literature contain errors, it remains unclear which factors contribute to these errors and how error rates can be reduced. One solution that has been proposed is the ‘co-pilot’ model, where at least two of a paper’s authors independently run all of the analyses in order to verify the results. In our previous work, employing a relatively small sample of psychology articles (n = 697) and using survey methodology to measure co-piloting, we failed to find support for the effectiveness of co-piloting to reduce error rates. In the current study, we examined all psychology articles ever published in PLOS ONE (n = 14,946) for statistical reporting errors using the automated procedure ‘statcheck’, and used the author contributions sections to derive author responsibilities. Although we confirmed the alarmingly high prevalence of statistical reporting errors, we found no support for a relationship between reporting errors and shared responsibility for the analyses. Secondary analyses failed to reveal additional relationships between reporting errors and the number of authors on a paper, the number of authors responsible for the analyses, or the first author being responsible for the analyses. We discuss several potential best practices that may contribute to the reduction of both errors and biases in psychological research, including use of statcheck, pre-registration, well-organized project management, rigorous data documentation and archiving, and use of programs that automatically insert all statistical results into a text file, such as R Markdown.

The most widely used method of statistical inference in psychological research is null hypothesis significance testing (NHST; Cohen, 1994; Hubbard & Ryan, 2000; Krueger, 2001; Nickerson, 2000; Wetzels et al., 2011). The results of statistical tests using NHST typically contain three elements: the test statistic, the degrees of freedom of the test statistic, and the *p*-value associated with the combination of the test statistic and the degrees of freedom. When one or more of these three elements are incorrectly reported, the statistical result will be inconsistent. Such an inconsistency can vary from a difference that does not have any effect on the conclusions based on the statistical test (further called an 'error'; Bakker & Wicherts, 2011; Bakker & Wicherts, 2014a; Veldkamp et al., 2014; Wicherts et al., 2011) to a difference affecting the overall conclusions of a study (further called 'gross error'; Bakker & Wicherts, 2011; Bakker & Wicherts, 2014a; Veldkamp et al., 2014; Wicherts et al., 2011), but recently also called ''gross inconsistency' (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Errors and gross errors in the reporting of statistical results can occur in different ways: the elements in the reported test result can be incorrectly retrieved from the output of the statistical software used to conduct the analysis, they can be incorrect due to mere typographic errors or rounding errors, or they can be incorrect due to confirmation bias or intentional misreporting (Agnoli et al., 2017; John et al., 2012).

Over the last decade, it has become clear that many of the statistical test results reported in the psychological literature contain (gross) errors (Bakker & Wicherts, 2011; Bakker & Wicherts, 2014a; Berle & Starcevic, 2007; Caperos & Pardo, 2013; Garcia-Berthou & Alcaraz, 2004; Nuijten et al., 2016; Veldkamp et al., 2014; Wicherts et al., 2011). Bakker and Wicherts (2011) manually retrieved and recomputed 4,077 statistical results from 281 psychology articles and found around 18% of the results to be erroneous, with 1.2% of results showing a gross error. More recently, Epskamp and Nuijten (2015) developed 'statcheck', software to automatically retrieve and check statistical results from articles that contain results that are reported according to the publication manual of the American Psychological Association (2010). Statcheck is used to examine the error rates in the psychology literature (Hartgerink, van Aert, Nuijten, Wicherts, & van Assen, 2015; Nuijten et al., 2016; Veldkamp et al., 2014), is currently being employed at several journals (including *Psychological Science* and *PsychOpen*) to screen for errors during the peer review process, and is undergoing developments allowing use for articles written according other reporting styles. The largest and most recent of studies employing statcheck (Nuijten et al., 2016) concluded that almost 50% of psychology articles published between 1985 and 2013 contained at least one error and almost 13% at least one gross error. Of all statistical results reported, almost 10% constituted an error, and 1.4% constituted a gross error. Errors in reported results may not only have consequences for the validity of conclusions based on these results, but also for meta-analyses that calculate effect sizes on the basis of the reported results (Bakker & Wicherts, 2011; Gøtzsche, Hróbjartsson, Marić, & Tendal, 2007; Levine & Hullett, 2002; Nuijten et al., 2016) and for meta-research using reported *p*-values (e.g. Simonsohn et al., 2014b; van Assen, van Aert, & Wicherts, 2015). Generally, the occurrence of reporting errors negatively affects trust in the accuracy of the results in published articles.

To date, it remains unclear which factors contribute to these errors and how error rates can be reduced. Veldkamp et al. (2014, Chapter 3) investigated whether 'co-piloting' applied to statistical analyses, as proposed by Wicherts (2011), was associated with reduced error rates in articles published in the flagship journals of six major areas of psychological science. The idea of ço-piloting on statistical analyses is that at least two of a paper's authors store copies of the data and independently run all of the analyses in order to verify the results (Wicherts, 2011; Wicherts et al., 2011). Such a rigorous practice requires proper documentation of the data and analyses to allow co-authors (and peers) to double check the results (Wicherts et al., 2011). Veldkamp et al. hypothesized that more than one author having been involved in the statistical analyses implied that the results had been double checked and sufficiently documented to allow verification, and that involvement of more than one author in the analyses would therefore be associated with a lower likelihood of errors in the reported results.

While Veldkamp et al. confirmed the high reporting error rates in psychology and found support for the notion that involvement of more than one author in the statistical analysis is rather uncommon among psychologists, they failed to find an association between shared involvement and error rates. However, Veldkamp et al.'s data had several limitations that may have restricted the study's ability to detect this potential association. First, estimation of collaboration on statistical analysis was based on a questionnaire asking authors to recall who had been involved in the statistical analyses reported in a paper they had published over a year ago, which may have yielded inaccurate responses. Second, the data resulting from the questionnaire could have been subjected to response biases and social desirability, as some of the survey respondents indicated that they considered the issue of reporting errors to be rather sensitive. Third, the study potentially suffered from lack of statistical power due to the use of a relatively small sample of articles, and the choice to consider only the results reported in the first (or only) study reported in these articles and collaboration that had occurred in these first (or only) studies. For these reasons, we decided to conduct another study to investigate the relationship between co-piloting and errors in reported statistical results without any of these limitations; we made use of a considerably larger set of articles, and did not make use of questionnaire but rather used publically available indicators of the number of authors being responsible for the statistical analyses.

In studies of reporting errors it has become customary to examine both the probability that an article contains at least one (gross) error, and the probability that a statistical result constitutes a (gross) error (Bakker & Wicherts, 2011; Berle & Starcevic, 2007; Garcia-Berthou & Alcaraz, 2004; Nuijten et al., 2015; Veldkamp et al., 2014). In line with this, we here report all of these probabilities whenever relevant. Specifically, we first examined whether the probabilities that an article contains at least one (gross) error, and the probability that a statistical result constitutes a (gross) error were lower when more than one author was responsible for the analyses than when only one author was responsible for the analyses. We then investigated the potential role of the first author. It is widely believed that the chance that an article will be published is higher when hypotheses are confirmed than when hypotheses are not confirmed (e.g. Franco, Malhotra, & Simonovits, 2014), making statistical analysis prone to confirmation bias, which may in turn lead to reporting errors. As the first author typically has a higher interest in getting the article published than co-authors have, we examined whether the probabilities of reporting errors were higher when the first author was responsible for the analyses than when the first author was not responsible for the analyses. All confirmatory hypotheses were pre-registered on the Open Science Framework and can be accessed through https://osf.io/8n7mj/ along with all data and scripts used in this study.

We also exploratively examined two additional factors that are potentially related with the probability that reporting errors occur. These two factors are related to diffusion of responsibility (Wallach, Kogan, & Bem, 1964) and the potential for 'social loafing' (Karau & Williams, 1993; Latane, Williams, & Harkins, 1979), in which the mere presence of many others and the expectation that others will bear certain responsibilities might lower individual's diligence. In the present context, having too many people responsible for the statistical analyses or having a very large number of authors on the article may increase the probability of making errors in the reporting of statistical results. In these explorative analyses, we therefore examined the associations between the probabilities of statistical reporting errors and the number of authors responsible for the statistical analyses and between the probabilities of statistical reporting errors and the number of authors on the article.

METHOD

Articles published in journals of the Public Library of Science (PLOS) are Open Access and provide an overview of author contributions in a standard format, including which authors conducted the analyses of a given article. Therefore, the full collection of articles published in PLOS journals provides an excellent large-scale database to study the relation between co-piloting of the analyses and reporting errors. To this end, we first downloaded all available meta-data of all articles ever published in the journals of PLOS until 31 December 2016 using the R package 'rplos' (v0.6.4; Chamberlain, Boettiger, & Ram, 2015)). The full script is available through https://osf.io/hq43m/. From the obtained metadata we extracted for each article how many authors were responsible for the statistical analyses, whether the first author was among those being responsible for the analyses, and how many authors were listed. We also retrieved data on declarations of competing interests to study whether the presence of a competing interest increased the probability of gross errors due to confirmation bias. However, as it proved infeasible to automatically determine whether a competing interest had been declared due to the unlimited number of ways in which authors could declare the absence or presence of a competing interest, we did not pursue this in the current study (thereby deviating somewhat from our pre-registration). Second, we downloaded the full texts of all articles that had the tag 'psychology', and employed statcheck (v1.2.2; Epskamp & Nuijten, 2015) to retrieve and check the reported statistical results in these psychology articles.

From the downloaded meta-data of all articles published in PLOS journals and from the statcheck results based on the downloaded full texts of the psychology articles, we created four data files. The first file ('all meta') contains for each article published in all PLOS journals how many authors were responsible for the statistical analyses, whether the first author was among those being responsible for the analyses, how many authors were listed, whether a conflict of interest had been declared, the number of statistical results that had been retrieved by statcheck (if any), the number of statistical results flagged by statcheck as constituting a (gross) error (if any results had been retrieved), and whether statcheck flagged at least one statistical result as a (gross) error (if any results had been retrieved). This file (which also contains additional meta-data that are not used in the current study) can be found through https://osf.io/7tsxf/. The second file ('psych meta') contains a subset of the data in the first file: the meta-data for PLOS psychology articles only. This file can be found through https://osf.io/p9jm6/. The first and second file have as many cases (or rows) as articles. The third file ('all statcheck') contains the same meta data of all articles as the first file, but also includes all statistical results retrieved from each article by statcheck and accompanying data on whether the result was flagged as constituting a (gross) error. This file can be found through https://osf.io/wm6f3/. Finally, the fourth file ('psych statcheck') contains a subset of the data in the third file: the same data but then for PLOS psychology articles only. This file can be found through https://osf.io/u279t/. The third and fourth file have as many cases (or rows) as statistical results in all articles combined.

ANALYSIS PLAN

Our pre-specified analysis plan was pre-registered (see https://osf.io/8n7mj/). As pre-registered, we only included psychology articles published in PLOS ONE in our confirmatory analyses because over 90% of the psychology articles published in PLOS journals is published in PLOS ONE. Since we included *all* psychology papers ever published in PLOS ONE and thereby examined the whole population of (retrievable) statistical results ever reported in all psychology articles published in PLOS ONE, statistical testing (based on sampling) is superfluous in our study. Because of the use a population and lack or sampling, the error probabilities calculated by our models cannot be generalized to other journals or other scientific fields. Missing values in our study only occurred in the form of any of the relevant meta-data being unavailable for articles or no statistical results being retrieved from an article, or a combination of both. As pre-registered, when this occurred, the missing data points were excluded from the analyses based on pairwise deletion. Finally, although we overlooked this issue while pre-registering our study, we decided to delete (the same) three articles from all four data files: Bakker & Wicherts, 2014; Veldkamp at al., 2014; Wicherts et al., 2011. These were articles about statistical reporting errors, which deliberately listed examples of such errors. Leaving these articles in the data set would thus inadvertently have biased the calculations of the probabilities that reporting errors occur. We did not exclude any other data points as no obvious errors occurred during the extraction of the data (which was our pre-specified criterion for data exclusion). All scripts used to clean and analyze the data can be found through https://osf.io/pxdtv/.

RESULTS

Shared responsibility for statistical analyses in PLOS articles

Table 4.1 displays for each PLOS journal the number of articles, the number of authors per article, the number of authors responsible for the analyses per article, the percentage of articles that were co-piloted (i.e., where more than one author was responsible for the analyses), and the percentage of articles where the first author was responsible for the analyses.

Number of articles

In total, we obtained meta data of 187,163 articles. Out of these articles, 164,150 (87.7%) were published in PLOS ONE. Of the articles published in PLOS ONE, 14,946 (9.1%) had the tag 'psychology'. We verified that across all journals the vast majority of articles with the tag 'psychology was published in PLOS ONE (93.6%), justifying our pre-registered decision to solely focus on psychology articles published in PLOS ONE.

Number of authors per article

Across all fields and journals, the mean number of authors per article was 7.0 (SD = 4.9). In PLOS ONE, the mean number of authors per article was 6.9 (SD = 4.2). In psychology articles published in PLOS ONE, the mean number of authors was 5.3 (SD = 3.4).

Number of authors responsible for the analyses

Across all PLOS journals, the mean number of authors responsible for the analyses per article was 3.5 (SD = 2.2). In articles published in PLOS ONE, the mean number of authors responsible for the analyses was 3.4 (SD = 2.2). In psychology articles published in PLOS ONE, the mean number of authors responsible for the analyses was again somewhat lower, namely 2.7 (SD = 1.7).

Co-piloting

We considered an article 'co-piloted' when more than one author was responsible for the analyses. Across all journals, the percentage of articles that were co-piloted was 85.8%. In PLOS ONE, the percentage of articles that were co-piloted was 85.6%. The percentage of psychology articles in PLOS ONE that were co-piloted was 76.4%. The way we measured co-piloting in the current study yielded a percentage that was more than twice as high than the percentage yielded by the way we measured co-piloting in our previous work (39.7%; Veldkamp et al., 2014). There, we used a survey in which we asked first authors of the published articles included in our study to indicate how many people had been involved in the first or only study reported in their article. In the current study, we used the publicly available author contribution sections of the published articles included in our study and were not able to distinguish between the first study and other studies reported in the article.

The first author being responsible for the analyses

We also examined the percentage of articles in which the first author was responsible for the analyses. Across all PLOS journals, the first author was among those listed as responsible for the analyses in the vast majority of articles: in 91.21%. In PLOS ONE, the first author was responsible for the analyses in 90.80% of the articles, and in PLOS ONE psychology articles, this was 92.38%.

Table 4.1 *Relevant author contributions in PLOS articles.*

Journal	No. of articles	Mean no. of authors (SD)	Mean no. responsible for analyses (SD)	% of articles co-piloted	% of articles with 1st author responsible for analyses
PLOS Biol.	2,275	7.4 (7.6)	4 (2.6)	87.11	98.22
PLOS Comput. Biol.	4,404	4.4 (2.8)	2.6 (1.6)	75.14	94.43
PLOS Genet.	5,874	9.7 (13.1)	4.4 (3.1)	92.26	96.34
PLOS Med.	1,288	10.2 (9.9)	3.7 (3.4)	84.43	88.94
PLOS Neglected Trop. D.	4,230	8.6 (4.3)	3.9 (2.5)	89.31	93.79
PLOS Pathog.	4,942	8.5 (5.3)	4.7 (2.9)	93.47	96.45
PLOS ONE	164,150	6.9 (4.2)	3.4 (2.2)	85.55	90.80
PLOS ONE psych. articles	14,946	5.3 (3.4)	2.7 (1.7)	76.44	92.38
Total	187,163	7 (4.9)	3.5 (2.2)	85.76	91.21

Errors in statistical results reported in psychology articles published in PLOS ONE

Numbers of statistical results retrieved by statcheck

Statcheck retrieved statistical results from 4,178 (28%) of the psychology articles published in PLOS ONE, yielding a total of 48,496 statistical results. Across these psychology articles, the mean number of statistical results retrieved per article was 11.61 (SD = 12.60), and the median number of retrieved results was 7. The mean and median retrieved numbers of results are somewhat lower than in most major psychology journals, but in line with those found in flagship journals such as the *Journal of Applied Psychology, Psychological Science,* and the *Journal of Consulting and Clinical Psychology* (Nuijten et al., 2016; Veldkamp et al., 2014).

The probability that a psychology article in PLOS ONE contained at least one (gross) error.

The probability that a psychology article published in PLOS ONE contained at least one error was 40.57%, and the probability that a psychology article published in PLOS ONE contained at least one gross error was 9.77%. These results align with earlier results documented for the psychology journals *Journal of Applied Psychology, Psychological Science,* and the *Journal of Consulting and Clinical Psychology*, where the mean and median numbers of statistical results per article were similar to those in PLOS ONE (Bakker & Wicherts, 2011; Nuijten et al., 2016; Veldkamp et al., 2014). As the probability that an article contains at least one (gross) error is higher in journals with higher numbers of retrieved results per article, comparison of our results at the article level with results at the article level in journals with a different number of retrieved results is cumbersome.

The probability that a statistical result reported in a psychology article in PLOS constituted a (gross) error

At the level of the statistical results, there are three methods to calculate the probability that a statistical result reported in a psychology article in PLOS ONE constitutes a (gross) error. First, the simplest model (M1) simply calculates the probability as the total number of errors divided by the total number of statistical results. The problem with M1 is that it does not take into account the nested structure of the data. The second method (M2) calculates the probability by dividing for each article the number of errors by the number of reported statistical results, and then takes the mean value across the articles. The problem with M2 is that it weighs all articles equally, even though some articles may contain many more results than others. The third method (M3) calculates the probability by fitting a multilevel logistic regression null model to the data. Because transforming the fixed intercept of the logistic model to a probability does not correspond to the model's implied a probability when the variance of the random effect (i.e. of the article) exceeds zero (Raudenbush & Bryk, 2002), we integrated out the random effect to calculate the probability of an error. The problems with M3 are that its assumption of a normally distributed random effect may be violated, and recent simulation results highlighted that its estimation may not work well if both the probability of an error and the sample size at the article level are very small (Nuijten, 2017). To conclude, the probability of an error can be calculated with three different methods that are all meaningful and problematic at the same time, and may provide diverging results. We therefore report the results of all three methods, even though in our pre-registration (predating Nuijten, 2017), we only foresaw the use of M3.

The probabilities that a statistical result reported in a psychology article in PLOS ONE constituted an error as calculated by M1, M2, and M3, were 11.48%, 11.02%, and 11.09%, respectively. These probabilities are quite close to those found in the major psychology journals (9.7% according to M1 and 10.6% according to M2 (Nuijten et al., 2016)). The probabilities that a statistical result constituted a gross error as calculated by M1, M2, and M3 were 1.48%, 1.56%, and 5.16%, respectively. The probabilities as computed by M1 and M2 are again close to the probabilities in flagship psychology journals (1.4% and 1.6% respectively; Nuijten et al., 2016). The probability as computed by M3 cannot be compared to estimates in the literature as M3 was not used in previous research (other than in Chapter 3 of this dissertation). Because of the large differences in probabilities yielded by M1 and M2 on the one hand, and M3 on the other hand, the implausible implication of M3's probability that almost half of the errors are gross errors, and problems with fitting the multilevel logistic regression model to similar data on gross errors as ours (Nuijten, 2017), we have more faith in the probabilities of gross errors found by M1 and M2.

Associations between the probability of errors and author contributions

For clarity, we will first present our results at the level of the articles, followed by the results at the level of individual statistical results. To compute the probabilities of error in relation to author contributions at the level of the article, we fitted logistic regression models with the article containing at least one (gross) error (yes/no) as the dependent variable, and the type of author contribution variable (co-piloting / the first author being responsible for the analyses / the number of authors on the paper / the number of authors responsible for the analyses) as predictor. Subsequently, we converted the intercepts from the models to probabilities. To compute the probabilities at the level of the statistical results, we fitted multilevel logistic regression models with the article as random factor and the type of author contribution variable as fixed factor. Here, we integrated out the random effects from the models to calculate the probabilities of an error in the reported statistical result.

As explained in the section ‘analysis plan’, we refrained from statistical testing. However, for the sake of completeness and to align with our pre-registration, we provided results for tests of relevant differences in notes below the tables summarizing the results at the level of the article (Table 4.2) and at the level of the statistical result (Table 4.3).

Associations between the probability of errors and author contributions at the level of the article

The probability that an article contained at least one (gross) error and co-piloting (primary hypothesis).

Contrary to our hypotheses, the probability that an article contained at least one error was higher when co-piloting occurred (40.48%) than when co-piloting did not occur (37.11%). The same held for gross errors: the probability that an article contained at least one gross error was (marginally) higher when co-piloting occurred (9.70%) than when co-piloting did not occur (9.30%).

The probability that an article contained at least one gross error and the first author being responsible for the analyses (secondary hypothesis)

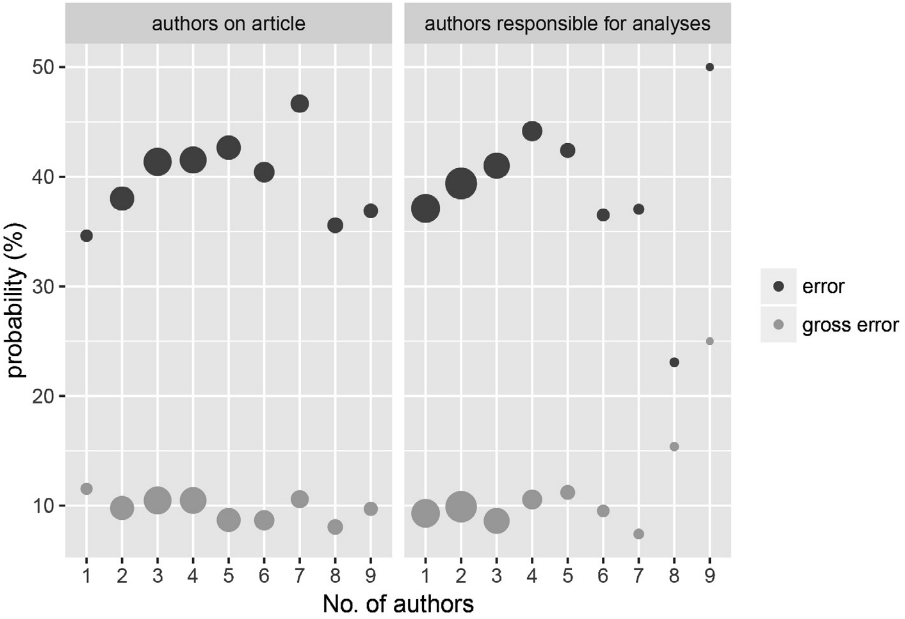
In line with our hypothesis, the probability that an article contained at least one gross error was slightly higher when the first author was responsible for the analyses (9.65%) than when the first author was not responsible for the analyses (8.21%).

Table 4.2 *Probabilities that an article contained at least one (gross) error and author contributions.*

Error type	Author Contribution Variable	Probability of error (%)	No. of articles	No. of articles with at least one (gross) error
error	Not co-piloted	37.11[i]	1,032	383
	Co-piloted	40.48	2,722	1,102
gross error	Not co-piloted	9.30[ii]	1,032	96
	Co-piloted	9.70	2,722	264
gross error	1st author not responsible	8.21[iii]	207	17
	1st author responsible	9.65	3,556	343

Note: [i] Wald Z = 1.89, ***p*** = .059; [ii] Wald Z = .37, ***p*** = 0.713; [iii] Wald Z = 0.68, ***p*** = .496.

Figure 4.1 *Probabilities that an article contained at least one (gross) error in relation to the number of authors and the number of authors responsible for the analyses.*



Note: The size of the dots represents the number of articles used to compute the probability.

The probability that an article contained at least one (gross) error and diffusion of responsibility (exploratory hypotheses)

We also examined the potential role of diffusion of responsibility by looking at the association of the probability that an article contained at least one (gross) error with both the number of authors on an article and the number of authors being responsible for the analyses. As can be seen in Figure 4.1, we found no evidence for these relationships. Please note that we cut the number of authors off at > 9, as the number of articles in which more than 9 authors were listed and/or responsible for the analyses strongly dropped to no more than a few observations.

Associations between the probability of errors and author contributions at the level of the statistical result

The probability that a statistical result constituted a (gross) error and co-piloting (primary hypothesis)

Contrary to our hypotheses, the probability that a result constituted an error was slightly higher when co-piloting occurred (11.87%, 11.36%, and 11.52% according to methods M1, M2, and M3 respectively) than when co-piloting did not occur (11.11%, 9.42%, and 9.66%). We found mixed results with respect to the hypothesis that the probability that a result constituted a gross error would be lower when co-piloting occurred than when co-piloting did not occur: according to M1 and M2, the probability that a result constituted a gross error was lower when co-piloting occurred (1.39% and 1.43% respectively) than when co-piloting did not occur (1.82% and 1.82%). However, according to M3, the probability that a result constituted a gross error was marginally higher when co-piloting occurred (5.29%) than when co-piloting did not occur (5.22%). Yet, as explained above, model M3 suffers from problems (Nuijten, 2017) that lead us to focus on the results based on models M1 and M2.

The probability that a statistical result constituted a gross error and the first author being responsible for the analyses (secondary hypothesis)

Contrary to our hypotheses, the probability that a result constituted a gross error was lower when the first author was responsible for the analyses (1.45%, 1.51%, and 5.25% according to methods M1, M2, and M3 respectively) than when the first author was not responsible for the analyses (2.80%, 1.92%, and 5.64%).

The probability that a statistical result constituted a gross error and diffusion of responsibility (exploratory hypotheses)

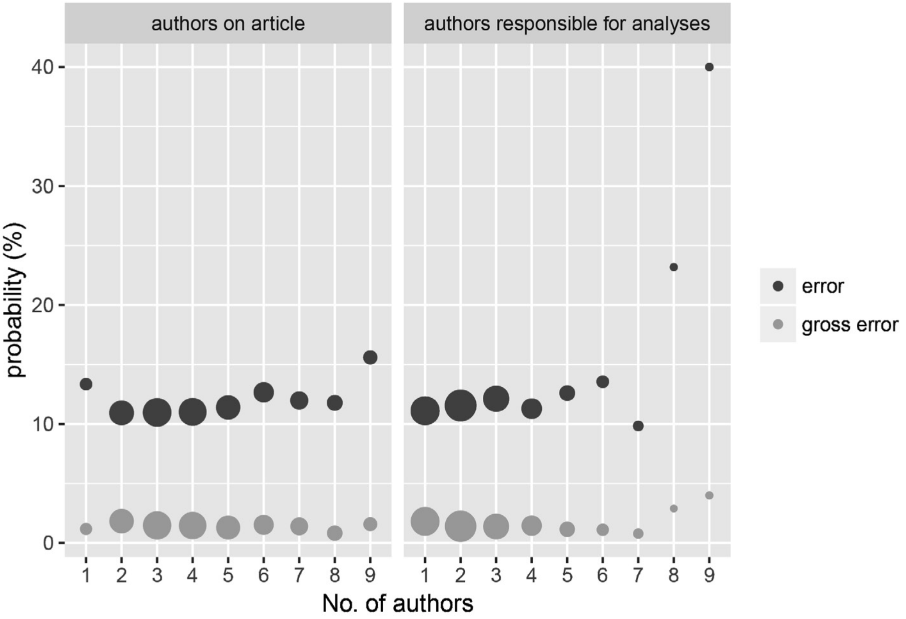
We also examined the potential role of diffusion of responsibility by looking at the association of the probability that a statistical result constituted a gross error with both the number of authors on an article and the number of authors responsible for the analyses. As can be seen in Figure 4.2, there was no strong evidence for these relationships. We again cut the number of authors off at > 9, as there were only few articles in which more than 9 authors were listed and/or responsible for the analyses.

Table 4.3 *Probabilities that a statistical result constitutes a (gross) error and author contributions.*

Error type	Author contribution variable	Probability of error M1	Probability of error M2	Probability of error M3	Variance random effect (M3)	No. of results	No. of results constituting a (gross) error
error	Not co-piloted	11.11	9.42	9.66[i]	2.89	11,952	1,328
	Co-piloted	11.87	11.36	11.52	2.89	31,078	3,688
gross error	Not co-piloted	1.82	1.82	5.22[ii]	22.62	11,952	217
	Co-piloted	1.39	1.43	5.29	22.62	31,078	433
gross error	1st author not responsible	2.80	1.92	5.64[iii]	22.65	1,819	51
	1st author responsible	1.45	1.51	5.25	22.65	41,263	599

Note: [i]: Wald Z = 2.9, ***p*** = .004; [ii]: Wald Z = 0.14, ***p*** = .885; [iii]: Wald Z =-0.36, ***p*** = .72.

Figure 4.2 *Probabilities that a statistical results constituted a (gross) error in relation to the number of authors and the number of authors responsible for the analyses.*



Note: The size of the dots represents the number of statistical results used to compute the probability. Although an increasing number of authors responsible for the analyses was statistically significantly associated with a higher probability that a statistical result constituted an error (Wald Z = 2.65, p = 0.008), this result is strongly influenced by the small number of articles in which 8 or 9 authors were responsible for the analyses. These yielded a total of only 69 and 75 statistical results respectively, compared to 1500- 11925 results yielded by articles with 1 to 5 authors responsible, 627 results yielded by articles with 6 authors responsible, and 254 results yielded by articles with 7 authors responsible.

DISCUSSION

We replicated error rates reported in earlier studies on statistical reporting errors in psychology (Bakker & Wicherts, 2011; Bakker & Wicherts, 2014a; Caperos & Pardo, 2013; Nuijten et al., 2016; Veldkamp et al., 2014; Wicherts et al., 2011) and medical fields (Berle & Starcevic, 2007; Garcia-Berthou & Alcaraz, 2004) in the full population of psychology articles published in PLOS ONE between 2003 and 2016. Just as in other psychology journals, the error rates in PLOS ONE psychology articles were alarmingly high: more than four out of then articles contained an error, and about one in ten articles contained an error that could have affected conclusions about statistical significance. These and previous findings suggest that the accuracy of results reported in the psychological literature cannot be taken for granted. However, we have no reason to assume that these error rates are unique to psychology. To date, the only other fields where the prevalence of reporting errors has been studied have been psychiatry (Berle & Starcevic, 2007) and medicine (Garcia-Berthou & Alcaraz, 2004), where the numbers were similar (albeit somewhat lower at the article level) to those in psychology. Studies of reporting errors in other scientific disciplines are therefore warranted. Currently, the software we used to automatically retrieve and re-compute statistical results, 'statcheck' (Epskamp & Nuijten, 2015), is undergoing developments that will allow its use in articles where results are reported according to reporting standards other than those of the American Psychological Association (American Psychological Association, 2010).

We found that co-piloting (here defined as declared shared responsibility for the statistical analyses) was ubiquitous in each of the PLOS journals, ranging from over 75% in PLOS Computational Biology to over 93% in PLOS Pathogens. In around 90% of the psychology articles in PLOS ONE, multiple authors had declared responsibility for the analyses. This result deviates from the earlier survey results (Veldkamp et al., 2014), where in only 39.7% of psychology articles multiple authors had been involved in the analyses of the data concerning the first or only study in the article.

We have not verified the author contributions via a survey or other means and we are not aware of any systematic study of the validity of author contribution statements among psychologists. In the medical literature however, concerns have been raised about the validity of the contribution disclosure forms (Bates, Anić, Marušić, & Marušić, 2004; Ilakovac, Fister, Marusic, & Marusic, 2007; Marušić, Bates, Anić, & Marušić, 2006). Hence it remains unclear why the percentages of co-pilot practices in the current study based on author contribution statements were so different than the percentages found in Veldkamp et al.'s (2014) survey. In the current study, we used the number of authors responsible for *all* analyses in the articles as a measure of co-piloting, while in our previous study (Veldkamp et al., 2014) we used the number of authors responsible for the *first or only study* reported in the article as a measure of co-piloting. Our current strategy allowed us to examine many more articles than our previous strategy (14,946 versus 346), but had the disadvantage that measuring the number of authors responsible for all analyses reported in a complete article does not tell us how many authors were responsible for each of the analyses of each of the individual studies within one article. That is, co-piloting as measured with author contribution statements does not exclude the possibility that some analyses in a paper were not checked by more than one author. Given the high prevalence of reporting errors, it is clear that more research into the precise (best and worst) practices in analyzing and checking of results is warranted.

We failed to find support for the notion that co-piloting reduced the probability of errors in the reporting of statistical results. Any differences between error probabilities in articles that were co-piloted and articles that were not co-piloted were marginal, and smaller than the differences in error probabilities that have been found between psychology journals (Bakker & Wicherts, 2011; Nuijten et al., 2016; Veldkamp et al., 2014). The marginal differences in combination with the inconsistent patterns rendered interpretation of these findings difficult. One interpretation of our results is, as previously alluded to, that many articles where more authors were responsible for the analyses did in fact not involve co-piloting, but rather having different authors bearing responsibility for different (types of) analyses in the article. However, when assuming that part of these articles actually did use co-piloting, we still would have found somewhat lower probabilities of (gross) errors if co-piloting had an effect. Another possible explanation of our findings may be that co-piloting does lower the probability of errors, but that this effect is outweighed by an opposite effect of diffusion of responsibility. That is, when sharing responsibility for the analyses, checking may occur in a less vigilant manner, resulting in multiple authors detecting as many errors as a single author who is fully responsible for all analyses. This alternative explanation is in line with our exploratory analyses examining potential effects of diffusion of responsibility, which yielded no support for the notion that an increasing number of authors or an increasing number of authors responsible for the analyses was associated with reduced probabilities of errors in reported statistical results.

We did not find consistent support for a higher probability of (gross) errors in papers where the first author was responsible for the analyses. The inconsistent patterns and the marginal differences again rendered interpretations of these findings difficult. However, here too one may reason that opposite forces cancelled out potential effects. On the one hand, there may be an adverse effect of confirmation bias affecting analyses by first authors more strongly than analyses by co-authors, leading to relatively more errors when the first author conducted the analyses. Yet on the other hand, first authors may have felt more responsible for the article and therefore also for accuracy of the reported results, leading them to verify and double check the results reported in the article when analyzing the data themselves. Given that we currently know very little about why errors or gross errors occur (notwithstanding surveys identifying misreporting as a questionable and possibly deliberate practice in the pursuit of significance; John et al., 2012; Agnoli et al., 2017), it remains difficult to interpret the (lack) of association between first author responsibility and reporting errors.

Reasons for our inability to find support for relationships between statistical reporting errors and shared responsibility for the statistical analyses may be sought also in the manner in which we defined the concept 'statistical reporting errors'. An inconsistency between a test statistic, its degrees of freedom, and its *p*-value may decrease trust in the article in which it is reported (and might affect meta-analyses and meta-research using the reported results), but is unlikely to reflect an error in the statistical analysis that yielded the result. Serious methodological and analytical flaws such as major violations of underlying assumptions or exploitation of degrees of freedom in the analyses in the quest for significance might not show up as inconsistencies in the reported results, while such flaws might still be countered by co-piloting. Having two instead of one expert involved in processes that are prone to human error has been found to help reduce the risk of error (Beaty, 2004; Lindvall et al., 2004; Reason, 1990; Wiegman & Shappell, 2003) and might well lower the risk of biased analytic results. In statistical analysis, this would translate to at least two authors openly discussing the analytical strategy to be taken, conducting the analyses collaboratively, and/or independently conducting the analyses while (cross-) checking the results. For this to work, the data file(s) need to be appropriately documented, analysis scripts need to be clear and properly annotated, and data files need to be shared between authors. This, in turn, will increase transparency, accountability, and the ease with which other researchers can verify the results reported in an article.

The use of NHST involves many, often arbitrary decisions that its users need to make (Bakker et al., 2012; Bakker & Wicherts, 2014a; John et al., 2012; Kriegeskorte, Simmons, Bellgowan, & Baker, 2009; Nieuwenhuis, Forstmann, & Wagenmakers, 2011; Simmons et al., 2011; Simonsohn et al., 2014a; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016; van Aert et al., 2016; Wagenmakers et al., 2011; Wicherts et al., 2016). Making as many of those decisions before the analyses are conducted or even before the study takes place forces authors to think through all potential pitfalls in the design and the analyses and to collaboratively formulate an analysis plan. When fully documented a priori, this strategy is known as study pre-registration (Chambers, 2013; de Groot, 1956/2014; Wagenmakers et al., 2012). Pre-registration may help reduce methodological and analytical error through fostering open discourse among co-authors. In an ideal pre-registration, the analysis script is also written (and ideally checked on a mock data set) prior to the execution of the analyses. Pre-registration can be challenging, and not all issues in a study can be foreseen, as for instance evidenced in our own pre-registration of the current study, where we did not anticipate the difficulty of retrieving data on competing interests and overlooked the need to remove our own articles about reporting errors from our data set. However, as long as all deviations from a pre-registration are transparently reported, the value of pre-registration remains high. User-friendly platforms such as the Open Science Framework (https://osf.io/) offer authors different versions of pre-registration, and a secure environment in which to manage research projects, safely archive all research files, and share files with co-authors, collaborators, or other researchers. The type of well-organized project management that these platforms offer, increases accountability and transparency and may thereby increase the likelihood of detection of methodological and statistical errors.

Returning to statistical reporting errors as examined in this study, we still consider finding ways to reduce these errors of utmost importance. Editors, reviewers, and readers of scientific articles need to be able to trust that the reported results are accurate. The human eye easily oversees errors and gross errors; even if all authors of a paper checked all results in their article, these errors may go unnoticed. Rather than requiring two or more authors to scan the results for inconsistencies, it may therefore be more effective and efficient to leave this scanning to computer programs such as statcheck. Any author can download this easy to use package, which comes with a clear manual, or use the website statcheck.io to readily check articles prior to submitting the article. In addition, journals can employ statcheck as part of the submission or peer review process. Statcheck is currently implemented as part of the peer review process by journals including Psychological Science and *PsychOpen*, and will soon be adopted by other journals, including PLOS ONE. Another way to reduce human error in reporting of statistical results is to compose manuscripts using programs in which scripts can be written that automatically insert all statistical results into a text file. An example of such a program is ‘R Markdown’ (http://rmarkdown.rstudio.com/), which is part of the open source software ‘R’. In R Markdown, authors write their complete manuscript text, or only the text of the results section of their manuscript, within the R environment. Instead of manually inserting individual results after viewing analysis output, authors insert the R code producing their results directly into the sentences reporting the results. For example, rather than writing ‘*p* = .123’, authors write ‘p = `r test_1$p.value`’. After writing a section like this, authors click a button that generates a Word file or pdf file containing the text with results corresponding to the inserted R code. This way, errors in retrieving and copying the right numbers from analysis output are avoided and results sections can be easily checked by co-authors.

To examine the effectiveness of the potentially error reducing strategies such as listed above, we suggest three possibly fruitful avenues for future research. First, we recommend comparisons of error rates in articles where statcheck was used during the peer review process with error rates in articles where statcheck was not used. The Tilburg University meta-research group is currently in the early phase of conducting such a study in collaboration with PLOS for psychology articles submitted to PLOS ONE. Second, studies in which error rates in articles published by researchers who pre-registered their study, or who rigorously managed their data and documented their analyses well on the Open Science Framework may be compared to error rates in articles published by researchers who failed to do this. Finally, error rates in articles that were composed using R Markdown may be compared to error rates in articles that were composed without the use of R Markdown. Preferably, such meta-research should employ randomization of articles or studies to different treatment conditions to preclude the possible alternative explanation that researchers who chose to use statcheck, pre-registration, rigorous management of data and analyses syntaxes, or R-markdown are also those researchers who tend to make less (or more) reporting errors, and to determine the actual effects of these best practices.

As in our earlier study (Chapter 3), we failed to find a clear relationship between co-piloting and statistical reporting errors. Despite the alarmingly high error rates, it remains unclear how psychologists currently analyze their data and which best practices may help in diminishing errors. We did however identify several potential best practices that may contribute to the reduction of both errors and biases in psychological research, including use of statcheck, pre-registration, well organized project management, rigorous data documentation and archiving, and use of programs that automatically insert all statistical results into a text file. These practices, especially when combined, would result in science with high levels of transparency and accountability. This way, co-piloting would entail creating a methodologically rigorous workflow wherein co-authors first agree on a pre-specified analysis plan or analysis script. The complete research process would be pre-registered, and all materials, data, and scripts would be shared among co-authors on a research project management platform. After data collection and analysis, at least two of the authors would independently run the analysis scripts (preferably using a program like R Markdown), and verify the results reported in the draft manuscript. Subsequently, statcheck would be run to check the consistency of the reported results. Finally, before or after submission of the manuscript, all well-documented materials, data, and scripts would be made public in such a way that others (peer reviewers or readers of the article) can readily and independently reanalyze the data and verify reported results. When data sharing is infeasible for ethical reasons, the data can be made shared in different ways to allow verification (Wicherts & Bakker, 2012).

CHAPTER 5

Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid *p*-hacking

This chapter is published as Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid *p*-Hacking. *Frontiers in Psychology, 7*(1832).

ABSTRACT

The designing, collecting, analyzing, and reporting of psychological studies entail many choices that are often arbitrary. The opportunistic use of these so-called researcher degrees of freedom aimed at obtaining statistically significant results is problematic because it enhances the chances of false positive results and may inflate effect size estimates. In this chapter, we present an extensive list of 34 degrees of freedom that researchers have in formulating hypotheses, and in designing, running, analyzing, and reporting of psychological research. The list can be used in research methods education, and as a checklist to assess the quality of preregistrations and to determine the potential for bias due to (arbitrary) choices in unregistered studies.

From the inception of the first study idea to the final publication, psychological studies involve numerous choices that are often arbitrary from a substantive or methodological point of view. These choices could affect the outcome of significance tests applied to the data, and hence the conclusions drawn from the research. These choices are also called researcher degrees of freedom (Simmons et al., 2011) in formulating hypotheses, and designing, running, analyzing, and reporting of psychological studies, and they have received considerable recent interest for two main reasons. First, researchers' opportunistic use of them greatly increases the chances of finding a false positive result (DeCoster et al., 2015; Ioannidis, 2005b; Simmons et al., 2011), or a Type I error in the language of Neyman-Pearson's variant of null hypothesis significance testing (NHST). Second, their strategic use in research may inflate effect sizes (Bakker et al., 2012; Ioannidis, 2008; Simonsohn et al., 2014a; van Aert et al., 2016). Hence, researcher degrees of freedom play a central role in the creation of (published) research findings that are both hard to reproduce in a reanalysis of the same data and difficult to replicate in independent samples (Asendorpf et al., 2013).

Among many potential solutions to counter inflated effects and elevated chances of finding false positive results caused by researcher degrees of freedom, one solution has received most attention: preregistration (Chambers, 2013; de Groot, 1956/2014; Wagenmakers et al., 2012). Preregistration requires the researcher to stipulate in advance the research hypothesis, data collection plan, specific analyses, and what will be reported in the paper. Although "planned research" more accurately describes this preregistered research, we will employ the commonly used term "confirmatory research" to describe it. An increasing number of journals now support preregistration for confirmatory research (e.g., Eich, 2014). In addition, over 35 journals now use a format of registered reports (Chambers, 2013) in which the registrations themselves are subject to peer review and revisions before the data collection starts, and the report is accepted for publication regardless of the direction, strength, or statistical significance of the final results. For instance, this format is now used in the journals *Cortex*, *Comprehensive Results in Social Psychology*, and *Perspectives on Psychological Science* (for Registered Replication Reports).

To disallow researchers to still use researcher degrees of freedom, it is crucial that preregistrations provide a *specific*, *precise*, and *exhaustive* plan of the study. That is, the ideal preregistration should provide a detailed description of all steps that will be taken from hypothesis to the final report (it should be specific). Moreover, each described step should allow only one interpretation or implementation (it should be precise). Finally, a preregistration should exclude the possibility that other steps may also be taken (it should be exhaustive). Hence, a preregistration specifies the project in such a way that all potential contingencies in formulating hypotheses, and designing, running, analyzing, and reporting are covered. For instance, the syntax for the statistical analyses should preferably be created in advance to be run (once) on the collected data to yield the final statistical results. Our own experiences with preregistration taught us that this specification is no easy task and that maneuverability remains if preregistrations are not sufficiently specific, precise, or exhaustive. For instance, just indicating one's use of a certain scale as the main outcome measure in an experiment typically does not preclude the researcher to attempt many different ways in how to score the items of the scale in his or her pursuit for statistical significance. A pre-registration should also be exhaustive because the stipulation that one will test Hypothesis A in a certain way does not preclude the possibility that one can *also* test Hypothesis B in the study. Therefore, for confirmatory aspects of the study, the word "only" is key (e.g., "we will only test Hypothesis A in the following manner").

The goal of this chapter is to present a list of researcher degrees of freedom that can be used in research methods education, as a checklist to assess the quality of preregistrations, and to determine the potential for bias due to (arbitrary) choices in unregistered studies. By pointing out many different researcher degrees of freedom, we hope to raise awareness of the risk of bias implicit in a lot of research designs in psychology and beyond. The list enables a charting of what Gelman and Loken (2014) dubbed the garden of forking paths in the analysis of data; i.e., the many different analytic decisions that could be or could have been made with a given data set. In what follows, we use the singular term researcher DF (degree of freedom) to mean a particular choice during the study, and the plural term researcher DFs when referring to multiple researcher degrees of freedom (or different types of choices).

Because NHST is by far the most used statistical framework used in psychology and related fields, we created the list of researcher DFs with NHST in mind. Other possible statistical frameworks are based on confidence intervals (Cumming, 2012), precision of effect size estimation (Maxwell et al., 2015), or Bayesian statistics (e.g., Kruschke, 2015). We note that most researcher DFs are relevant for all statistical frameworks. However, some researcher DFs need to be replaced or omitted (e.g. power analysis [D6], which is defined in the NHST framework) or added (e.g., selection of the prior, which is only used in Bayesian statistics) in approaches other than NHST. At this point, we therefore recommend using our list primarily for research using NHST.

We created the list in a qualitative manner; we as a group of methodologists studying researcher DFs, publication bias, meta-analysis, misreporting of results, and reporting biases, came up with a large list of researcher DFs, discussed these, and created a manageable list in several rounds of revision. We are aware that our list may not be exhaustive, but believe the list is a good starting point for a

Table 5.1 *Degrees of freedom in formulating the hypotheses, designing the study, collecting the data, analyzing the data, and reporting of psychological studies*

Code	Related	Type of Researcher Degree of Freedom
Hypothesizing		
T1	R6	Conducting explorative research without any hypothesis
T2		Studying a vague hypothesis that fails to specify the direction of the effect
Design		
D1	A8	Creating multiple manipulated independent variables and conditions
D2	A10	Measuring additional variables that can later be selected as covariates, independent variables, mediators, or moderators
D3	A5	Measuring the same dependent variable in several alternative ways
D4	A7	Measuring additional constructs that could potentially act as primary outcomes
D5	A12	Measuring additional variables that enable later exclusion of participants from the analyses (e.g., awareness or manipulation checks)
D6		Failing to conduct a well-founded power analysis
D7	C4	Failing to specify the sampling plan and allowing for running (multiple) small studies
Data collection		
C1		Failing to randomly assign participants to conditions
C2		Insufficient blinding of participants and/or experimenters
C3		Correcting, coding, or discarding data during data collection in a non-blinded manner
C4	D7	Determining the data collection stopping rule on the basis of desired results or intermediate significance testing
Data Analysis		
A1		Choosing between different options of dealing with incomplete or missing data on ad hoc grounds
A2		Specifying pre-processing of data (e.g., cleaning, normalization, smoothing, motion correction) in an ad hoc manner
A3		Deciding how to deal with violations of statistical assumptions in an ad hoc manner
A4		Deciding on how to deal with outliers in an ad hoc manner
A5	D3	Selecting the dependent variable out of several alternative measures of the same construct
A6		Trying out different ways to score the chosen primary dependent variable
A7	D4	Selecting another construct as the primary outcome
A8	D1	Selecting independent variables out of a set of manipulated independent variables
A9	D1	Operationalizing manipulated independent variables in different ways (e.g., by discarding or combining levels of factors)
A10	D2	Choosing to include different measured variables as covariates, independent variables, mediators, or moderators
A11		Operationalizing non-manipulated independent variables in different ways
A12	D5	Using alternative inclusion and exclusion criteria for selecting participants in analyses
A13		Choosing between different statistical models
A14		Choosing the estimation method, software package, and computation of SEs
A15		Choosing inference criteria (e.g., Bayes factors, alpha level, sidedness of the test, corrections for multiple testing)
Reporting		
R6	T1	Presenting exploratory analyses as confirmatory (HARKing)

checklist that serves the purpose to assess the degree to which preregistrations truly protect against the biasing effects of researcher DFs. Most of the researcher DFs have been described in previous publications (Bakker et al., 2012; Bakker & Wicherts, 2014b; Chambers, 2013; Francis, 2013; John et al., 2012; Kriegeskorte et al., 2009; Nieuwenhuis et al., 2011; Simmons et al., 2011; Simonsohn et al., 2014a; Steegen et al., 2016; van Aert et al., 2016; Wagenmakers et al., 2011). The researcher DFs on our list are invariably inspired by actual research we have encountered as researchers, replicators, re-analyzers, and critical readers of published and unpublished works. Although we have actual examples of the use of each of the researcher DFs we discuss below, we choose not to identify the researchers, projects, or papers involved; the issues are general and it would not help to focus on individual cases.

We discuss the different researcher DFs categorized under headers that refer to different phases in a study from its early theoretical underpinnings to its final publication (hypothesizing, designing, data collection, analyzing, and reporting), and indicate links between different researcher DFs at the different phases. Each researcher DF will be coded according to its phase. All researcher DFs are listed in Table 5.1. We focus on experiments as the most basic design used to study the (causal) effect(s) of independent variable(s) on the dependent variable(s). This design is the most widely used in psychology and entails a good archetype to discuss the many researcher DFs in a multitude of research designs, including quasi-experimental studies and correlational studies aimed to explain dependent variables on the basis of one or more predictor variable(s).

HYPOTHESIZING PHASE

The degree to which the researcher can find relevant statistically significant results in the data is already partly determined by the specificity of theoretical predictions that the study aims to address. A confirmatory study requires a clearly stated hypothesis to be tested, while more exploratory studies aimed at finding interesting (typically statistically significant) patterns in the data often lack a priori theorizing on (causal) relations. Such exploratory studies are virtually guaranteed to yield support for something interesting (Wagenmakers et al., 2011). Since most results in psychology are presented in the realm of the hypothetico-deductive model (Hubbard, 2015), it is tempting to present exploratory findings incorrectly as having been hypothesized in advance. This practice, to which we return when discussing the reporting phase, is also called Hypothesizing After Results are Known or HARKing (Kerr, 1998). The relevant researcher DF during the theorizing phase, namely *T1: conducting explorative research* pervades many of the researcher DFs that we describe below in the later phases of the study. HARKing yields statistical evidence for found patterns that is often much weaker than it appears to be. The reason is that the evidence should be seen in the context of the size and breadth of the explorations, allowing for appropriate corrections for multiple testing. Unfortunately, data explorations are often presented without such necessary corrections. T1 could be dealt with by specifying the independent variable and the dependent variable of interest before running the study, preferably in a preregistration of the study (Wagenmakers et al., 2012). Note that even a preregistered (confirmatory) study can include some exploratory analyses, which is unproblematic as long as these explorations are clearly distinguished from the confirmatory analyses.

However, even if there is a vague prior notion about the relation between the independent and dependent variables, hypotheses that fail to indicate the direction of an effect (or relation) enable later flexibility in the analysis and interpretation of results (Schaller, 2016). If the hypothesis is merely “X is related to Y” or “X affects Y”, the researcher can later analyze the data in two alternative ways; one way to obtain a positive effect and another way to obtain a negative effect of X on Y, which entails a strategy that is a special case of HARKing. The researcher DF is *T2: studying a vague hypothesis that fails to specify the direction of the effect.* Note that specifying the direction of the hypothesized effect is relevant for the decision to use a one- or two-tailed test. One-tailed tests can only be used to reject the null hypothesis when the a priori hypothesis was directional and the result was in the predicted direction. Testing hypotheses requires specificity and precision regardless of whether one uses one- or two-tailed tests. Consequently, a preregistered hypothesis needs to specify the direction of the effect or relation. Because of the need for proper power analyses (discussed below under D6), it is also important to have a prior sense of the size of the effect or strength of the relation.

DESIGN PHASE

Although most researcher DFs discussed in the literature pertain to the analysis of the data, both the theoretical predictions and the design of an experiment (or other types of studies) already allow the researcher to create options for flexible analyses in later phases of the study. A psychological experiment can be set up to have a certain degree of redundancy in the design that creates maneuverability in the collection of data, analysis of data, and reporting of results. This redundancy applies to both independent variables and dependent variables.

Independent variable(s)

We distinguish here between manipulated and non-manipulated independent variables. Manipulated independent variables are those manipulated in the design of the experiment, and typically involve randomization. In contrast, non-manipulated independent variables are based on measures of behavior or individual differences that (could) pertain to the research question. These independent variables are the main focus in correlational studies, but are also widely used in studying (moderation of) experimental effects. Moreover, additional measures taken after the manipulation or in correlational studies could later be used as mediators in explaining variance in the dependent variable, but an underspecified preregistration enables researchers to use these variables as primary dependent variables as well (see our discussion of D4).

Experiments can involve multiple manipulated independent variables (i.e., experimental factors), that are often crossed and that researchers can select or discard in later analyses based on particular (preferred) outcomes. Dropping of experimental conditions has been found to be quite common in a survey among psychological researchers (John et al., 2012) and in a study that considered psychological studies from a register (Franco, Malhotra, & Simonovits, 2016). Specifically, a researcher can discard a factor in a multifactorial experiment by pooling the data over the levels of that factor, or the researcher can select certain levels of a discarded factor. For instance, in a two-by-two factorial design studying the effects of both ostracism (including vs. excluding someone socially) and group composition (being in- or excluded by either a social in-group or a social out-group) on participants' mood, the researcher could ignore group composition either by pooling in- and outgroup levels, or by selecting one of the levels of group composition (say, the in-group) in the later analyses. Moreover, a given experimental factor involving more than two levels can later be analyzed in different ways. For instance, if an experimental factor has three conditions (say, 0, 1, and 2), the researcher can focus on all three levels, but also select two out of the three in the later analyses. Or the researcher can combine conditions 0 and 1 to compare it with condition 2, etc. In this way, this simple three level factor already yields seven different operationalizations for the analysis, from which the researcher can later choose one(s) that yielded the "best" result. So the design of manipulated independent variables offers the following researcher DF*: D1 creating multiple manipulated independent variables and conditions.* Like all researcher DFs, this researcher DF becomes more relevant as the number of scoring options increases, like with complex mixed designs involving multiple between-subject and within-subject factors featuring multiple levels. Consequently, preregistrations of such studies should specifically and precisely delineate how independent variables are later used in testing the focal hypotheses.

Non-manipulated independent variables based on the observed characteristics of participants, are also measured in most research designs. These non-manipulated independent variables such as personality characteristics, IQ, age, gender, ethnicity, political preference, etc. offer great flexibility in the later analyses of the data; one can use them as main predictor, but also as moderators to study potential interactions with manipulated factors, or as control variables or covariates as in ANCOVA. For instance, measured age can assume any of these roles in later analyses: e.g., for studying age differences, for testing whether age moderates the effects of any of the other independent variable(s), or as a control variable to explain some of the within-condition variation in the dependent variable. Moreover, measures taken after the manipulations can be used in later mediation analyses to explain variance in the dependent variable. This entails *D2: Measuring additional variables that can be selected later as covariates, independent variables, mediators, or moderators*. Obviously, adding more measures offers multiple ways to find interesting patterns in later stages of the study. Just as manipulated independent variables can often be operationalized in different ways, many non-manipulated independent variables, once selected, offer flexibility in how they will be used in the analyses. Participants can be assigned to different levels of those independent variables on the basis of flexible thresholds or category assignments (Steegen et al., 2016). For instance, age can be used to create two age groups (young and old), or three age groups (young, middle-aged, and old) on the basis of many different age-based category assignments. However, age can also be used as a continuous factor, covariate or moderator in later analyses. Similar flexibility applies to designs that involve demographic variables (e.g., income, SES, educational level, ethnicity, mother tongue, relationship status) or psychological individual differences (e.g., IQ, extraversion, diagnostic criteria, etc.) and is discussed below in the context of the analyses.

In sum, a research design that is littered with research DFs related to independent variables is complex and offers room for selecting and operationalizing these variables in multiple ways. An ideal preregistration, then, specifically and precisely specifies which manipulated independent variables and non-manipulated independent variables will be used in the analyses and also indicates how both types of variables are to be operationalized, and that no other variables are to be used in the confirmatory analyses. We again emphasize that these specifications are only necessary for the confirmatory analyses; a potential exploratory analyses section of a paper is not at all problematic, as long as the preregistration and the paper clearly distinguish between these very different types of analyses.

Dependent variable(s)

The measurement of human behavior is often complex and is seldom done in a single predefined manner. A design prone to bias due to researcher DFs offers multiple dependent measures of the same construct. This enables the researcher to choose among different outcome measures the one(s) that offer(s) statistical significance. The relevant researcher DF in the design phase is *D3: measuring the same dependent variable in several alternative ways.* For instance, anxiety can be measured with various self-report scales, or with physiological measures (e.g., galvanic skin response, heart rate variability).

Another particularly flexible design allows the researcher to choose among several dependent variables that concern *different* constructs in cases where the originally targeted primary outcome failed to show statistically significant effects. Among the research practices studied by John et al., this practice of not reporting all dependent measures showed quite high prevalence estimates (John et al., 2012). Additionally, direct evidence indicates that psychological researchers often choose among different outcome measures (Franco et al., 2016; LeBel, Borsboom, Giner-Sorolla, Hasselman, Peters, Ratliff, & Smith, 2013). In the medical literature on randomized clinical trials, this researcher DF is often called outcome switching and the bias it introduces is called outcome reporting bias (Chan, Hrobjartsson, Haahr, Gotzsche, & Altman, 2004; Kirkham et al., 2010; Weston et al., 2016). For instance, outcomes that were initially designated as secondary outcome variables appeared as primary outcome variables in the published article. Or a variable that was originally viewed as a potential mediator of an effect might replace the original main outcome variable if the latter failed to show an effect. Here we denote this researcher DF by *D4: measuring additional constructs that could potentially act as primary outcomes.*

Thus in the design of studies, the researcher can already create many researcher DFs that allow for opportunistic use in later phases of the research process, relating to using multiple measures of the same construct (D3), and creating opportunities to find additional effects by adding measures of additional constructs besides the one(s) that were the original focus of interest (D4). D4 allows HARKing (Kerr, 1998), whereas D3 is aimed at the same targeted construct and related to how the primary outcome will be used in later analyses. It is clear that the ideal preregistration should specify which dependent variable(s) will be used in testing particular hypotheses. However, as we discuss below, even specifying the measure (say, the Rosenberg Self-Esteem Scale) that is to be used as primary outcome is *not* specific and precise enough to avoid *p*-hacking during analyses, because often the scores on such measures can be computed in different ad hoc ways.

Excluding participants

Adding numerous measures besides the main independent and dependent variables to the design offers yet another researcher DF: background variables (e.g., age, gender, ethnicity) or other individual differences can be used to later discard participants in an ad hoc manner from the analysis. For instance, a researcher might decide to exclude older-aged participants for some reason that might actually not be independent of the effect of the exclusion on the final analysis. Such exclusion of cases on the basis of measured variables often comes across as ad hoc because if that decision rule had been a priori, these older-aged participants should not have completed the study in the first place.

Other types of measures that can be used to discard participants include awareness checks, as often used in priming research (e.g., the funnel debriefing; Bargh & Chartrand, 2000), checks for alertness in responding like the blue dot task (Oppenheimer, Meyvis, & Davidenko, 2009), or even the simple question like "did you participate seriously in this study?". Decision rules on how to deal with these questions need to be pre-specified to avoid them becoming a researcher DF. Similarly, manipulation checks (i.e., measures of the independent variable) can also be implemented in the design, offering a way not only to assess the strength of the manipulation, but also to discard particular participants from the analyses for not showing any desired response to the manipulation. These decisions in the data selection offer great flexibility in choosing whom to include in the analysis. *D5: measuring additional variables that enable later exclusion of participants from the analyses (e.g., awareness and manipulation checks).* Therefore, an ideal preregistration specifically and precisely describes which types of participants will be excluded from the analyses, and also explicates that the stated rules of exclusion will be the only ones that will be used to discard participants (it should be exhaustive). The reason is that only stating a particular exclusion rule in the preregistration does not preclude the possibility to also exclude participants on other ad hoc grounds.

Power and sampling plan

Despite the core importance of statistical power in NHST, most studies using NHST fail to report a formal power analysis (Bakker et al., 2012; Cohen, 1990; Sedlmeier & Gigerenzer, 1989). This is problematic because researchers' intuitions about power are typically overly optimistic (Bakker, Hartgerink, Wicherts, & van der Maas, 2016) and studies in psychology are often underpowered. More importantly, underpowered studies are themselves more susceptible to bias (Bakker et al., 2012), particularly in combination with the use of many of the other researcher DFs that we describe here. The reason is that the sampling variability is larger and many decisions made in analyzing the data will have proportionately larger effects when sample sizes are smaller. In other words, using researcher DFs to obtain statistically significant results is typically more effective with smaller samples. Low power can create bias and hence D6: *Failing to conduct a well-founded power analysis* is a researcher DF in designing studies.

A rigorous preregistration not only provides a good rationale for the sample size in the form of a power analysis, but also should describe the complete sampling plan, i.e. the targeted sample size, when the data collection starts and ends, and how the participants are to be sampled. The sampling plan should specify the population from which sampling occurs, the procedure of sampling, and the end point of data collection. The sampling plan should also specify when additional participants are to be sampled in cases where the targeted sample size is not met (e.g., due to drop-out or data exclusions). The sampling plan should be specific and precise to disallow researchers to conduct intermediate tests during data collection. If not, a researcher can decide to collect more data after witnessing a non-significant result or to cease data collection earlier than planned if the result is already significant (see also C4), both of which affect the Type I error rate. The sampling plan should also preclude the researcher to conduct a particular study multiple times, and only present the “best” study (i.e., the one with the most desirable results). The use of multiple small studies instead of a larger one is an effective (yet problematic) strategy to find at least one statistically significant result (Bakker et al., 2012) and small underpowered studies can also be pooled by means of a meta-analysis in an ad hoc manner to obtain a statistically significant result (Ueno, Fastrich, & Murayama, 2016). Hence, we call the following researcher DF, *D7: Failing to specify the sampling plan and allowing for running (multiple) small studies.*

DATA COLLECTION PHASE

During the collection of experimental data, it is possible to act in certain ways that enhance the probability of finding a statistically significant result. Most of the issues are methodological, although some are statistical and bear on issues of multiple testing and sequential analyses. In our discussion, we assume that the design itself is internally valid and that the measures are construct valid, in the sense that the experiment does not involve any confounds or artifacts and uses appropriate measures. This, of course, does not always mean that the actual study does not suffer from threats to internal validity or construct validity.

Non-random assignment

Although methodological textbooks are clear on the benefits of random assignment, the randomization techniques used to assign participants to conditions are often not specified in research articles. Using non-random assignment could greatly affect differences between conditions in personal characteristics or other factors that could affect the outcome. For instance, an experimenter might (purposively or not) only run treatment participants in the evening, thereby creating a potential confound, or the assignment could be based on observable personal characteristics that might bear on the outcome measure (e.g., a particularly slow moving participant is assigned to the condition that aligns with slowness). In other words, the randomization technique should be specifically and precisely stipulated in advance and followed throughout the experiment, thereby avoiding *C1: the failure to randomly assign participants to conditions.*

Incomplete blinding

It is widely recommended to employ double-blinding techniques to avoid demand characteristics and placebo effects on part of participants as well as experimenter expectancy effects during data collection (Rosenthal, 1966). Participants are blinded if the design prevents them from knowing to which condition they have been assigned or from knowing the research hypotheses. Experimenters are blinded if they do not know to which condition a participant is allocated at any time. There are several ways in which both types of blinding can be unsuccessful, potentially leading experimenters to treat participants (unwillingly) differently across conditions, or participants to act in ways that yield invalid support for the research hypothesis. Hence, *C2: Insufficient blinding of experimenters and/or participants* could potentially introduce bias. For instance, experimenters could use non-naïve participants (e.g., a fellow student) or (in)avertedly convey information about what is expected from participants in a given condition. The pre-registration study should specifically and precisely describe the procedure of how participants and experimenter(s) are blinded, if applicable.

Discarding, correcting, and coding data

If experimenters are involved in coding or other ways of data handling, incomplete blinding concerning condition assignment or hypotheses could introduce bias. Working with participants is a social process in which experimenters have information about participants or their behavior that might enable them to predict scores on the dependent variable for individual participants. For instance, an experimenter may witness a slowly working student in a condition that is expected to yield quick responses and might decide to discard that participant for not participating seriously even though there is no clear experimental protocol that dictates such a decision. This creates biases during the data collection, and such biases are particularly problematic in experiments involving coding of behavior in a non-blinded manner. Similarly, missing values or incorrectly filled out answers on a questionnaire or test could also be corrected or filled out during data collection by someone who is not blind to condition (or the hypotheses) and hence might make biased decisions. For instance, the experimenter could decide to correct or fill in the answer of a participant who indicated the incorrect gender or no gender on a questionnaire. Although making such corrections or imputations deliberately might go beyond questionable and so might entail falsification (a violation of research integrity), doing this without awareness in a poorly structured research setting might nonetheless cause considerable bias. A specific, precise, and exhaustive research protocol can help avoid this researcher DF. *C3: correcting, coding, or discarding data during data collection in a non-blinded manner.*

Intermediate significance testing

The decision whether or not to continue with data collection could be dependent on intermediate analyses of the data. This is reflected by the common practice to continue data collection after witnessing a statistically nonsignificant result or by quitting data collection earlier than planned after witnessing a statistically significant result (John et al., 2012). It is well known that this type of sequential testing is problematic without any formal correction for multiple testing (Wagenmakers, 2007) and increases Type 1 error rates. *C4: determining the data collection stopping rule on the basis of desired results or intermediate significance testing.* A specific and precise a priori sampling plan could ameliorate this, and so this researcher DF is related to D7 described above.

ANALYSIS PHASE

In the analysis phase, the researcher directly witnesses the effects of choices on the statistical outcome. It is surprising that blinding to conditions and hypotheses of experimenters, coders, and observers is considered to be crucial during data collection, while in practice, the analyses are typically conducted by a person who is not only aware of the hypotheses, but also benefits directly from corroborating them (commonly by means of a significance test). Together with the many researcher DFs during the analyses, these factors do not entail the most optimal mix for objective and unbiased results.

Data cleaning and processing

Before running the focal analyses, experimental data often need to be cleaned and prepared for analysis. Data cleaning involves many choices related to missingness, outliers, or violations of the distributional assumptions. Because of potential drop-out, data collection problems, or a lack of full responses for other reasons (e.g., participants' inattention or refusal to answer some questions), some data might be missing entirely for participants or for some or many of the variables of interest. Missing data can be dealt with by listwise deletion, pairwise deletion, multiple imputation, full information methods, and other methods (Schafer & Graham, 2002). This choice creates a researcher DF, namely *A1: choosing between different options of dealing with incomplete or missing data on ad hoc grounds.*

Neuroimaging techniques (e.g., signals from fMRI, EEG, MEG) and other data-intense measurement procedures require extensive pre-processing steps that entail considerable maneuverability in the analysis of the data (Kriegeskorte, Lindquist, Nichols, Poldrack, & Vul, 2010; Kriegeskorte et al., 2009; Poldrack et al., 2016). For instance, with neuroimaging data, decisions related to regions of interest, dealing with head motions, corrections for slice timing, spatial smoothing, and spatial normalization can create a large number of different ways to analyze the data (Poldrack et al., 2016). The processing of such data can be done based on whether they provide preferred results, which offers *A2: Specifying pre-processing of data (e.g., cleaning, normalization, smoothing, motion correction) in an ad hoc manner.*

Tests have assumptions related to how the data are distributed. The typical assumption in the *F*-family of parametric tests is that the data are independently normally distributed and that variances of different groups are homogenous. There are various ways to deal with violated assumptions of such statistical tests: one could use non-parametric analyses, transform the data in various ways to approach normality or simply ignore the violations. Moreover, violations of variance homogeneity in ANOVAs or *t*-tests, non-normality, non-linearity in linear models, or heteroscedascity in regression could be dealt with in several alternative ways (Wilcox, 2012). When done in data-driven manner, this creates: *A3: Deciding on how to deal with violations of statistical assumptions in an ad hoc manner.*

Dealing with outliers is a particularly vexing issue that warrants specifically, precisely, and exhaustively described protocols in a preregistration. Outliers can be operationalized and detected in various ways (Bakker & Wicherts, 2014b; Barnett & Lewis, 1994; Wilcox, 2012) and they can be deleted or kept on the basis of many alternative criteria that could be chosen based on whether they lead to significance. Alternatively, the researcher can choose to conduct analyses that are less sensitive to outliers, like non-parametric or robust analyses. This creates *A4: deciding on how to deal with outliers in an ad hoc manner.*

Dependent variable(s)

Statistical analyses of experimental data boil down to predicting scores on the outcome measure chosen in the analysis on the basis of predictors (typically factors, but also covariates, mediators, and/or interaction terms). While running the analysis, the researcher can choose between different measures or operationalizations of the same construct implemented in the design of the study in an effort to find the measure that shows the preferred or best results. This practice, which is paired with the use of various measures in the design (D3), concerns the following researcher DF: *A5 selecting the dependent variable out of several alternative measures of the same construct.*

The dependent variable, once selected, can often be scored or operationalized in various ways. There also exist degrees of freedom even if there is only one overall measure or scale. Discarding, weighting, selecting, or redefining scoring rules of individual items can offer flexibility in analyses even if the items are based on a commonly used scale. For example, items of a scale that are originally measured on a five-point Likert scale can be dichotomized, or some items might be discarded from the scale score for showing low or negative item-rest correlations. Moreover, the scale score can be based on an unweighted sum of item scores or on weighting of items based on an item response model, or by running a principal components analysis and choosing among alternative ways to estimate the factor scores. So even for existing scales, flexibility exists in operationalizing the scores in the analyses. The use of response time data involving responses to many stimuli also involves many choices in dealing with slow response times, and how to summarize the major outcome variable. The researcher DF is *A6: trying out different ways to score the chosen primary dependent variable.*

Finally, researchers can choose to measure additional constructs next to the one(s) originally targeted as the main dependent variable (or primary outcome) in the design (see D4). During the analyses this creates *A7: selecting another construct as the primary outcome*.

Independent variables

If we consider ANOVA as a regression model, the use of independent variables means selecting among numerous predictors and/or interaction terms to predict the outcome, and hence different regression models. Without specific preregistration, a researcher often has numerous options to choose between different regression models. The researcher can also typically operationalize the non-manipulated and manipulated in various ways, particularly in flexible designs. During the analysis, the researcher can employ *A8: select independent variables out of a set of manipulated independent variables (paired with D1).* Similarly, even for a given manipulated variable, the researcher can often choose to discard or combine different levels of factors, which creates *A9: operationalizing the manipulated independent variables in different ways (e.g., by discarding or combining levels of factors; paired with D1).*

Furthermore, during the analyses, the researcher can make opportunistic use of a host of additional non-manipulated measures (D2), as well as possible mediator variables measured during the study, thereby creating *A10: choosing* *to include different measured variables as covariates, independent variables, mediators, or moderators* in the analysis. The number of different analytic options considered for finding some statistically significant result (mediation, moderation, main effect) on the basis of measured variables can be quite large. For instance, adding big five personality measures to a simple one-way experimental design enables the researcher to seek for the moderation of effects by all of these personality traits. However, these big five traits can also be used as covariates, or simply as independent variables to help explain variance in the outcome measure(s). More degrees of freedom are added if the researcher does not specifically and precisely describe in advance how these measured variables are to be used and scored in the analysis (Steegen et al., 2016). For example, a measure of extraversion could be used as a linear predictor based on unweighted sum of individual item scores or some estimate of the factor score reflecting the underlying construct. However, the researcher can also compare participants with some (arbitrarily chosen) high or low score on the scale used to measure extraversion (and even there, the researcher could discard some items because they showed low item-rest correlations). This creates *A11: operationalizing non-manipulated independent variables in different ways.*

An exceptionally flexible analysis involves many different regression models based on a host of different combinations of predictors (main effects, interactions, control variables or covariates), and alternative ways to operationalize these predictors, leading to a very large number of regressions (Sala I Martin, 1997) in some designs. For instance, a researcher might add age as a moderator during the analysis and check whether different ways to categorize age groups yields some interesting results. Running so many regressions creates a massive multiple testing problem that can be solved in statistical ways or with a sufficiently detailed preregistration.

Selection criteria

One can also change the analysis by altering the sample size on the basis of different criteria to (de)select participants. This yields *A12: Use of alternative inclusion and exclusion criteria in selecting participants for use in the analysis*. This researcher DF is paired with D5, i.e. the design choice to include many additional variables related to manipulation checks or awareness checks or any other personal characteristics that can be used as selection criteria. There are many bases to select or deselect participants for the analysis, including performance (e.g., many alternative levels of the percentage of items answered correctly on some task that measures the manipulation), awareness questions, or any personal characteristics. A specific, precise, and exhaustive plan to not include particular participants in the final data analyses not only avoids this researcher DF, but could also spare resources by simply not collecting any (additional) data for participants who fail to meet the inclusion criteria. For instance, if a linguistics researcher is not interested in participants who are not native speakers of the language of interest, he or she would be better off not running these participants at all, instead of excluding their data only at the analysis phase.

Statistical model, estimation, and inference

Even for relatively straightforward experiments, many different statistical models can be used to analyze experimental data, including linear regression, ANOVA, MANOVA, or robust or non-parametric analyses. Hence, an obvious researcher DF is *A13: choice of the statistical model*. However, choosing the statistical model (say, a regression with three predetermined predictors), often does not preclude additional statistical choices. Specifically, statistical models need to be estimated and this can frequently be done in several ways. Even with a given estimation method, the researcher can choose between different corrections to the standard errors (SEs) of parameters. For instance, one could choose for robust standard errors instead of the standard SEs. Moreover, different statistical software packages (e.g., SPSS, R, SAS) often implement the same estimation techniques and correction methods in slightly different ways, leading to diverging results. These alternative estimation methods, software packages, and correction methods might lead to different outcomes and hence entail a researcher DF. *A14: the choice for estimation method, software package, and computation of SEs.* To wit, even a standard ANOVA requires a choice between different types of sum of squares, three of which are available in SPSS (this choice is typically not described in articles). This problem is particularly vexing for more advanced analyses, that can be estimated with Maximum Likelihood (ML), Ordinary Least Squares, Weighted Least Squares, Mean and Variance Adjusted Weighted Least Squares, Partial Least Squares, or Restricted ML, with or without robust standard errors (to name just a few options).

Finally, without a specific and precise registration, a researcher can choose inference criteria (e.g., Bayes factors, alpha level, sidedness of the test, corrections for multiple testing) in different ways, and on the basis of analytic outcomes. Thus *A15: choosing inference criteria (e.g., Bayes factors, alpha level, sidedness of the test, corrections for multiple testing)* is another researcher DF. For instance, a researcher can choose to use a one-sided test if this is the only way to obtain significance, or employ more lenient corrections for multiple testing if the need arises. Preregistrations should explicate these criteria.

REPORTING PHASE

In the reporting of results, the researcher is faced with the freedom to report details of the a priori hypotheses, design, data collection, and analysis of the study. Here, the potential exploitations of the many researcher DFs discussed above can or cannot be reported, which renders the reporting of such details crucial (Wigboldus & Dotsch, 2016). For instance, the researcher could report only a subset of many analyses that showed the researcher's most desirable results. The comprehensive reporting of the study design and results is necessary for both reproducibility (reanalyzing the study using the same data) and replicability (rerunning the study as similar as possible in a new sample) (Asendorpf et al., 2013). It is obvious that the many researcher DFs can be hidden for readers (or critical reviewers) by failing to report some independent variables, some dependent variables, missing data, data exclusions, or other relevant choices made during the analyses. Reproducibility requires a verification of the steps taken from the data collection to the final report, including choices made during the collection and analysis of the data, such as pre-processing of the data, the statistical model, the estimation technique, software package, and computational details, data exclusions, dealings with missing or incomplete data, violated distributional assumptions, and outliers. This offers the following researcher DF in the reporting phase: *R1: failing to assure reproducibility (verifying the data collection and data analysis).* The preferred way to assure reproducibility is to share data and analytic details (computer syntaxes/code) in or alongside the paper (Nosek et al., 2015).

The exploitation of researcher DFs creates bias, which might lower replicability of earlier results in novel samples (Open Science Collaboration, 2015). To allow later (direct) replications of a study, it is crucial that the report (or its supplements) include sufficient details on the data collection, including procedures and all materials used (stimuli, instructions, manipulations, and measures). Nowadays, such information can be shared via online repositories or platforms such as the Open Science Framework. Failing to do this impedes replications, and so we consider this another researcher DF during the reporting of studies, namely *R2: failing to enable replication (re-running of the study).* Although both reproducibility and enabling replication are considered matters of reporting here, a preregistration of the study could already specifically and precisely indicate what information is going to shared and in what manner.

Furthermore, for preregistered studies, there exists an additional researcher DF related to reporting of results. Specifically, the researcher(s) could *R3: fail to mention, misrepresent, or misidentify the study preregistration.* Studies of preregistrations of randomized clinical trials highlight that preregistrations in the medical literature are often not followed in the final report (Chan, Hrobjartsson, et al., 2004). This problem can be avoided by having reviewers compare the preregistration to the (submitted) research article.

Moreover, researchers could fail to present relevant unpublished work in their final publication. This creates *R4: failing to report so-called "failed studies" that were originally deemed relevant to the research question*. Note that failed studies are often those that showed no statistically significant results, which is a main reason for authors for not publishing the results (Cooper, DeNeve, & Charlton, 1997). However, if the study was seen in advance as valid and methodologically rigorous, the study cannot be considered "failed" and should be considered as adding relevant evidence. This is the idea underlying the article format of registered reports, in which the rationale and methods of a study are reviewed and the final study is accepted for publication regardless of the (statistical significance of the) final result (Chambers, 2013; de Groot, 1956/2014; Simons, Holcombe, & Spellman, 2014).

There are two more researcher DFs in the reporting of studies that bear on the results or the rationale for the study, respectively. First, researcher(s) could *R5: misreport results and p-values* (Bakker & Wicherts, 2011), for instance by presenting a statistically nonsignificant result as being significant. This practice and similar practices of misreporting of results (e.g., incorrectly stating a lack of moderation by demographic variables) are quite common (John et al., 2012; Nuijten et al., 2016). Second, researchers can choose to *R6: hypothesize after the results are known (HARKing).* They can falsely present results of data explorations as though they were confirmatory tests of hypotheses that were stipulated in advance (Wagenmakers et al., 2011), which is related to lack of clear hypotheses (T1) and appears to be quite commonly practiced by psychologists (John et al., 2012). Both types of misreporting lower trust in reported findings and potentially also the replicability of results in later research.

DISCUSSION

We created a list of 34 researcher DFs, but our list is in no way exhaustive for the many choices that need be made during the different phases of a psychological experiment. Some of the researcher DFs are clearly related to others, but we nonetheless considered it valuable to list them separately according to the phase of the study. One can envision many other ways to create bias in studies, including poorly designed experiments with confounding factors, biased samples, invalid measurements, erroneous analyses, inappropriate scales, data dependencies that inflate significance levels, etc. Moreover, some of the researcher DFs on our list do not apply to other statistical frameworks, and our list does not include the specific

DF associated with those frameworks (e.g., specifying priors in Bayesian analyses). Here we focused on the researcher DFs that are often relevant even for well-designed and rigorously conducted experiments and other types of psychological studies that use NHST to test their hypotheses of interest.

We sympathize with Gelman and Loken's (2014) argument that the term questionable research practices in relation to researcher's use of researcher DFs is not always necessary, because the majority of the researcher DFs we describe involve choices that are arbitrary: researchers just need to decide between these different options and *could but not necessarily will* use these researcher DFs in an opportunistic manner. What matters is that the data could be collected and analyzed in different ways and that the final analyses reported in the research article could have been chosen differently if the results (based on these different choices and bearing on statistical significance) had come out differently. The issue, then, is not that all researchers try to obtain desirable results by exploiting researcher DFs but rather that the researcher DFs have strong potential to create bias. Such potential for bias is particularly severe for experiments that study subtle effects with relatively small samples. Hence, we need an appropriate way to deal with researcher DFs; one way to assess the relevance of choices is to report all potentially relevant analyses either as a traditional sensitivity analyses or as a multiverse analysis (Steegen et al., 2016). Another solution is that the data are available for independent reanalysis after publication, although this is not always possible due to low sharing rates (Wicherts et al., 2011). However, preventing bias is better than treating it after it has occurred. Thus, the preferred way to counter bias due to researcher DFs is to preregister the study in a way that no longer allows researchers to exploit them.

The ideal preregistration of a study provides a *specific*, *precise*, and *exhaustive* story of the planned research, that is, it describes all steps, with only one interpretation, and excludes other possible steps. Our list can be used in research methods education, as a checklist to assess the quality of preregistrations, and to determine the potential for bias due to (arbitrary) choices in unregistered studies. We conducted a study (see Chapter 6) focusing on the quality of a random sample of actual pre-registrations on the Open Science Framework in which we used a scoring protocol based on our checklist to assess the degree to which these pre-registrations avoid any potential *p*-hacking. The protocol assesses the preregistration's specificity, precision, and completeness at the level of each researcher DF; a score of 0 is assigned if the DF is not limited, whereas 1 and 2 are assigned if the description is partly or fully specific and precise, respectively. A score of 3 is assigned if it is also exhaustive, i.e. if it excludes other steps. By applying the protocol, authors can also score their own preregistration, enabling them to improve their preregistration, and reviewers of registered reports and registered studies can use the protocol as well. Both authors and reviewers can thus use the protocol to limit potential *p*-hacking in planned studies.

We suggest a few avenues for future research. First, while most of the researcher DFs in our list are relevant to other statistical frameworks as well, the list should be adapted for studies planning to use confidence intervals and certain precision of effect size estimates (Cumming, 2012, 2014; Maxwell et al., 2015), or Bayesian analyses (Kruschke, 2015). Second, where we focused on preregistrations and assessing their quality, it is likewise urgent to develop and assess protocols for using 'open materials', 'open data', and 'open workflows' (Nosek et al., 2012). These transparent practices have many benefits and are currently gaining traction (e.g., Eich, 2014; Kidwell et al., 2016), but are often insufficiently detailed, documented or structured to allow other researchers to reproduce and replicate results (e.g., reuse of open data requires solid documentation and meta-data; Wicherts, 2017). While we believe all these open practices strengthen research, a lot can still be gained by creating protocols that provide specific, precise, and exhaustive descriptions of materials, data, and workflow.

CHAPTER 6

Restriction of opportunistic use of researcher degrees of freedom in pre-registrations on the Open Science Framework

After all authors have approved this manuscript, it will be submitted as Veldkamp, C. L. S., Bakker, M., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Soderberg, C. K., Mellor, D., Nosek, B. A., & Wicherts, J. M. Restriction of opportunistic use of researcher degrees of freedom in pre-registrations on the Open Science Framework.

ABSTRACT

In psychological science, researchers face many, often seemingly arbitrary choices in formulating the hypotheses of their study, designing their experiment, collecting their data, analyzing their data, and reporting their results. Opportunistic use of these 'researcher degrees of freedom' aimed at obtaining statistical significance however increases the likelihood of obtaining false positive results and overestimating effect sizes, and lowers the reproducibility and replicability of published results. Here, we compared the effectiveness of two types of pre-registration (i.e. stipulating all planned aspects of a study in advance) as a solution to restrict opportunistic use of researcher degrees of freedom (or *p*-hacking). Both types (Standard Pre-Data Collection Registrations and Prereg Challenge Registrations) are currently available on the Open Science Framework and differ in the extent to which they provide authors with detailed instructions and requirements on how to write the pre-registration. Results of comparing random samples of 53 pre-registrations from each type indicate that neither of the two types of pre-registrations sufficiently restricted opportunistic use of researcher degrees of freedom. However, on average, Prereg Challenge Registrations, which follow a format providing more detailed instructions and requirements, worked better than Standard Pre-Data Collection Registrations, which follow a basic format hardly providing any instructions. We discuss the benefits and limitations of pre-registration, and provide suggestions on how to improve pre-registration formats.

Attempts to replicate original research findings seem uncommon in psychology (Asendorpf et al., 2013; Mahoney, 1985; Makel, Plucker, & Hegarty, 2012; but also see Neuliep & Crandall, 1993a; Neuliep & Crandall, 1993b), but when replications are conducted, they produce weaker evidence in the large majority of cases, or even no evidence for the original findings in many cases (Asendorpf et al., 2013; Chang & Li, 2015; Ioannidis, 2005a, 2007; Marsman et al., 2017; Mobley, Linder, Braeuer, Ellis, & Zwelling, 2013; Open Science Collaboration, 2015). In recent years, many psychologists have expressed their concerns about the size and the gravity of these problems, even referring to it as a replication crisis (Nosek & Lakens, 2015; Open Science Collaboration, 2015; Pashler & Harris, 2012; Spellman, 2015).

The difficulty of replicating published results is believed to be due to several interrelated problems. First, statistically significant results have a higher probability of getting published (Dwan et al., 2008; Fanelli, 2010b, 2012; Sterling, 1959; Sterling et al., 1995), a problem commonly known as publication bias. As a consequence of this selection for significance, many published effect sizes over-estimate population effect size (Bakker et al., 2012; Ioannidis, 2008; Simonsohn et al., 2014a; van Aert et al., 2016), and many published statistically significant results may constitute false-positive findings (Ioannidis, 2005b). Second, analyses of (psychological) data often involve many (often arbitrary) choices that have to be made during data analysis that researchers could use opportunistically when confronted with an (undesired) non-significant result (DeCoster et al., 2015; Ioannidis, 2005b; Nuzzo, 2015; Sijtsma, Veldkamp, & Wicherts, 2015; Simmons et al., 2011; Wicherts et al., 2016). This use may result in statistically significant findings after all, (denoted *p*-hacking) and may also result in overestimated effect sizes and dissemination of false positive results. In psychology, the opportunistic use of the so-called ‘Researcher Degrees of Freedom’ (Simmons et al., 2011; Wicherts et al., 2016) constitutes a large problem because its occurrence is estimated to be high (Agnoli et al., 2017; Fanelli, 2009; Fiedler & Schwarz, 2015; Franco et al., 2016; John et al., 2012; LeBel, Borsboom, Giner-Sorolla, Hasselman, Peters, Ratliff, & Tucker Smith, 2013; O’Boyle, Banks, & Gonzalez-Mulé, 2014) and its effects are particularly strong for underpowered (small sample) studies that are very common in psychology (Bakker et al., 2012; Cohen, 1962; Maxwell, 2004).

We believe that the tendency to use Researcher Degrees of Freedom opportunistically can largely be attributed to human cognitive biases such as confirmation bias, motivated reasoning, hindsight bias, and the inclination to see patterns in any data (Mahoney & DeMonbreun, 1977; Mynatt, Doherty, & Tweney, 1977; Nickerson, 1998), combined with the common practice (if not requirement) to report positive results in research papers (Fanelli, 2010). Just like other humans, scientists do not always recognize their own biases: they can ‘fool themselves’ (Nuzzo, 2015). We also believe that many scientists do not perceive their data-analytical choices as opportunistic, but rather as being 'justified by the data'. However, making data-analytical decisions after seeing the data and results of analyses on these data violates the principles of confirmatory null hypothesis significance testing or NHST (Wagenmakers et al., 2012), which is the most widely used statistical framework in psychology (Hubbard & Ryan, 2000; Wetzels et al., 2011) and creates options to present data-driven hypotheses as confirmatory hypotheses (Kerr, 1998). Human biases combined with the many choices in how to design studies, collect data, analyze data, and report results likely inflate the likelihood of both false-positive findings and exaggerated effect sizes. In environments where competition is high (Anderson, Ronning, De Vries, & Martinson, 2007; Fanelli, 2010a; Martinson et al., 2005) and where the pursuit of significant results is strongly incentivized by publications of positive and "novel" results in high impact journals (Asendorpf et al., 2013; Nosek et al., 2012), reporting false-positive results and/or over-estimated effect sizes may be beneficial for researchers in the short run. In the long run, however, researchers might be confronted with the irreproducibility of their results, and the credibility of their field will be negatively affected. Interestingly, despite the risks for bias associated with small samples, high impact journals in psychology report relatively more studies with smaller and hence underpowered samples than do lower-ranked journals (Fraley & Vazire, 2014).

A widely discussed solution to reduce opportunistic use of Researcher Degrees of Freedom or *p*-hacking (and publication bias) in psychology is study pre-registration (Asendorpf et al., 2013; Chambers & Munafo, 2013; de Groot, 1956/2014; Nosek et al., 2015; Nosek & Bar-Anan, 2012; Nosek et al., 2012; Wagenmakers et al., 2012). The idea of pre-registration is to specify all planned aspects of a study in advance, thereby making a clear distinction between 'hypothesis testing' (confirmatory research) and 'hypothesis generating' (explorative research') (Wagenmakers et al., 2012). Pre-registration thus entails stipulating the confirmatory hypotheses, the study design used to test these hypotheses, the data collection plan, the (confirmatory) analysis plan, and what will be reported in the manuscript prior to the execution of the study. It is important to note that pre-registrations do not prohibit exploratory analyses; they merely make a *distinction* between confirmatory and exploratory analyses, and provide the planned steps of the study before actually carrying out the study.

Currently, there are a number of emerging forms of pre-registration in psychology. For example, researchers can register their study on the Open Science Framework (https://osf.io) or on the AsPredicted website (https://aspredicted.org/) and refer to these pre-registrations in the manuscript reporting on the relevant project. One of the leading psychology journals, *Psychological Science*, now actively supports and rewards pre-registration (Eich, 2014). In addition, pre-registration can occur in the form of ‘registered reports’ (Chambers, 2013; Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Jonas & Cesario, 2015) or ‘registered replication reports’ focused on replicating earlier (seminal) findings (Simons et al., 2014). In these pre-registered (replication) reports, the introduction and methods section of a paper are submitted and peer reviewed prior to the execution of the study. When these sections have been approved by the reviewers and editor, the authors receive a so-called ‘in principal acceptance’, meaning that acceptance of the manuscript for publication will be conditional only on the pre-specified methods and analysis plan being followed, and not on the study outcome. This format has already been adopted by more than 35 journals (see https://cos.io/rr/) across different disciplinary fields.

In order for pre-registrations to be efficient in limiting or even eliminating opportunistic use of researcher degrees of freedom, pre-registrations must be *specific*, *precise*, and *exhaustive* (Wicherts et al., 2016, Chapter 5). By specific, we mean that the pre-registration is detailed in its description of all phases of the research process from the design of the study to what will be reported in the manuscript. By precise we mean that each aspect of the research plan is open to one single interpretation only. Finally, by exhaustive we mean that each aspect of the pre-registered research plan explicitly excludes the possibility of deviations from the pre-registered research plan. For example, a specific, precise, and exhaustive description of a sampling plan would state the exact procedure or number of participants to be recruited, describe the protocol of sampling in all its details (i.e., including the exact number of people to be approached, the exact time frame and situation in which participants will be invited), list the inclusion and exclusion criteria for selecting participants or data points, specify how many and how additional participants or data points will be sampled when the pre-set sample size is not reached, and explicitly state no other sampling procedure(s) will be used than those listed. A pre-registered sampling plan that simply states that at least 50 participants will be recruited in each condition leaves room to continue recruiting until intermediate testing yields a significant effect and to exclude participants for ambiguous reasons, and allows for the exclusion of certain participants for whatever reason. Hence, such a poorly specified sampling plan might still create many options to inflate false positive rates in NHST (e.g., Wagenmakers, 2007).

As another example, we explain how a pre-registered description of a dependent variable that consists of ‘we will use the Positive and Negative Affect Schedule (PANAS)’ leaves ample room to select only those items (e.g., the negative affect items) that correlate with the independent variable for the analyses, to remove items from the scale that decrease the correlation with the independent variable, to construct the composite score in a manner that yields the most favorable effects (e.g., high versus low rather than continuous scores), or to examine whether pairwise deletion, list wise deletion, or imputation of missing values, etc. yields the most favorable results. In other words, merely stating the use of a particular scale in a pre-registration fails to limit numerous remaining degrees of freedom in how to collect and analyze the data. In order to provide a specific, precise, exhaustive description of a composite measure (e.g.,. an affect scale) of a dependent variable, it should explain the protocol to administer the items, the scoring of the items, and the procedure to construct the composite score from the items (arithmetic mean, weighted mean, sum, etc.). In addition, it should specify how deviating individual items, incorrect values, and missing values on individual items will be handled, and should explicitly mention that no other procedure(s) will be used for measuring the dependent variable.

Finally, we emphasize that pre-registrations should indicate what will be reported in the manuscript following the study. Research in the field of randomized controlled trials in medical science, where pre-registration has been used for years, suggests that there are often crucial discrepancies between the pre-registration and the article reporting the trial. For example, the primary outcome measures reported in articles have been found to differ from the primary outcomes listed in the pre-registration alarmingly often (Chan & Altman, 2005; Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2004; Goldacre, 2016), and final sample sizes reported in articles have been found to not always match those in the pre-registrations of the same trials (Chan, Hróbjartsson, Jørgensen, Gøtzsche, & Altman, 2008). In order for pre-registrations to be effective, pre-registrations thus also need to be permanently archived and accessible (Chan, 2008), and pre-registrations should be explicitly compared to studies reported in manuscripts during peer review. In registered reports, this is an essential part of the review process (Chambers, 2013).

The currently emerging forms of pre-registration in psychology vary widely in the extent to which they help researchers compose a specific, precise, and exhaustive pre-registration. For example, the website AsPredicted facilitates a form of preregistration that prioritizes ease and time-saving by requiring researchers to briefly answer a total of eight questions about their study. These eight questions ask researchers to state the main research question, describe the key dependent variable(s), state to how many and which conditions participants will be assigned, specify the exact analyses to test the main hypothesis, indicate whether and which secondary analysis will be conducted, stipulate how many observations will be collected, indicate whether there is anything else they would like to pre-register, and indicate whether they have already collected data. The website then generates a private or public PDF (depending on the author's choice), that will be archived by the website, can be downloaded by the authors, and can be verified by reviewers.

The Open Science Framework website features a number of different types of pre-registrations, ranging from formats that hardly offer any instructions to fixed formats instructing authors to provide a high level of detail about many aspects of their study. At the time of the start of the current study, three major types of pre-registrations were available: 'Open-ended Registrations', 'Standard Pre-Data Collection Registrations', and 'Prereg Challenge Registrations'. 'Open-ended Registrations' are the broadest and most general version, where researchers are only asked "to provide a narrative summary of their project", 'Standard Pre-Data Collection Registrations' have the same basic format as 'Open-ended Registrations', but the difference is that here, researchers are also asked to indicate whether they have already collected or looked at the data before composing the pre-registration. Researchers can choose how they (permanently) archive their pre-registration on the Open Science Framework (publicly, privately, or a combination of both, e.g. by keeping it private until publication of the results).

The third type, 'Prereg Challenge Registrations', are of a rather different format. Here, authors are asked to fill out a form containing 26 questions and to provide a high level of detail in answering the questions. The questions pertain to general information about the study (title, authors, research questions and hypotheses), to the sampling plan (whether existing data are used, explanation of existing data, data collection procedure, sample size, sample size rationale, stopping rule), to the variables (manipulated variables, measured variables, indices), to the design plan (study type, blinding, study design, randomization), to the analysis plan (statistical models, transformations, follow-up analyses, inference criteria, data exclusion, missing data, and (optional) exploratory analyses), and to the scripts that will be used (optional). This format was developed for and is currently used in the 'Preregistration Challenge' (or 'Prereg Challenge'), a competition held by the Center of Open Science in order to promote experience and education with pre-registration. To be eligible for one of the 1000 prizes of $1000, participants must submit a fully completed 'Prereg Challenge Registration' form for review to the Center for Open Science) and publish their manuscript in one of the participating Open Access journals.

Finally, pre-registration in the form of 'Registered Reports' (Chambers, 2013; Chambers et al., 2014; Jonas & Cesario, 2015; Nosek & Lakens, 2015) also requires a high level of detail, but the exact requirements depend on the journal. Journals welcoming or requiring Registered Reports, such as Comprehensive Results in Social Psychology, usually provide extensive instructions for submission at the pre-study stage (i.e. the introduction, the method section, and the analysis plan) and the submission at the post-study stage (i.e. the introduction, method section, and analysis plan approved in the pre-study stage now combined with the results and discussion). For example, the Journal of Comprehensive Results in Social Psychology offers a detailed manual (see http://www.tandf.co.uk/journals/authors/rrsp-submission-guidelines.pdf) instructing authors to provide an introduction section “motivating the research question and a full description of the experimental aims and hypotheses”. Next, authors are instructed to write a method section that includes a full description of the sample characteristics (“including criteria for subject inclusion and exclusion, and detailed description of procedures for defining outliers), a description of experimental procedures (“in sufficient detail to allow another researcher to repeat the methodology”), the proposed analysis pipeline (“including all preprocessing steps, and a precise description of all planned analyses, including appropriate correction for multiple comparisons”), and a power analysis (for studies using NHST, where “effect sizes for power analysis should be justified with reference to the existing literature”), or a specification of the predictions of the theory so that a Bayes factor can be calculated (for studies using Bayesian hypothesis testing, indicating “what distribution will be used to represent the predictions of the theory and how its parameters will be specified”).

In sum, pre-registration formats differ greatly in the extent to which they take the author by the hand in writing a pre-registration that is sufficiently specific, precise, and exhaustive. We therefore believe that the extent to which pre-registrations restrict opportunistic use of researcher degrees of freedom may therefore largely depend on the quality of the instructions and requirements imposed by the medium offering the opportunity for pre-registration.

Our own experiences with previous pre-registrations and the pre-registration of the current study have shown that writing a specific, precise and exhaustive pre-registration that truly limits all manoeuvrability with respect to confirmatory hypothesis testing is difficult and requires considerable time and energy. Such a pre-registration is lengthy and detailed, and needs to be extensively reviewed by all authors involved. However, we have also learned that planning every detail of a study ahead is eventually more efficient in that it helps reduce the number of unforeseen issues and errors that can occur in the design of a study, during data collection, or in analyses, and in that it reduces time spent writing the introduction and method sections of manuscripts. Moreover, examining our pre-registration after data collection and data analysis often confronted us with the presence of our own hindsight bias and tendency for motivated reasoning, which for us confirms the value of pre-registration in limiting our human tendency to ‘fool ourselves’ (Nuzzo, 2015).

In the current study, we evaluated to what extent current pre-registrations efficiently restricted opportunistic use of 29 out of the 34 researcher degrees of freedom listed in Chapter 5 (Wicherts et al., 2016) by means of a detailed scoring protocol. Specifically, we evaluated pre-registrations of two of the pre-registration formats described above: Open Science Framework ‘Standard Pre-Data Collection Registrations' and Open Science Framework 'Prereg Challenge Registrations'. As explained above, these types of pre-registrations differ with respect to the amount of instructions provided and the level of detail required. Evaluating these two types of pre-registration could shed light on our three research questions: (i) whether pre-registrations that were written following more detailed and specific requirements (Prereg Challenge Registrations) restricted opportunistic use of researchers degrees of freedom more than pre-registrations that were written following less detailed and less specific requirements (Standard Pre-Data Collection Registrations), (ii) which of the 29 researchers degrees of freedom tended to be covered better by the preregistrations than others, and (iii) how to create pre-registration guidelines that do adequately restrict opportunistic use of researcher degrees of freedom. In our study, we tested one, pre-registered confirmatory hypothesis: that Prereg Challenge Registration would receive on average higher 'restriction scores' (see method section) than Standard Pre-Data Collection Registrations. The complete preregistration of our study of pre-registrations can be found here. Note that although our own pre-registration followed the Prereg Challenge Registration format, it was not entered into the Prereg Challenge competition because two of the authors of the current article are involved in the organization of the Prereg Challenge. We wrote our own pre-registration according to the standards we set for the ideal pre-registration (see our scoring protocol described in the method section) and continued revision until it received full marks from one of the members of the coding team (EC) who was not involved in the creation of the scoring protocol.

METHOD

Researcher Degrees of Freedom Assessed

In order to evaluate the extent to which Standard Pre-Data Collection Registrations (SPR) and Prereg Challenge Registrations (PCR) restricted potential opportunistic use of researcher degrees of freedom (researcher DFs), we employed the list of researcher DFs we presented in Chapter 5 (Wicherts et al., 2016), which aimed to raise awareness of the risk of bias implicit in a lot of research designs in psychology and other fields. The list can be used in research methods education, as a tool to assess potential bias in unregistered studies, and as a checklist to assess the quality of preregistrations. The latter use was implemented in the current study, where we constructed a coding protocol based directly on this list. As explicated in Chapter 5 (Wicherts et al., 2016), the list itself was created in a qualitative manner based on both the existing literature (Bakker et al., 2012; Bakker & Wicherts, 2014b; Chambers, 2013; Francis, 2013; John et al., 2012; Kriegeskorte et al., 2009; Nieuwenhuis et al., 2011; Simmons et al., 2011; Simonsohn et al., 2014a; Steegen et al., 2016; van Aert et al., 2016; Wagenmakers et al., 2011) and additional considerations.

The items on the list are categorized into five phases of the research process: formulating the hypotheses, designing the study, collecting the data, analyzing the data, and reporting. Although the list is by no means exhaustive, it covers a wide range of researcher DFs that can play a role in psychological research. It is important to note that the list was composed with the use of NHST in mind, and with a particular focus on experimental study designs (although it also applies to non-experimental studies). In the current study, we only included 29 out of the 34 researcher DFs from Wicherts et al. (2016) because five of the researcher DFs only concern the actual reporting phase of a study and therefore cannot be assessed based only on a pre-registration. Table 6.1 presents the list of 29 researcher DFs that we included in the current study. It provides the codes used in the original list (Chapter 5, Wicherts et al., 2016), it indicates to which other researcher DFs each researcher DF is related, the description of the researcher DF identical to the original list, and short labels describing the researcher DFs that we use later when describing our results.

Table 6.1 *Degrees of freedom in formulating the hypotheses, designing the study, collecting the data, analyzing the data, and reporting of psychological studies*

Code	Related	Type of Researcher Degree of Freedom	Label
Hypothesizing			
T1	R6	Conducting explorative research without any hypothesis	Hypothesis
T2		Studying a vague hypothesis that fails to specify the direction of the effect	Direction hypothesis
Design			
D1	A8	Creating multiple manipulated independent variables and conditions	Multiple manipulated IVs
D2	A10	Measuring additional variables that can later be selected as covariates, independent variables, mediators, or moderators	Additional IVs
D3	A5	Measuring the same dependent variable in several alternative ways	Multiple measures DV
D4	A7	Measuring additional constructs that could potentially act as primary outcomes	Additional constructs
D5	A12	Measuring additional variables that enable later exclusion of participants from the analyses (e.g., awareness or manipulation checks)	Additional IVS exclusion
D6		Failing to conduct a well-founded power analysis	Power analysis
D7	C4	Failing to specify the sampling plan and allowing for running (multiple) small studies	Sampling plan

Table 6.1 *Continued*

Code	Related	Type of Researcher Degree of Freedom	Label
Data collection			
C1		Failing to randomly assign participants to conditions	Random assignment
C2		Insufficient blinding of participants and/or experimenters	Blinding
C3		Correcting, coding, or discarding data during data collection in a non-blinded manner	Data handing/collection
C4	D7	Determining the data collection stopping rule on the basis of desired results or intermediate significance testing	Stopping rule
Data Analysis			
A1		Choosing between different options of dealing with incomplete or missing data on ad hoc grounds	Missing data
A2		Specifying pre-processing of data (e.g., cleaning, normalization, smoothing, motion correction) in an ad hoc manner	Data pre-processing
A3		Deciding how to deal with violations of statistical assumptions in an ad hoc manner	Assumptions
A4		Deciding on how to deal with outliers in an ad hoc manner	Outliers
A5	D3	Selecting the dependent variable out of several alternative measures of the same construct	Select DV measure
A6		Trying out different ways to score the chosen primary dependent variable	DV scoring
A7	D4	Selecting another construct as the primary outcome	Select primary outcome
A8	D1	Selecting independent variables out of a set of manipulated independent variables	Select IV
A9	D1	Operationalizing manipulated independent variables in different ways (e.g., by discarding or combining levels of factors)	Operationalizing manipulated IVs
A10	D2	Choosing to include different measured variables as covariates, independent variables, mediators, or moderators	Include additional IVs.
A11		Operationalizing non-manipulated independent variables in different ways	Operationalizing non-manipulated IVs
A12	D5	Using alternative inclusion and exclusion criteria for selecting participants in analyses	In/exclusion criteria
A13		Choosing between different statistical models	Statistical model
A14		Choosing the estimation method, software package, and computation of SEs	Method and package
A15		Choosing inference criteria (e.g., Bayes factors, alpha level, sidedness of the test, corrections for multiple testing)	Inference criteria
Reporting			
R6	T1	Presenting exploratory analyses as confirmatory (HARKing)	HARKing

The Scoring Protocol

For the current study, we created a protocol (see Appendix C) assessing to what extent a random selection of Standard Pre-Data Collection Registrations (SPR) and a random selection of Prereg Challenge Registrations (PCR) restricted potential opportunistic use of these 29 researcher DFs. The sample details are provided in the next section. Using the protocol, we assigned scores ranging from 0 to 3 to each of the 29 researcher DFs in each pre-registration in our sample. These scores had the following meaning: opportunistic use of the researcher DF was 0) not restricted at all, 1) restricted to some extent, 2) completely restricted (i.e., it was 'precise'), or 3) completely restricted by an explicit statement concerning the sole manner in which it was restricted (i.e., it was 'exhaustive' - an explicit statement that no deviation from the way it was registered would occur). As an example, we provide the protocol question to score researcher DF A4: 'Deciding on how to deal with outliers in an ad hoc manner'. This question and its answer categories were as follows:

Does the pre-registration indicate how to detect outliers and how they should be dealt with?

- **NO** not described at all → **A4 = 0**
- **PARTIAL** described but not reproducible on at least one of the following two aspects: what objectively defines an outlier (e.g., particular Z value, values for median absolute deviation statistic (MAD), interquartile range (IQR), Mahalanobis distance) and how they are dealt with (e.g., exclusion, method of Winsorisation, type of non-parametric test, type of robust method, bootstrapping) → **A4 = 1**
- **YES** reproducible on both aspects (objective definition of outliers & method of dealing with outliers) → **A4 = 2**
- **YES** like previous AND explicitly excluding other methods of dealing with outliers ("we will only use") → **A4 = 3**

This question was used to only score researcher DF A4, because A4 was is not related to other researcher DFs. As can be seen in the full protocol given in the Appendix C, scores on 13 of the researcher DFs showed some dependencies across different researcher DFs.

Sample

At the start of our study (August 17, 2016), 5,829 publicly available pre-registrations were listed on the pre-registrations search page on the Open Science Framework. These registrations included all types of pre-registrations then available on the Open Science Framework. As it was not possible to view how many belonged in each category without visually inspecting the individual pre-registrations, the Center for Open Science provided us with the URLs to the 122 public Prereg Challenge Registrations (PCRs) that had been submitted until August 16, 2016. Following the procedure described in the next section, we randomly selected a total of 53 PCRs and 53 SPRs. This sample size was based on our pre-registered power analysis conducted in GPower 3.1 for the key test of the differences in mean scores (outlined in the section 'variables' below) between the two types of pre-registrations using a Wilcoxon-Mann-Whitney U-test. This power analysis yielded a total required sample size of 106 (53 per group) for a power of .80 to detect a medium effect size of Cohen's d = 0.50. While we had no previous literature to base the estimated effect size on, we considered a medium effect size to be an indication of a practically relevant difference between the two types of pre-registrations. As pre-testing of our protocol indicated that the necessary total sample size of 106 already required a total of at least 212 hours of coding, time constraints and financial restrictions did not allow us to aim for a higher number than the number yielded by the power analysis.

Our approach followed our pre-registration with the following exception: we started the coding phase of our study (see paragraph 'Coding Procedure' in the next section) with 53 SPRs and 53 PCRs, but during the coding phase one of the PCR (with pre-registration number 54) turned out to have been withdrawn by the authors of the pre-registration. As the coding phase could not be finalized for this pre-registration, we excluded this pre-registration from our data file. Our final sample thus consisted of 53 SPRs and 52 PCRs.

Procedure

Selection of pre-registrations

Because the two types of pre-registrations were collected from different databases and because they differed in their standard features, we had slightly different selection procedures for each type of pre-registration. Both procedures followed our own pre-registration in all regards, except for an unforeseen issue discussed in Footnote 1. We now outline both procedures.

Standard Pre-Data Collection Registrations (SPRs)

Because it was impossible to determine how many of the 5,829 public preregistrations on the Open Science Framework were SPRs, we first selected a random set of 250 pre-registrations from the total of 5,829 results as a pre-selection. We did this by means of an R script (R version 3.2.4) that generated 250 random numbers between 1 and 5,829 and using these numbers to select the pre-registrations that corresponded to these numbers. Although the pre-registrations on the Open Science Framework were not numbered, the result pages were numbered and each contained 10 pre-registrations. We could thus use the numbering of the pages to locate the pre-registrations that corresponded to the generated random numbers. We then copied the hyperlinks of the selected pre-registrations into an Excel file in the order of the generated random numbers. Subsequently, we created two copies of these files and two coders (EC and CV) started independently at the top of the rows in their file and checked all 250 selected pre-registrations for eligibility on the basis of the selection criteria listed below. Next, the coded files of both coders were compared, and the first 53 pre-registrations for which both coders agreed on inclusion were included[1]. Finally, we allocated an identifying number (1 to 53) to each of the selected registrations in the order of which they were (randomly) selected by the code above.

Selection criteria for SPRs. The URLs to the preregistrations in our Excel file (see previous paragraph) sometimes referred to the main page of a pre-registered project, while in other cases they referred to sub-components of a pre-registered project (e.g. only to the analysis plan). The selection criteria were meant to be applied to the main page of a project. Thus, when the URL turned out to refer to a sub-component (observed by checking whether the front page of the registration contained a name before the slash, see the example in our own pre-registration), we copied the URL to the main page of the project (i.e. the URL of the page reached when clicking on the name before the first slash) in a new column next to the columns with the original URL. Then we only retained the URLs to the main page of each project and applied the following selection criteria, for Standard Pre-Data Collection Registrations:

- Only pre-registrations that were labeled as an *OSF-Standard Pre-Data Collection Registration* were included (the label was visible in the field 'Registration Supplement' on the main page of each pre-registration).
- Pre-registrations of replications studies were excluded. We excluded these because researcher DFs are typically more restricted in replications than in

[1] As a much smaller percentage of the registrations were Standard Pre-Data Collection Registrations (see selection criteria) than anticipated, this selection only yielded 31 registrations to include in our final sample. For this reason, we randomly selected a second sample of 250 pre-registrations with the same code used for the first sample of 250 pre-registrations. We repeated the described selection process once again with this second pre-selection of 250 pre-registrations (coders CV and MB). Because in this second sample we only had to continue until we had 22 pre-registrations that had been found eligible by both coders, it was agreed that the two files would be compared when both coders had found at least 30 pre-registrations eligible, rather than to continue coding all 250 pre-registrations. We considered 33 pre-registrations before we obtained the required 29. We added these 29 pre-registrations from the second round to the 31 pre-registrations from the first round, and then checked for duplicates in the total sample. We found six pre-registrations of which the hyperlink referred to the same project as another pre-registration's hyperlink. We removed the occurrences of these six pre-registrations that were lower on the list than the one that we retained, leaving us with 54 pre-registrations. We then selected the first 53 out of 54 to include in our final sample.

original studies as replication studies tend to follow the exact protocol of the original study. If one or more of the words ‘replication’, ‘RRR’, or ‘RPP’ were mentioned in the title of the pre-registration, the pre-registration was not considered eligible. If the title did not mention one of these words but an abstract (if present) or summary (if present) explicitly stated that the project was a replication, the study was not considered eligible either.

- Only pre-registrations written in English were included.
- Only fully accessible preregistrations were included. We excluded preregistrations to which access turned out to be restricted (e.g., when clicking on the link to the pre-registration yielded the message ‘this action is forbidden’).
- Only pre-registrations that contained at least one statistically testable hypothesis were included. We excluded pre-registrations in which no hypotheses could be tested by means of statistical analysis.
- “Test pre-registrations” we excluded. We evaluated whether a pre-registration was a real pre-registration or a test case by examining the title of a pre-registration for words like ‘test’, ‘test registration’, or ‘pre-registration demonstration’.
- Pre-registrations of pre-tests of materials and of pilot studies were excluded. If the title of the pre-registration contained the words ‘pre-test’ or ‘pilot’, the pre-registration was not considered eligible.
- Pre-registrations that were written by any of the coders were excluded. If the name of one of the coders appeared on the pre-registration, the pre-registration was not considered eligible.
- Pre-registrations that were part of the set of three pre-registrations that we used to test our protocol were excluded.

Prereg Challenge Registrations (PCRs)

We checked the 122 public Prereg Challenge entries that we received from the Center for Open Science for eligibility in a manner almost identical to the manner in which we checked the SPR for eligibility. The only difference was that we checked all 122 for eligibility, because the population was so much smaller than the population of SPR. We did however check the 122 PCR in randomized order, determined again by an R script (R version 3.2.4). The first 53 PCR that were considered eligible by two coders (MB and EC) were included in the sample.

Selection criteria for PCR. A number of the selection criteria for the other type of pre-registration (SPC) had by definition been met by public PCR because for entry into the Prereg Challenge they had already been reviewed by the Center for Open Science on the following criteria: being a Prereg Challenge Registration, being written in English, being fully accessible, not being a test registration, and containing at least one testable hypothesis. We therefore only had to apply the following remaining selection criteria (as stipulated in our own pre-registration):

- Pre-registrations of replications studies were excluded (see selection criteria for SPR).
- Pre-registrations of pre-tests of materials and of pilot studies were excluded (see selection criteria for SPR).
- Pre-registrations that were written by any of the coders were excluded (see selection criteria for SPR).
- Pre-registrations that were part of the set of three pre-registrations that we used to test our protocol were excluded.

Coding Procedure

Each pre-registration was coded independently by two coders, according to a scheme generated by an R script (R version 3.2.4). The coders first entered the number of hypotheses they encountered in the pre-registration into their coding sheet, and then followed the protocol to score each of the 29 researcher DFs. In the early phase of coding, scoring one pre-registration took each coder on average 45 to 60 minutes, but this length decreased as the coders gained more experience. When all coders finished their coding, the scores given to each pre-registration by each pair of coders were automatically compared by means of another R script (R version 3.2.4). We computed the percentage of scores on which the coders agreed using yet another R script (R version 3.2.4), and we computed an agreement percentage for the number of hypotheses that the coders had counted.

Across all pre-registrations and coding pairs, coders agreed on the number of hypotheses in only 14.29% of the scores given. Across SPR only, this agreement percentage was 15.09%, and across PCR only, this was 13.46%. The agreement percentages were much higher for the coding of the researcher DFs: across all researcher DFs, pre-registrations and all coding pairs, the same score had been given in 74.84% of the cases. For SPR only, the same score had been given in 77.75% of the cases, and for PCR in 71.88%.

In case of discrepancies on the coding of researcher DFs, the two coders discussed their coding until they agreed on a final score. As these discussions turned out to solve all discrepancies, it was not necessary to ask a third coder to solve any disagreements. We did not attempt to resolve discrepancies with respect to the number of hypotheses that had been counted as the specific hypotheses were not part of our analyses but merely served as an indication of clarity and specificity of the pre-registrations. As the protocol instructed coders to evaluate each hypothesis in a pre-registration and to then assign to each researcher DF the lowest score for each hypothesis, the scores were based on the worst pre-registered hypothesis in each pre-registration regardless of how many hypotheses had been counted.

Variables of interest

In line with our own pre-registration, we computed two mean scores: 1) a score indicating to what extent a pre-registration restricted opportunistic use of researcher DFs: the Registration Restriction Score (RRS), and 2) a score indicating to what extent each researcher DF was restricted across pre-registrations: mean score per researcher DF. The RSS was computed as the unweighted arithmetic mean of the scores (0-3) of all 29 researcher degrees of freedom in our protocol. As there were dependencies between some of the researcher degrees of freedom (see Table 6.1), some researcher DFs carried more weight than others. The restriction score per researcher DF was calculated as the unweighted arithmetic mean of the scores (0-3) across each set of (53 and 52) pre-registrations. These are presented in Table 6.2 as the means per researcher DF.

Statistical analyses

To test our hypothesis that PCRs restrict the potential for opportunistic use of researcher DFs to a greater extent than SPRs, we compared the median RSS of the PCRs to the median RSS of the SPRs by means of a one-tailed two-group Wilcoxon Rank Sum Test (i.e. equivalent to a Wilcoxon-Mann-Whitney U test). The reason we pre-registered the use of a non-parametric test was that we expected the Pre-registration Restriction Scores to be non-normally distributed, and this non-parametric test is robust against non-normality while still being relatively powerful (Bakker & Wicherts, 2014b). We then conducted explorative follow-up analyses to investigate on which of the researcher degrees of freedom the two types of pre-registrations differed. To this end, we conducted 29 two-tailed Wilcoxon Mann Whitney U tests to compare the median scores of the two groups per researcher degree of freedom. In all analyses we maintained an inference criterion of alpha = 0.05 (as pre-registered).

As a measure of effect size of the differences in medians, we employed Cliff's Delta (Cliff, 1993, 1996). Cliff's Delta is a measure of effect size for comparing the central tendency of two distributions, say U and V (Macbeth, Razumiejczyk, & Ledesma, 2011), and linearly related to P(U > V), i.e. the probability that a randomly drawn observation from U exceeds a randomly drawn observation from V. Cliff's Delta varies from-1 to 1. We used Cliff's Delta because it is suitable to assess effect size for comparing central tendency of ordinal variables, without making any assumptions on the distributions of the two variables, and because it is easily interpretable. Values under 0.147 are considered 'negligible', values between 0.147 and 0.330 are considered 'small', values between 0.330 and 0.474 are considered 'medium', and values larger than 0.474 are considered 'large' (Romano, Kromrey, Coraggio, & Skowronek, 2006), see https://cran.r-project.org/web/packages/effsize/effsize.pdf.

Table 6.2 *Mean scores per researcher DF and distributions of scores per researcher DF for Standard Pre-data Collection Registrations (SPR) and Prereg Challenge Registrations (PCR), and differences in median scores between SPR and PCR per researcher DF.*

		Standard Pre-Data Collection Registrations (SPR)						Prereg Challenge Registrations (PCR)						Differences in Median	
Researcher DF		Mean (SD)	0	1	2	3	NA	Mean (SD)	0	1	2	3	NA	Test	Cliff's D
Hypothesizing															
T1	Hypothesis	1.98 (0.31)	1.9	-	96.2	1.9	-	2.02 (0.14)	0.0	-	98.1	1.9	-	W = 1404, p = 0.571	0.02
T2	Direction hypothesis	1.60 (0.84)	20.8	-	77.4	1.9	-	1.54 (1.20)	34.6	-	42.3	23.1	-	W = 1422, p = 0.749	0.03
Design															
D1	Multiple manipulated IVs	0.38 (1.02)	64.2	-	0.0	9.4	26.4	1.03 (1.42)	46.2	-	1.9	23.1	28.8	W = 880, p = 0.026	0.22
D2	Additional IVs	0.00 (0.00)	100	-	-	0.0	-	0.12 (0.58)	96.2	-	-	3.8	-	W = 1431, p = 0.155	0.04
D3	Multiple measures DV	1.25 (0.98)	37.7	-	62.3	0.0	-	1.62 (0.80)	19.2	-	80.8	0.0	-	W = 1633, p = 0.037	0.19
D4	Additional constructs	0.00 (0.00)	100	-	-	0.0	-	0.00 (0.00)	100	-	-	0.0	-	NA	0.00
D5	Additional IVs exclusion	0.87 (0.92)	45.3	26.4	24.5	3.8	-	1.23 (0.70)	13.5	51.9	32.7	1.9	-	W = 1729.5, p = 0.017	0.26
D6	Power analysis	0.72 (0.91)	58.5	11.3	30.2	0.0	-	0.96 (0.99)	50.0	3.8	46.2	0.0	-	W = 1551, p = 0.212	0.13
D7	Sampling plan	0.47 (0.58)	56.6	39.6	3.8	0.0	-	0.71 (0.58)	34.6	57.7	5.8	0.0	-	W = 1641, p = 0.034	0.21
Data collection															
C1	Random assignment	0.27 (0.67)	66.0	1.9	9.4	0.0	22.6	0.86 (0.92)	34.6	11.5	25.0	0.0	28.8	W = 1028.5, p = 0.001	0.36
C2	Blinding	1.00 (1.00)	3.8	1.9	3.8	0.0	90.6	0.02 (0.14)	92.3	1.9	0.0	0.0	5.8	W = 50.5, p < 0.001	-0.59
C3	Data handing/collection	0.04 (0.19)	96.2	3.8	0.0	0.0	-	0.04 (0.19)	96.2	3.8	0.0	0.0	-	W = 1379, p = 0.992	0.00
C4	Stopping rule	0.47 (0.58)	56.6	39.6	3.8	0.0	-	0.71 (0.58)	34.6	57.7	5.8	0.0	-	W = 1641, p = 0.034	0.21
Data Analysis															
A1	Missing data	0.19 (0.39)	81.1	18.9	0.0	0.0	-	0.76 (0.55)	28.8	63.5	5.8	0.0	-	W = 2065.5, p <0.001	0.53
A2	Data pre-processing	0.50 (0.84)	9.4	-	1.9	0.0	88.7	0.50 (0.93)	11.5	-	3.8	0.0	84.6	W = 23, p = 0.935	-0.04
A3	Assumptions	0.04 (0.19)	96.2	3.8	0.0	0.0	-	0.18 (0.48)	84.6	9.6	3.8	0.0	-	W = 1488, p = 0.070	0.10

Table 6.2 *Continued*

		Standard Pre-Data Collection Registrations (SPR)						Prereg Challenge Registrations (PCR)						Differences in Median	
Researcher DF		Mean (SD)	0	1	2	3	NA	Mean (SD)	0	1	2	3	NA	Test	Cliff's D
A4	Outliers	0.25 (0.62)	84.9	5.7	9.4	0.0	-	0.69 (0.92)	57.7	19.2	19.2	3.8	-	W = 1751, p = 0.003	0.27
A5	Select DV measure	1.25 (0.98)	37.7	-	62.3	0.0	-	1.62 (0.80)	19.2	-	80.8	0.0	-	W = 1633, p = 0.037	0.19
A6	DV scoring	0.55 (0.70)	56.6	32.1	11.3	0.0	-	0.65 (0.65)	44.2	46.2	9.6	0.0	-	W = 1519, p = 0.317	0.10
A7	Select primary outcome	0.00 (0.00)	100	-	-	0.0	-	0.00 (0.00)	100	-	-	0.0	-	NA	0.00
A8	Select IV	0.59 (1.19)	58.5	-	1.9	13.2	26.4	1.14 (1.48)	44.2	-	0.0	26.9	28.8	W = 853.5, p = 0.083	0.18
A9	Operationalize manipulated IVs	1.05 (1.26)	41.5	-	18.9	13.2	26.4	1.92 (1.19)	17.3	-	25.0	28.8	28.8	W = 982.5, p = 0.004	0.36
A10	Include additional IVs	0.00 (0.00)	100	-	-	0.0		0.12 (0.58)	96.2	-	-	3.8	-	W = 1431, p = 0.155	0.04
A11	Operationalize non-manipulated IVs	0.43 (0.66)	28.3	11.3	3.8	0.0	56.6	0.63 (0.67)	26.9	25.0	5.8	0.0	42.3	W = 405, p = 0.229	0.17
A12	In/exclusion criteria	0.87 (0.92)	45.3	26.4	24.5	3.8	-	1.21 (0.72)	15.4	50.0	32.7	1.9	-	W = 1710.5, p = 0.024	0.24
A13	Statistical model	0.85 (0.77)	37.7	39.6	22.6	0.0	-	1.31 (0.51)	1.9	65.4	32.7	0.0	-	W = 1846, p = 0.001	0.34
A14	Method and package	0.08 (0.38)	96.2	0.0	3.8	0.0	-	0.13 (0.44)	90.4	5.8	3.8	0.0	-	W = 1455.5, p = 0.254	0.06
A15	Inference criteria	0.17 (0.43)	84.9	13.2	1.9	0.0	-	1.08 (0.33)	1.9	88.5	9.6	0.0	-	W = 2516, p < 0.001	0.83
Reporting															
R6	HARKing	0.00 (0.00)	100	-	0.0	0.0	-	0.00 (0.00)	100	-	0.0	0.0	-	NA	0.00

Note: Distribution of scores are given in percentages. Not all percentages add up to exactly 100% due to rounding to 1 decimal of each individual percentage. A '-' sign indicates that this score was not available.

Our final data file contained 35 columns: one column with the pre-registration number, one column with the title of the pre-registration, one column with the hyperlink to the pre-registration, one column with group number (0 for pre-registrations from the group ‘Standard Pre-Data Collection Registrations’ and 1 for pre-registrations from the group ‘Prereg Challenge Registrations’), two columns with the number of hypotheses that the coders counted in the pre-registration (one column for coder 1, and one column for coder 2), and 29 columns with one researcher degree of freedom each. For our analyses, we used an R script (R version 3.2.4), which was an elaborated version of our pre-registered analysis script. Besides our decision not to code pre-registration no. 54 (see sample description above), we did not exclude any data. After coding of the data, the data file was checked for missing values, and coders who left values missing were asked to code the variables they missed. Values coded as 99 indicated that the specific variable was not applicable in the coded registration. As we indicated in our pre-registration, we dealt with missing values in the following way. For our confirmatory analysis, we employed two-way imputation, i.e. based on the corresponding row and column means. For our follow-up analyses (conducted at the column-level), we employed pairwise deletion of missing values. The tables in this manuscript were created using an R Markdown script. Please note however that in these tables, the researcher DF labels were added manually, and some column names were adapted. The R Markdown script can be used directly when loading the workspace of the analysis script.

RESULTS

Overall difference between Standard Pre-Data Collection Registrations (SPRs) and Prereg Challenge Registrations (PCRs)

Our first research question was whether pre-registrations that were written following more detailed and specific requirements (PCR) restricted opportunistic use of researchers degrees of freedom more than pre-registrations that were written following less detailed and specific requirements (SPR). In line with our pre-registered confirmatory hypothesis, Prereg Challenge Registrations received higher median Registration Restriction Scores (RSS): the median RRS was significantly higher in PCRs (Mdn = 0.81) than in SPRs (Mdn = 0.57), U = 2053, p = < .001. The difference was large (Cliff’s Delta = 0.49). Although RRS scores could range from 0 to 3, the highest RRS received by SPRs was 1.05, while the highest score received by PCRs was 1.47. The median and maximum RRS scores in both types of pre-registrations suggest that the extent to which opportunistic use of researcher DFs was restricted was overall low in both SPRs and PCRs. However, it should be taken into account that our protocol was rather strict.

Differences between Standard Pre-Data Collection Registrations (SPR) and Prereg Challenge Registrations (PCR) at the level of the individual researcher DFs

As pre-registered, we conducted follow-up analyses to examine on which of the researcher DFs there were statistically significant differences in the distribution of scores between SPRs and PCRs. In Table 6.2, we report for each type of pre-registration the mean score of each researcher DF, and the distribution of the scores for each researcher DF. In addition, we provide for each researcher DF the results of the Wilcoxon Rank Sum Test (equivalent to the Wilcoxon Mann Whitney U test), and Cliff's Delta as a measure of effect size. Note that it can be observed from Table 6.2 that scores of 3 did not occur or were rare for most researcher DFs, confirming that the extent to which opportunistic use of researcher DFs was restricted was low in both types of pre-registration at the level of researcher DFs. Due to the differences in mean scores between many of the researcher DFs within each phase (hypothesizing, designing, data collecting, analyzing, and reporting), it is not feasible to compare how well the researcher DFs in each phase were restricted. What we can say, however, is that the researcher DFs pertaining to the hypothesizing were relatively well restricted in both types of pre-registrations, and that the researcher DF pertaining to reporting was not restricted at all in either of the two types of pre-registrations. We address which researcher DFs tended to be better restricted than others while we report the differences between SPRs and PCRs in mean score on individual researcher DFs below.

In total, there were 15 researcher DFs on which there was no significant difference between the scores obtained by PCRs and SPRs, while 13 researcher DFs were significantly better restricted by PCRs than by SPRs. On eight of these 13 researcher DFs the size of the difference was small in terms of Cliff's Delta (0.147-0.330), on three researcher DFs the size of the difference was medium (0.330-0.474), and on two researcher DFs the difference in size was large (larger than 0.474). Finally, there was one researcher DF that was better restricted by SPR than by PCR. We now zoom in on the researcher DFs in each of these categories.

Researcher DFs showing no differences between PCRs and SPRs

Many of the researcher DFs that failed to show a difference between the two types of pre-registrations were poorly restricted throughout. Researcher DFs related to HARKing were restricted by neither the PCRs nor the SPRs (D4, A7, and R6). That is, none of the pre-registrations in our sample explicitly specified that the confirmatory analysis section of the paper would *not* include another dependent variable than the ones specified in the preregistration. These three researcher DFs are among the most obvious researcher DFs that pre-registrations are meant to restrict, and their seemingly poor restriction in the pre-registrations in our sample might be reason for concern. Researcher DFs pertaining to measuring additional variables (D2 and A10) were also poorly restricted. Although the PCR format asks authors to list all variables in the study, it does not ask for what purpose variables are measured that are not included in the analyses. Similarly, treating of data during data collection (C3) was hardly restricted by the pre-registrations and indeed neither format addresses this issue. Finally, two researcher DFs concerning the statistical model (A3 and A14) were poorly restricted. Despite the PCR format being rather specific about which aspects should be considered in the description of the statistical model, it does not mention violations of statistical assumptions, estimation method, or software to be used.

Researcher DFs with small differences between PCRs and SPRs.

Higher scores were obtained by PCRs than by SPRs on researcher DFs that pertained to the operationalization of the variables in the study: D1, D3 and A5, and D5 and A12. These differences indicate that the operationalization of variables was slightly better restricted through the PCR format, which may be due to the PCR format explicitly asking authors to be very specific in the description of all their variables, while the SPR format provides no instruction with respect to the description of variables. Most of these researcher DFs were relatively well restricted compared to other researcher DFs in both types of pre-registration, although the researcher DF pertaining to measuring the independent variable (D1) was generally not as well restricted as the other four. Researcher DFs that relate to the sampling plan and data collection stopping rule (D7 and C4) were also better restricted by the PCR format. Since the PCR format leaves room for authors to state that they will recruit a sample of ‘at least’ a certain size and that therefore continued sampling after intermediate testing is not precluded might explain that the difference was rather small for the sampling plan and data collection stopping rule. Compared to other researcher DFs, these researcher DFs were moderately well restricted. Finally, there was a small difference on how to deal with outliers (A4) between the two types of preregistrations. In PCRs outliers were mentioned more often than in SPRs, as they are explicitly mentioned as part of the section ‘data exclusion’ in the PCR format. However, although this researcher DF was relatively more restricted than other researcher DFs, most of the PCRs did not specify *how* they would define outliers still leaving some room for defining outliers in different ways.

Researcher DFs with medium differences between PCR and SPR

There were three researcher DFs where the size of the difference in medians was medium. The first of these pertained to randomization (C1). Restriction was not great in either set of pre-registration, but clearly better in the PCR format, which prompts authors to at least mention randomization. However, if the sample of PCRs contained more survey studies using software such as Mturk or Qualtrics, the difference may also be explained by our decision to give a score of 3 when such studies stated that randomization would occur through the software's randomization device.

The second researcher DF that showed a medium difference between PCR and SPR concerned the operationalizing of independent variables (A9). This researcher DF was well restricted in the PCR format, but only to some extent with the SPRs. However, we note that the mean scores on this researcher DF were heavily influenced by the relatively high number of pre-registrations that received a score of 3, particularly among PCRs. The reason for the occurrence of this unusually high score was that when the analysis concerned a t-test (or a non-parametric equivalent), we considered this researcher DF excluded by definition and therefore assigned a score of 3. t-tests were rather common, and may have been more common among PCRs. An explanation for the difference between PCRs and SPRs may lie in that a number of PCRs provided analysis scripts (leading to a score of at least 2), whereas this never occurred in SPRs. Finally, the researcher DF related to the choice of statistical model (A13) was, compared to other researcher DFs, relatively more restricted in both types of pre-registration, but the separate section dedicated to the data analyses in the PCR format may have been the cause of clearly better restriction in PCR than in SPR.

Researcher DFs with large differences between PCR and SPR

There were two researcher DFs on which the difference between PCRs and SPRs was large. The first pertained to dealing with missing data (A1), which was not too well restricted in either type of pre-registration. In the PCR format there was a question about dealing with missing values, prompting authors to at least mention missing values, while the SPR format lacked such a prompt. However, most PCRs were not particularly detailed about what constituted a missing value, or how missing values were to be dealt with, rendering the scores on this researcher DF lower than other researcher DFs in the Prereg Challenge.

The second researcher DF concerned the inference criteria (A15); in the PCR format, this researcher DF was moderately well restricted, whereas among SPRs, this researcher DF was hardly restricted at all. Again, the PCR format included a specific question asking authors to indicate which inference criteria they would use, while in the SPR format, merely asking authors to provide a summary of their project apparently did not induce most authors to mention inference criteria.

Researcher DFs that were better restricted by SPR than by PCR

Lastly, there was one researcher DF that was better restricted by SPRs than by PCRs. This researcher DF pertained to blinding (C2), and the difference was large. In a very large number of SPRs this researcher DF was coded as 'Not Applicable', whereas it was coded as 0 in the vast majority of PCRs. This difference is explained by the way our coding protocol was formulated. The protocol question consisted of two parts. First, it asked whether blinding of participants and/or experimenters was mentioned. In SPRs the word 'blinding' hardly ever appeared, resulting in most SPR being coded as 'not applicable', leaving few pre-registrations of this type to be coded. Second, if blinding was mentioned, the coder checked whether the pre-registration described procedures to blind participants to and/or experimenters to conditions. This was coded as 0 in the vast majority of PCRs because authors responded that blinding was not applicable or that no blinding occurred in their study to the question in the PCR format asking authors to describe blinding procedures in their study. We decided to strictly adhere to our pre-registered protocol, and answer 'yes' to the question of whether blinding was mentioned and then almost always coded '0', because responses such as 'no blinding occurred' were not considered to describe blinding procedures. We realize that this was a choice that resulted in scores that were unrepresentative of how well pre-registrations of the PCR type restricted this researcher DF, and that the PCR scores on this researcher DF should not be over-interpreted.

DISCUSSION

In this study, we investigated whether pre-registrations that were written following more detailed and specific requirements (Prereg Challenge Registrations) restricted opportunistic use of researcher's degrees of freedom more than pre-registrations that were written following less detailed and specific requirements (Standard Pre-Data Collection Registrations). As expected, PCRs generally restricted opportunistic use of researcher DFs better than SPRs. In addition, we examined for each researcher DF separately whether it was restricted better in PCRs or in SPRs. A majority of researcher DFs were more restricted by PCRs than by SPRs, although the difference was only large for two researcher DFs pertaining to missing values and to inference criteria, which were explicitly asked in the PCR format but not in the SPR format. Finally, we attempted to clarify which researcher DFs were better restricted than others, but found no systematic overall differences between restrictions of the researcher DFs appearing in different phases of a research project. However, researcher DFs pertaining to the hypothesizing phase were generally best restricted, and those pertaining to HARKing were worst restricted. Although our design is correlational and hence cannot entirely preclude confounding factors in comparing the pre-registration formats, comparison of the two formats suggests that differences in the specific requests in these formats underlie most differences; in line with our hypotheses, more specific requests are associated with more restricted researcher DFs.

Based on these results, we argue that reminding authors of all details to be described in a pre-registration produces pre-registrations that are better at doing what they are supposed to do: protecting authors from their own biases in confirmatory hypothesis testing. Specific guidelines and requirements are necessary. Without these specific guidelines and requirements, authors are free to leave as many researcher DFs unrestricted as they wish, rendering the label 'pre-registered' meaningless.

To illustrate what it could mean in practice when pre-registrations obtain low scores, we provide an example. Suppose that a pre-registration receives a score of 0 on six out of the 15 researcher DFs concerning the data analyses (which occurred quite often). If we conservatively assume that each of these researcher DFs is associated with two options to analyze data, this offers the researcher a total of 2^6 = 64 ways to analyze the data of their pre-registered study. It is clear therefore, that many current pre-registrations in both formats are insufficiently specific and still allow at least some ways of *p*-hacking. Thus even better, more specific and stricter instructions are needed to help researchers produce pre-registrations that further lower the risk of bias in (psychological) studies.

Although our hypothesis that PCRs restricted researcher DFs to a larger extent than SPRs, we had expected larger differences between the two types of pre-registrations. More specifically, we did expect the SPRs to receive low scores, but we (informally) expected PCRs to receive higher scores than they eventually did. Especially at the level of the individual researcher DFs, the scores for most of the PCRs were rather disappointing. On many researcher DFs, the PCR scores were hardly better than the SPR scores. Most often this was the case when a researcher DF was not explicitly addressed in the PCR format, supporting the idea that the quality of pre-registrations indeed depends on the format according to which it is written. Nonetheless, the low scores on both types of pre-registrations were also partly attributable to the strictness of our protocol. That is, each coded researcher DF could only obtain a '3' if it explicitly excluded other options than those mentioned in the pre-registration (referring to exhaustiveness of the pre-registration). Because none of the pre-registrations in our sample did this explicitly, scores of '3' were only very rarely assigned.

One might argue that pre-registering a study in itself is a form of explicitly excluding researcher DFs. However, we have learned from studies on registrations of randomized controlled trials that 'outcome switching' and HARKing are rather common even in publications of studies that were pre-registered (Chan, 2008;

Chan & Altman, 2005; Chan, Hrobjartsson, et al., 2004; Chan et al., 2008; Goldacre, 2016). Hence, by including the exhaustiveness criterion in our protocol we assessed if authors explicitly prohibited changes in their protocol when carrying out and reporting their study. The idea is that if a pre-registration explicitly excludes changes in protocol, actual changes can no longer be defended by reasons such as by "lack of statistical significance", "journal space restrictions", or "lack of clinical importance" (Chan, Hrobjartsson, et al., 2004)". Moreover, it is then no longer possible to deny any deviations from the protocol. In the study of outcome switching mentioned earlier (Chan, Hrobjartsson, et al., 2004), 86% of authors who were asked whether there had been any outcome measures that were not reported in their published article denied the existence of such measures, despite the registrations of their study providing clear evidence for these outcome measures.

In addition to being strict, our protocol was also far from perfect. Coders agreed on the coding in 74% of the scores, indicating that scoring based on the protocol was not unambiguous. During the discussions to resolve discrepancies, we learned that the coders sometimes had different interpretations of the protocol. We tried to reach agreement on each difference of interpretation in order to code all pre-registrations in a consistent manner. In some cases, it proved difficult to apply (some parts of) the protocol. For instance, when pre-registrations pertained to studies in which secondary data analysis occurred, the researcher DFs related to data collection and some parts of the design were impossible to code. When pre-registrations pertained to studies in which other estimation methods were used, such as Bayesian statistics or point estimation, many of the questions in our protocol were redundant. All issues regarding the protocol were recorded in notes kept by the coders, and these notes will be used to improve the protocol for future use. Another reason for the 74% agreement was that several pre-registrations were so ill structured and loosely written that it was hard to make sense of the planned research. For example, it was often very difficult to understand what the main hypotheses were, as demonstrated by the percentage of agreement on the number of hypotheses in a pre-registration being shockingly low (around 15%). Since all coders are all psychologists trained in formulating hypotheses, we interpret this finding as evidence of researchers' difficulty and inability to clearly state testable hypotheses and expectations. Our interpretation is in line with recent findings of Hartgerink, Wicherts, & van Assen (2017), who found that only in 15 out of 178 gender effects reported in published studies it was clearly stated whether these effects were as expected or hypothesized. Similarly, Motyl et al. (2017) often wrongly selected statistical results as focal results (Nelson, Simmons, & Simonsohn, 2017), and researchers in the Open Science Collaboration experienced some difficulties in specifying the main hypothesis in the 100 primary studies included in the Reproducibility Project Psychology (Open Science Collaboration, 2015). Thus, psychological researchers should improve the formulation of the hypotheses in both pre-registrations as well as in their published studies.

Another potential limitation is that our findings do not provide causal evidence on whether pre-registering studies restrict the opportunistic use of researcher degrees of freedom. Direct evidence might be obtained if the same studies are randomly assigned to both a 'pre-registration' condition and a 'no pre-registration' condition and then analyzed and reported by two different teams.

Based on our results, we can suggest a number of ways in which to improve pre-registration formats so that they result in pre-registrations that better restrict opportunistic use of researcher degrees of freedom. First, we deem it clear that pre-registration formats need to help authors in writing pre-registrations that are specific, precise, and exhaustive. It is extremely difficult to write a good pre-registration, and authors need proper guidance. We therefore deem it unlikely that formats as proposed by AsPredicted, which consists of eight rather broad questions, produce pre-registrations that effectively restrict opportunistic use of the many researcher degrees of freedom that exist in most studies. We would have liked to assess this empirically, but were not able to. In the first design of our own study, we aimed to include pre-registrations from AsPredicted too, but the hosts of the website declined our request to share the public pre-registrations created on their site. Currently, some of the AsPredicted registrations are shared on the Open Science Framework, but at the time of the start of this study, we had no (legal) means of including these pre-registrations in our study.

In addition to offering specific guidelines and requirements, pre-registrations will likely improve if formats offer a way for authors to number their hypotheses, and to detail the way in which they will test each hypothesis separately. Thus, rather than having authors first list all their hypotheses and sub-hypotheses, followed by the design, procedure, variables, and analyses, it will be clearer to first specify the design of the study, and then specify for each hypothesis how this design will be used to test the hypothesis, which variables are measured in what manner for the test of this hypothesis, and how the hypothesis will be tested statistically. Furthermore, pre-registrations will benefit from peer review. The PCRs in our sample had been reviewed by the Center for Open Science, but only in terms of whether it met all entry criteria. This form of review might have made authors of PCRs take writing the pre-registration more seriously than authors of SPRs, which are not reviewed at all. Review by peers who are knowledgeable on the research topic of the pre-registration, as occurs in Registered Reports (Chambers, 2013, 2015; Chambers et al., 2014) will likely significantly improve pre-registrations. Moreover, when peer reviewers assess a pre-registration prior to the execution of a study, they can actively contribute to improving the study design at precisely the right moment, thereby increasing the benefit of pre-registration even further.

In addition to reducing the effects of opportunistic use of researcher degrees of freedom in the published literature, pre-registration also benefits researchers directly. Although writing a good pre-registration takes time and effort, it eventually results in a more efficient research cycle. Writing the pre-registration of the current study was a lengthy process, but the extensive discussions during the process of writing our pre-registration helped us identify many of the potential issues in our study. During the analysis phase, we did not have to think about which analysis to conduct, how to deal with missing values, or what to do with outliers; the analysis script was a good as ready to be used. Writing the introductory and method section of the current manuscript was considerably faster and more straightforward than in our previous manuscripts reporting about un-registered studies or badly-registered studies. In addition, pre-registrations teach you a lot about all the issues you did not or could not foresee. As in our own pre-registration of the current study, there are always issues that were not anticipated, or which appear to have been a poor decision in hindsight. However, deviations from pre-registrations are not forbidden, as long as we are open and transparent about them. Pre-registrations are meant to reduce *opportunistic* use of researcher degrees of freedom, not to eliminate researcher degrees of freedom from research altogether. Authors of pre-registered studies are free to conduct as many (pre-registered or non pre-registered) exploratory analyses as they see fit and can even make changes to their confirmatory analyses, as long as they openly report (and discuss) these changes, and rightly distinguish exploratory analyses from the truly confirmatory analyses. These, and many other benefits of pre-registration (Wagenmakers & Dutilh, 2016) renders pre-registration a choice that does not tie your hands, but guides your hands towards more trustworthy and replicable results.

CHAPTER 7

Epilogue

Recent studies have highlighted that not all published findings in the scientific literature are trustworthy (e.g., Baker, 2016; Open Science Collaboration, 2015; Ioannidis, 2005), suggesting that currently implemented control mechanisms such as high standards for the reporting of research methods and results, peer review, and replication, are not sufficient. In psychology in particular, solutions are sought to deal with poor reproducibility and substandard replicability of research results (e.g. Munafò et al., 2017) . These problems are believed to be due to bias resulting from failure to publish all relevant research findings (publication bias, Dwan et al., 2008; Ioannidis, Munafo, Fusar-Poli, Nosek, & David, 2014), common errors in the reporting of statistical results (e.g.,, Nuijten et al., 2016), and flexibility in the way data are analyzed (Simmons et al., 2011; Wagenmakers et al., 2011). In this dissertation project I considered these problems from the perspective that the scientific enterprise must better recognize the human fallibility of scientists (Mahoney, 1976, 1979). First, I studied perceptions of the characteristics of scientists. Then I examined the prevalence of statistical reporting errors, and whether collaboration on statistical analyses might reduce these errors. Finally, I presented an overview of many potential biases in hypothesizing, designing, collecting, analyzing, and reporting of psychological experiments, and evaluated the promising method of pre-registration as a way to deal with these biases.

Perception of the characteristics of scientists

In Chapter 2 we aimed to learn to what extent the human fallibility of scientists is currently recognized. Specifically, we studied the degree to which scientists and lay people believe in the storybook image of the scientist: the image that scientists are more objective, rational, open-minded, intelligent, honest and communal than other human beings. In four surveys among highly-educated lay people and scientists, we found that in both groups, belief in this storybook image of the scientist was strong. Moreover, scientists perceived larger differences between scientists and other professionals than lay people did. In addition, we found indications that scientists may be prone to the human tendency to attribute higher levels of desirable traits to people in one's own group than to people in other groups. We concluded that scientists do not readily recognize their own fallibility and may believe that other scientists are more fallible than themselves.

The survey methodology we employed in Chapter 2 was subject to a number of limitations that are typical for (online) surveys: a low response rate and potential selection bias may have limited the generalizability of our findings. In addition, our sample of highly-educated lay people from the United States may not have been fully representative of the population of highly-educated lay people. Our research designs also presented a number of limitations. First, as objective data on higher levels of objectivity, rationality, open-mindedness, intelligence, integrity or communality among scientists is lacking, we cannot exclude the possibility that scientists and lay people are indeed correct in their positive perceptions of scientists. Second, our method of studying whether scientists recognize their own fallibility was rather indirect. A valuable avenue for future research may therefore lie in studies that compare the actual levels of objectivity, rationality, open-mindedness, integrity, intelligence, and communality that scientists and other professionals display. One may also more directly study how scientists value imposed system and policy changes: by conducting surveys or structured interviews asking scientists (1) how they perceive their own fallibility and that of fellow scientists, (2) to what extent they regard human fallibility to be a problem in science, and (3) whether they feel that solutions to better deal with human fallibility are needed. We deem it important that scientists acknowledge their own fallibility, because such self-reflection is the first line of defense against potential human error aggravated by confirmation bias, hindsight bias, motivated reasoning, and other human cognitive biases that could well play a negative role in the scientific process.

Human error in psychological science

In Chapters 3 and 4 we zoomed in on psychological science and focused on human error in the use of null hypothesis significance testing (NHST), and considered a potential best practice of collaboration on statistical analysis that might help reduce the likelihood of these errors

In Chapter 3, we surveyed authors of 697 empirical articles published in six flagship psychology journals to investigate whether the authors had worked alone on the analysis or whether they had worked together in a so-called ‘co-pilot model of statistical analysis’ (Wicherts, 2011) in the first or only study reported in the article. We examined this by asking how many of the authors of the article had been involved in various aspects of the data analysis and the reporting of the results. Despite the fact that more than 99% of the articles had multiple authors, in the majority of these articles (around 60-75 %), only one author had been involved in conducting the analyses, writing down the sample details, and writing up the results. Moreover, in the majority of articles, only one author had access to the data at the time we conducted our survey (about a year after publication). This suggests that co-piloting or shared responsibility for the statistical analyses, is uncommon in psychology. We also documented errors in the statistical results reported in the 697 articles by means of the automated procedure ‘statcheck’. As in previous research, we found alarmingly high error rates: overall, 63% of the articles contained at least one statistical reporting error, and 20% of the articles contained at least one *p*-value that was erroneous to such a degree that it may have affected decisions about statistical significance. The probability that a given *p*-value was inconsistent was over 10%. Although we expected co-piloting to lower error rates, we failed to find an association between co-piloting and statistical reporting errors in the first or only study in the articles.

Potential limitations of the study reported in Chapter 3 arose because of its use of an (online) survey: the survey may have suffered from response bias, with authors of articles with more errors being less likely to respond than authors of articles with less errors. In addition, our survey required authors to remember author contributions for an article that had been published at least one year earlier, lowering the probability that the answers were accurate. In addition, our choice to include only those statistical results reported in the first or only study in the articles rendered the total number of statistical results to be associated with co-piloting practices to be relatively low. These limitations were dealt with in the study reported in Chapter 4, where we employed a larger set of articles and used a different method to determine whether co-piloting had occurred in these articles.

In Chapter 4, we scanned the full population of psychology articles ever published in the multidisciplinary Open Access journal PLOS ONE (14,946) for statistical reporting errors, again using statcheck. To measure whether co-piloting had occurred in these articles, we made use of the author contribution section of the articles stating whom of the authors had been responsible for the data analyses. Here, we found error prevalences that were highly similar to those established in Chapter 3, but the percentage of articles in which co-piloting occurred according to these author contribution sections was considerably higher than according to our survey in Chapter 3: 76.4% versus 39.7%. Despite having more statistical power and despite using a method that was not subject to the limitations of survey methodology, we again failed to find a relationship between a reduced likelihood of statistical reporting errors and collaboration on the statistical analyses.

Even though the study reported in Chapter 4 addressed a number of limitations of the study reported in Chapter 3, there was another set of potential limitations in the study reported in Chapter 4. First, the contribution disclosure forms used for the author contribution sections in published articles may not always yield a reliable picture of actual author contributions (Ilakovac et al., 2007). Second, we used the number of authors responsible for *all* analyses in the articles as a measure of co-piloting, while in the study reported in Chapter 3, we used the number of authors responsible for the *first or only study* reported in the article as a measure of co-piloting. The strategy in Chapter 4 allowed us to examine many more articles than our strategy in Chapter 3, but had the disadvantage that measuring the number of authors responsible for all analyses reported in a complete article could not tell us how many authors had been responsible for the analyses of each of the individual studies within one article.

Then there were a number of limitations that applied to both studies reported in Chapters 3 and 4. The discrepancy between the two methods of measuring the extent to which co-piloting had been employed suggests that we should be concerned about the validity of both our co-piloting measures. Neither of the measures may have truly captured what we consider co-piloting: that the statistical analyses are conducted independently by at least two co-authors, stipulating double-checks of the analyses and the reported results, open discussions on analytic decisions, and improved data documentation that facilitates later verification of the analytical results by (independent) peers (Wicherts, 2011). Given the many indicators for the presence of error and the potential for biases in statistical analyses, more research into how researchers work when analyzing their data is clearly warranted. In addition, statistical reporting errors may not be the type of errors that are reduced by the co-pilot model. The use of null hypothesis significance testing (NHST) itself is prone to error (e.g. Huxley, 1986) and misinterpretation (e.g. Hoekstra, Finch, Kiers, & Johnson, 2006), and it may only be serious methodological and analytical flaws that co-piloting (or collaboration with methodologists; Sijtsma (2015)) helps avoid. We focused on NHST because it is a dominant approach in psychology and other fields (Krueger, 2001), but it would be interesting to consider errors and biases with other inferential approaches such as use of confidence intervals (Cumming, 2014) or Bayesian statistics (e.g. Kruschke, 2015; Wagenmakers, 2007).

The high prevalence of statistical reporting errors suggests that the accuracy of published results cannot be taken for granted, and that further research studying solutions is warranted. Here, we propose several ways to deal with human error in (psychological) research and suggest how the effectiveness of the proposed practices might be studies in future research. First, we suggest leaving detection of statistical reporting error to statcheck (Epskamp & Nuijten, 2015), which can be applied prior to and during peer review. Since currently statcheck can only detect statistical results reported in APA format, we both recommend more widespread APA reporting across journals, and consistent use of APA reporting in each article rather than the use of uninformative reports such as ‘p < .05’ or merely the sole p-values. Another way to reduce human error in reporting of statistical results is to compose manuscripts using programs in which scripts can be written that automatically insert all statistical results into a text file, such as R Markdown. Third, we maintain that shared responsibility for the statistical analysis, as inherent to co-authorship, comprises the obligation to critically examine all results reported in a manuscript, the raw data files, and the analysis scripts before submission, to document and archive the data in such a way that independent peers can verify the results reported in the published article, and to archive the data in a manner that ensures that all co-authors retain access to the data after publication of the manuscript. This kind of workflow is facilitated by platforms such as the Open Science Framework (https://osf.io), and we highly recommend its use.

Fourth, we propose that in order to reduce the likelihood of methodological and statistical flaws, authors consult with a methodologist and/or statistician prior to conducting their study and analyzing the data (Sijtsma, 2015; Sijtsma et al., 2015). Finally, we recommend making as many of the analytical decisions as possible before the analyses are conducted or even before the study takes through the use of pre-registration (de Groot, 1956/2014; Wagenmakers et al., 2012). Thinking through all potential pitfalls in the design and the analyses and collaboratively formulating an analysis plan may help reduce methodological and analytical error through fostering open discourse among co-authors. In addition, pre-registration is meant to reduce bias due to opportunistic use of the many researcher degrees of freedom that researchers have in conducting psychological research, as was discussed in Chapters 5 and 6. The effectiveness of these potential best practices could be studied by comparing error rates in articles in which these practices were and were not used. Randomized experiments would be ideal, as they would preclude the possible alternative explanation that researchers who already employ these practices are also those researchers who tend to make less (or more) reporting errors.

Human bias in psychological science

Psychological data can often be analyzed in many different ways. The often arbitrary choices that researchers face in analyzing their data are called researcher degrees of freedom (Simmons et al., 2011). Researchers might be tempted to use these researcher degrees of freedom in an opportunistic manner in their pursuit of statistical significance (often called *p*-hacking), which is understandable from the perspective that scientists are prone to human cognitive biases. However, opportunistic use of researcher degrees of freedom leads to biased inferences, dissemination of results that potentially constitute false positives, and inflated effect size estimates of genuine effects. Together, this lowers not only the reproducibility of analytical results, but also the likelihood that results can be replicated in new samples of the same population. Therefore, it is important to deal with researcher degrees of freedom effectively.

In Chapter 5, we created a list of researcher degrees of freedom that psychological scientists have in formulating their hypotheses, designing their experiment, collecting their data, analyzing their results, and reporting their results. The list is not exhaustive and was compiled with the use of NHST in experiments in mind, but several of the researcher degrees of freedom are applicable to other inferential techniques and designs. Future work pertaining to our checklist might include adaptions for studies using other frameworks and designs, and the development of similar lists for use in different disciplinary fields. In its current shape it can be used to assess the level of potential bias in published research and as a tool in research methods education. Its most interesting use however is to apply it as a checklist to evaluate whether pre-registrations are sufficiently specific, precise, and exhaustive to effectively restrict opportunistic use of research degrees of freedom. The list may be used by authors of pre-registration, reviewers of both pre-registrations as well as registered reports.

In Chapter 6, we assessed the extent to which current pre-registrations restricted opportunistic use of the researcher degrees of freedom on the list presented in Chapter 5. Specifically, we employed a scoring protocol to compare two types of pre-registration formats currently available for psychological research that involve different levels of guidelines and requirements. As expected (and following our own pre-registration), we found that pre-registrations that were written following an elaborate set of guidelines and requirements restricted opportunistic use of researcher degrees of freedom considerably better than basic pre-registrations that were written following a limited set of guidelines and requirements. However, neither of the types of pre-registrations were sufficiently specific, precise, and exhaustive to deal with all researcher degrees of freedom listed in Chapter 5. This led us to conclude that better instructions, specific questions, and stricter requirements are necessary in order for pre-registrations to do what they are supposed to do, namely protect authors from their own biases in confirmatory hypothesis testing.

We concede that pre-registration is no easy task, and that not all issues in a study can be foreseen. We therefore argue that deviations from the pre-registered approach should not be forbidden, as long as they are transparently reported. Deviations can be judged by peer reviewers and readers on the degree to which they are defensible and could have created potential biases. Although strictly speaking, any deviation from the pre-registration yields an exploration, the degree to which such an exploration is data driven (thereby creating potential bias) should be discussed openly. For instance, in Chapter 4, we described how we deviated from our pre-registered analysis plan after realizing that one of the dependent variables was very difficult to measure. However, we believe that as long as such unforeseen issues are transparently reported, they serve to increase understanding of the conducted research and the reported results, to guide future research, and to illustrate that studies are not always as organized and clean as many research articles (unjustifiably) appear to suggest (Giner-Sorolla, 2012; Nosek & Bar-Anan, 2012).

Potential limitations of the study reported in Chapter 6 include that our findings do not provide causal evidence on whether pre-registering studies restrict the opportunistic use of researcher degrees of freedom. Direct evidence might be obtained if the same studies are randomly assigned to both a ‘pre-registration’ condition and a ‘no pre-registration’ condition, and the data are analyzed and reported by two different teams. Additionally, the protocol used to evaluate the pre-registration was rather strict. In addition, our protocol turned out not to be completely suitable for statistical frameworks other than NHST, and for studies using existing data. In the near future, we will improve our protocol so that it can be used in further evaluations of pre-registrations, and for designing pre-registered studies and writing registered reports. For example, other types of pre-registration formats, such as registered reports (Chambers, 2013) might be evaluated using our improved protocol, and authors of pre-registrations themselves might benefit from using our protocol as a checklist guiding the creation of their pre-registration. Finally, we hope to be able to examine the future articles thast will report the eventual pre-registered studies in our sample, and compare these to the pre-registrations to see whether any deviations from the pre-registrations have occurred.

Dealing with our human fallibility

As an anonymous science Nobel Prize Laureate who participated in one of the studies in Chapter 2 proclaimed, “Scientists are human, and so sometimes do not behave as they should as scientists.”. By demonstrating how (psychological) science is prone to error and bias and by studying best practices I hope to help researchers break existing taboos on errors and bias that hamper open discussion about the way we (should) conduct our research. An honest reflection on research practices requires us to accept the fact that we as scientists are human and not immune to human fallibility. We can only deal with our fallibility and strengthen science effectively if we acknowledge the role of human factors and (continue to) study ways to reduce errors and bias in research.

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), e0172792.

American Psychological Association. (2002). American Psychological Association ethical principles of psychologists and code of conduct. Retrieved from http://www.apa.org/ethics/code/index.aspx?item=11

American Psychological Association. (2010). *Publication Manual of the American Psychological Association. Sixth Edition*. Washington, DC: American Psychological Association.

Anderson, M. S., Horn, A. S., Risbey, K. R., Ronning, E. A., De Vries, R., & Martinson, B. C. (2007). What do mentoring and training in the responsible conduct of research have to do with scientists' misbehavior? Findings from a national survey of NIH-funded scientists. *Academic Medicine, 82*(9), 853-860.

Anderson, M. S., Martinson, B. C., & De Vries, R. (2007). Normative dissonance in science: Results from a national survey of U.S. scientists. *Journal of Empirical Research on Human Research Ethics, 2*(4), 3-14.

Anderson, M. S., Ronning, E. A., De Vries, R., & Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics, 13*(4), 437-461.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Nosek, B. A. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*(2), 108-119.

Bacon, F. (1621/2000). *Novum Organum* (L. Jardine & M. Silverthorne, Trans.). Cambridge: Cambridge University Press.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452-454.

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' Intuitions About Power in Psychological Research. *Psychological Science*.

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.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*, 666-678.

Bakker, M., & Wicherts, J. M. (2014a). Outlier Removal and the Relation with Reporting Errors and Quality of Psychological Research. *PLoS One, 9*, e103360.

Bakker, M., & Wicherts, J. M. (2014b). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods, 19*(3), 409.

Balon, R. (2005). By whom and how is the quality of research data collection assured and checked? *Psychotherapy and psychosomatics, 74*(6), 331-335.

Bargh, J. A., & Chartrand, T. L. (2000). Studying the mind in the middle; A practical guide to priming and automaticity research. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 253-285). Cambridge, UK: Cambridge University Press.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. Chichester: John Wiley & Sons.

Basalla, G. (1976). Pop science: the depiction of science in popular culture. In G. Holton & W. Blanpied (Eds.), *Science and its public*. Dordrecht, the Netherlands: D. Reidel.

Bates, T., Anić, A., Marušić, M., & Marušić, A. (2004). Authorship criteria and disclosure of contributions: comparison of 3 general medical journals with different author contribution forms. *JAMA, 292*(1), 86-88.

Beardslee, D. C., & O'dowd, D. D. (1961). The college-student image of the scientist. *Science, 133*(3457), 997-1001.

Beaty, D. (2004). *The naked pilot. The human factor in aircraft accidents*. Ramsbury, England: The Crowood Press.

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science. *Circulation research, 116*(1), 116-126.

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.

Bettencourt, B., Charlton, K., Dorr, N., & Hume, D. L. (2001). Status differences and in-group bias: a meta-analytic examination of the effects of status stability, status legitimacy, and group permeability. *Psychological bulletin, 127*(4), 520.

Bouter, L. M. (2015). Commentary: Perverse incentives or rotten apples? *Accountability in research, 22*(3), 148-161.

Brown, R., & Smith, A. (1989). Perceptions of and by minority groups: The case of women in academia. *European Journal of Social Psychology, 19*(1), 61-75.

Buyse, M., George, S. L., Evans, S., Geller, N. L., Ranstam, J., Scherrer, B., . . . Hutton, J. (1999). The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Statistics in medicine, 18*(24), 3435-3451.

Caperos, J. M., & Pardo, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema, 25*(3), 408-414.

Carlisle, J. (2012). The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia, 67*(5), 521-537.

Ceci, S. J. (1988). Scientists Attitudes toward Data Sharing. *Science Technology & Human Values, 13*(1-2), 45-52.

Ceci, S. J., & Walker, E. (1983). Private archives and public needs. *American Psychologist, 38*(4), 414-423.

Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences, 108*(8), 3157-3162.

Chamberlain, S., Boettiger, C., & Ram, K. (2015). rplos: Interface to the Search "API" for "PLoS" Journals Retrieved from http://CRAN.R-project.org/package=rplos

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

Chambers, C. D. (2015). Ten reasons why journals must review manuscripts before results are known. *Addiction, 110*(1), 10-11.

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., & Munafo, M. R. (2013, 5 June, 2013). Trust in science would be improved by study pre-registration. *The Guardian*. Retrieved from https://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration

Chambers, D. W. (1983). Stereotypic images of the scientist: The Draw-a-Scientist Test. *Science Education, 67*(2), 255-265.

Chan, A.-W. (2008). Bias, spin, and misreporting: time for full access to trial protocols and results. *PLoS Med, 5*(11), e230.

Chan, A.-W., & Altman, D. G. (2005). Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ, 330*(7494), 753.

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.

Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA, 291*(20), 2457-2465.

Chan, A.-W., Hróbjartsson, A., Jørgensen, K. J., Gøtzsche, P. C., & Altman, D. G. (2008). Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *BMJ, 337*, a2299.

Chang, A. C., & Li, P. (2015) Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say'Usually Not'. *Finance and Economics Discussion Series 2015- 083*. Washington, D.C.: Board of Governors of the Federal Reserve System.

Character traits: Scientific virtue. (2016). *Nature, 532*(7597), 139. Retrieved from http://www.nature.com/nature/journal/v532/n7597/full/nj7597-139a.html.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological bulletin, 114*(3), 494.

Cliff, N. (1996). *Ordinal Methods for Behavior Data Analysis*. Hillsdale, NJ: Erlbaum.

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

Cohen, J. (1990). Things I have learned (thus far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1994). The earth is round (P less-than.05). *American Psychologist, 49*(12), 997-1003.

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.

Cress, C. M., & Hart, J. (2009). Playing soccer on the football field: The persistence of gender inequities for women faculty. *Equity & Excellence in Education, 42*(4), 473-488.

Cronin, B., Shaw, D., & La Barre, K. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology, 54*(9), 855-871.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*: New York, NY: Routledge.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science, 25*(1), 7-29.

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 L. J. van der Maas]. *Acta Psychologica, 148*, 188-194.

DeCoster, J., Sparks, E. A., Sparks, J. C., Sparks, G. G., & Sparks, C. W. (2015). Opportunistic biases: Their origins, effects, and an integrated solution. *American Psychologist, 70*(6), 499.

Diekmann, A. (2007). Not the First Digit! Using Benford's Law to Detect Fraudulent Scientif ic Data. *Journal of Applied Statistics, 34*(3), 321-329.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., . . . Gamble, C. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One, 3*(8), e3081.

Eich, E. (2014). Business Not as Usual. *Psychological Science, 25*(1), 3-6.

Epskamp, S., & Nuijten, M. B. (2013). stacheck: Extract statistics from articles and recompute p-values. R package version 0.1.0. Retrieved from https://github.com/MicheleNuijten/statcheck

Epskamp, S., & Nuijten, M. B. (2015). R package “statcheck”: Extract statistics from articles and recompute values. Retrieved from https://github.com/MicheleNuijten/statcheck

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One, 4*(5), e5738.

Fanelli, D. (2010a). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PLoS One, 5*(4), e10271.

Fanelli, D. (2010b). "Positive" Results Increase Down the Hierarchy of the Sciences. *PLoS One, 5*(3), e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*(3), 891-904.

Fang, F. C., Bennett, J. W., & Casadevall, A. (2013). Males are overrepresented among life science researchers committing scientific misconduct. *MBio, 4*(1), e00640-00612.

Feist, G. J. (1998). Psychology of Science as a New Subdiscipline in Psychology. *Current Directions in Psychological Science, 20*(5), 330-334.

Fiedler, K., & Schwarz, N. (2015). Questionable research practices revisited. *Social Psychological and Personality Science, 7*(1), 45-52.

Fort, D. C., & Varney, H. L. (1989). How students see scientists: Mostly male, mostly white, and mostly benevolent. *Science and Children, 26*(8), 8-13.

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.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology, 57*(5), 153-169.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505.

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.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the Dependability of Research in Personality and Social Psychology. *Personality and Social Psychology Review, 18*(1), 3-12.

Garcia-Berthou, E., & Alcaraz, C. (2004). Incongruence between test statistics and P values in medical papers. *BMC Medical Research Methodology, 4*, 13.

Gauchat, G. (2012). Politicization of science in the public sphere a study of public trust in the United States, 1974 to 2010. *American Sociological Review, 77*(2), 167-187.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Department of Statistics. Columbia University. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460.

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. *BMJ: British Medical Journal, 327*(7417), 741.

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.

Goldacre, B. (2016). Make journals report clinical trials properly. *Nature, 530*(7588), 7-7.

Gøtzsche, P. C., Hróbjartsson, A., Marić, K., & Tendal, B. (2007). Data extraction errors in meta-analyses that use standardized mean differences. *JAMA, 298*(4), 430-437.

Hartgerink, C. H., Wicherts, J., & van Assen, M. (2017). Too good to be false: Nonsignificant results revisited. *Collabra: Psychology, 3*(1).

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2015). Distributions of p-values smaller than. 05 in Psychology: What is going on? *PeerJ PrePrints, 4*, e1642v1641.

Hassard, J. (1990). *Science experiences: Cooperative learning and the teaching of science*. Menlo Park, CA: Addison-Wesley.

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. (2015). *Corrupt research: the case for reconceptualizing empirical management and social science*. Thousand Oaks, CA: SAGE Publications.

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.

Huxley, P. (1986). Statistical errors in papers in the British-Journal-of-Social-Work - (Volumes 1-14). *British Journal of Social Work, 16*(6), 645-658.

Ilakovac, V., Fister, K., Marusic, M., & Marusic, A. (2007). Reliability of disclosure forms of authors' contributions. *Canadian Medical Association Journal, 176*(1), 41-46.

Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *JAMA, 294*(2), 218-228.

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

Ioannidis, J. P. A. (2007). Non-replication and inconsistency in the genome-wide association setting. *Human heredity, 64*(4), 203-213.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640-648.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7*(6), 645-654.

Ioannidis, J. P. A. (2014). How to Make More Published Research True. *PLoS Med, 11*(10), e1001747.

Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: evaluation and improvement of research methods and practices. *PLoS Biol, 13*(10), e1002264.

Ioannidis, J. P. A., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in cognitive sciences, 18*(5), 235-241.

Ipsos MORI. (2014). *Public attitudes to science 2014*. Retrieved from: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf

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.

Jonas, K. J., & Cesario, J. (2015). How can preregistration contribute to research in our field? *Comprehensive Results in Social Psychology*, 1-7.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 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 Biol, 14*(5), e1002456.

Kirkham, J. J., Dwan, K. M., Altman, D. G., Gamble, C., Dodd, S., Smyth, R., & Williamson, P. R. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ, 340*, c365.

Kornfeld, D. S. (2012). Perspective: Research misconduct: The search for a remedy. *Academic Medicine, 87*(7), 877-882.

Kriegeskorte, N., Lindquist, M. A., Nichols, T. E., Poldrack, R. A., & Vul, E. (2010). Everything you never wanted to know about circular analysis, but were afraid to ask. *Journal of Cerebral Blood Flow and Metabolism, 30*(9), 1551-1557.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience, 12*, 535-540.

Krueger, J. (2001). Null hypothesis significance testing - On the survival of a flawed method. *American Psychologist, 56*(1), 16-26.

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. London, UK: Academic Press.

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.

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*, 1-16.

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.

Levine, T. R., Weber, R., Hullet, C., Park, H. S., & Lindsey, L. L. M. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research, 34*(2), 171-U175.

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science, 26*(12), 1827-1832.

Lindvall, M., Muthig, D., Dagnino, A., Wallin, C., Stupperich, M., Kiefer, D., . . . Kahkonen, T. (2004). Agile software development in large organizations. *Computer, 37*(12), 26-34.

Macbeth, G., Razumiejczyk, E., & Ledesma, R. D. (2011). Cliff's Delta Calculator: A non-parametric effect size program for two groups of observations. *Universitas Psychologica, 10*(2), 545-555.

MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature, 526*(7572), 187-189.

Mahoney, M. J. (1976). *Scientist as Subject: The Psychological Imperative*. Cambridge, MA, US: Ballinger Publishing Company.

Mahoney, M. J. (1979). Psychology of the Scientist- Evaluative Review. *Social Studies of Science, 9*(3), 349-375.

Mahoney, M. J. (1985). Open exchange and epistemic progress. *American Psychologist, 40*, 29-39.

Mahoney, M. J., & DeMonbreun, B. G. (1977). Psychology of the scientist: An analysis of probem-solving bias. *Cognitive Therapy and Research, 1*, 229-238.

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.

Marsman, M., Schönbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A., & Wagenmakers, E. J. (2017). A Bayesian bird's eye view of ‘Replications of important results in social psychology’. *Royal Society Open Science, 4*(1), 160426.

Martinson, B. C., Anderson, M. S., Crain, A. L., & De Vries, R. (2006). Scientists' perceptions of organizational justice and self-reported misbehaviors. *Journal of Empirical Research on Human Research Ethics, 1*(1), 51-66.

Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). Scientists behaving badly. *Nature, 435*(7043), 737-738.

Marušić, A., Bates, T., Anić, A., & Marušić, M. (2006). How the structure of contribution disclosure statements affects validity of authorship: a randomized study in a general medical journal. *Current medical research and opinion, 22*(6), 1035-1044.

Marusic, A., Wager, E., Utrobicic, A., Rothstein, H. R., & Sambunjak, D. (2015). Interventions to prevent misconduct and promote integrity in research and publication. *The Cochrane Library*.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147-163.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist, 70*(6), 487.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., . . . Soderberg, C. K. (2016). How open science helps researchers succeed. *eLife, 5*, e16800.

Mead, M., & Metraux, R. (1957). Image of the scientist among high-school students. *Science, 126*(3270), 384-390.

Mendenhall, M., & Higbee, K. L. (1982). Psychology of the scientist: XLVIII. Recent trends in multiple authorship in psychology. *Psychological Reports, 51*(3), 1019-1022.

Merton, R. K. (1942). A Note on Science and Democracy. *Journal of Legal and Political Sociology, 1*, 115.

Miller, D. I., Eagly, A. H., & Linn, M. C. (2014). Women’s Representation in Science Predicts National Gender-Science Stereotypes: Evidence From 66 Nations. *Journal of Educational Psychology, 107*(3), 631-644.

Mitroff, I. I. (1974). *The subjective side of science. A philosophical inquiry into the psychology of the Apollo moon scientists*. Amsterdam, The Netherlands: Elsevier Scientific Publishing Company.

Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M., & Zwelling, L. (2013). A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. *PLoS One, 8*(5), e63221.

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., . . . Schönbrodt, F. D. (2016). The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review. *Royal Society Open Science, 3*(1), 150547.

Mosimann, J. E., Dahlberg, J., Davidian, N., & Krueger, J. (2002). Terminal digits and the examination of questioned data. *Accountability in Research: Policies and Quality Assurance, 9*(2), 75-92.

Mosimann, J. E., Wiseman, C. V., & Edelman, R. E. (1995). Data fabrication: Can people generate random digits? *Accountability in research, 4*(1), 31-55.

Motyl, M., Demos, A., Carsel, T., Hanson, B., Melton, Z., Mueller, A., . . . Wong, K. (2017). The State of Social and Personality Science: Rotten to the Core, Not So Bad, Getting Better, or Getting Worse? *Journal of Personality and Social Psychology, 113*(1), 34.

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.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *The Quarterly Journal of Experimental Psychology, 29*(1), 85-95.

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2017). Forthcoming in JPSP: A Non-Diagnostic Audit of Psychological Research. Retrieved from http://datacolada.org/60

Neuliep, J. W., & Crandall, R. (1993a). Everyone was wrong: There are lots of replications out there. *Journal of Social Behavior and Personality, 8*(6), 1.

Neuliep, J. W., & Crandall, R. (1993b). Reviewer bias against replication research. *Journal of Social Behavior and Personality, 8*(6), 21.

Newton, D. P., & Newton, L. D. (1992). Young children's perceptions of science and the scientist. *International Journal of Science Education, 14*(3), 331-348.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience, 14*, 1105-1107.

Nobelprize.org. (2014). Nobel Prizes and Laureates. Retrieved from http://www.nobelprize.org/nobel_prizes/lists/all/

Nosek, B. A., Alter, G., Banks, G., Borsboom, D., Bowman, S., Breckler, S., . . . Christensen, G. (2015). Promoting an open research culture: author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science (New York, NY), 348*(6242), 1422.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific Utopia: I. Opening Scientific Communication. *Psychological Inquiry, 23*(3), 217-243.

Nosek, B. A., & Lakens, D. (2015). Registered reports. *Social Psychology, 45*, 137-141.

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.

Nuijten, M. B. (2017). *Open Data and Reporting Inconsistencies Preregistration Study 3*. Retrieved from https://osf.io/538bc/

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*, 1-22.

Nuzzo, R. (2015). How scientists fool themselves-and how they can stop. *Nature, 526*(7572), 182-185.

ó Maoldomhnaigh, M., & Hunt, Á. (1988). Some factors affecting the image of the scientist drawn by older primary school pupils. *Research in Science & Technological Education, 6*(2), 159-166.

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2014). The Chrysalis Effect How Ugly Initial Results Metamorphosize Into Beautiful Articles. *Journal of Management*, *43*(2), 376-399.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867-872.

Over, R. (1982). Collaborative research and publication in psychology. *American Psychologist, 37*(9), 996.

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.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafo, M. R., . . . Yarkoni, T. (2016). Scanning the Horizon: Future challenges for neuroimaging research. Retrieved from http://dx.doi.org/10.1101/059188

Qualtrics. (2012). Qualtrics (Version 500235). Provo, Utah, USA.

Qualtrics. (2014). Qualtrics. Provo, Utah, USA.

R Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage Publications.

Reason, J. (1990). *Human error*. Cambridge, UK: Cambridge University Press.

Romano, J., Kromrey, J. D., Coraggio, J., & Skowronek, J. (2006). *Appropriate statistics for ordinal level data: Should we really be using t-test and cohen's d for evaluating group differences on the NSSE and other surveys?* Paper presented at the Annual meeting of the Florida Association of Institutional Research, Chicago, Illinois.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. East-Norwalk, CT, USA: Appleton-Century-Crofts.

Sala I Martin, X. X. (1997). I just ran two million regressions. *American Economic Review, 87*(2), 178-183.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Schaller, M. (2016). The empirical benefits of conceptual rigor: Systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). *Journal of Experimental Social Psychology, 66*, 107-115.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological bulletin, 105*(2), 309-316.

Shamoo, A. E., & Resnik, D. B. (2015). *Responsible conduct of research* (3rd ed.). New York: Oxford University Press.

Shen, H. (2013). Mind the gender gap. *Nature, 495*(7439), 22-24.

Sijtsma, K. (2015). Playing with Data—Or How to Discourage Questionable Research Practices and Stimulate Researchers to Do Things Right. *Psychometrika*, 1-15.

Sijtsma, K., Veldkamp, C. L. S., & Wicherts, J. M. (2015). Improving the conduct and reporting of statistical analysis in psychology. *Psychometrika*, 1-6.

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.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after p-hacking.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at Perspectives on Psychological Science. *Perspectives on Psychological Science, 9*(5), 552-555.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). p-Curve and effect size: correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666-681.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534.

Smith, T., & Son, J. (2013). General Social Survey 2012 final report: Trends in public attitudes about confidence in institutions. *NORC at the University of Chicago, Chicago, IL*.

Spellman, B. A. (2015). A Short (Personal) Future History of Revolution 2.0. *Perspectives on Psychological Science, 10*(6), 886-899.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702-712.

Steneck, N. H. (2013). Global research integrity training. *Science, 340*(6132), 552-553.

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.

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.

Sugimoto, C. R. (2013). Global gender disparities in science. *Nature, 504*(7479), 211-213.

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. Cambridge, England: Cambridge University Press.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. Austin (Eds.), *Psychology of intergroup relations* (pp. 7-24). Chicago: Nelson-Hall.

Thomson Reuters. (2014). Web of Science™. Retrieved from https://webofknowledge.com.

Tijdink, J. K., Bouter, L. M., Veldkamp, C. L. S., van de Ven, P. M., Wicherts, J. M., & Smulders, Y. M. (2016). Personality Traits Are Associated with Research Misbehavior in Dutch Scientists: A Cross-Sectional Study. *PLoS One, 11*(9), e0163251.

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

Tijdink, J. K., Vergouwen, A. C. M., & Smulders, Y. M. (2013). Publication pressure and burn out among Dutch medical professors: a nationwide survey. *PLoS One, 8*(9), e73381.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Cambridge, MA, USA: Blackwell.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin, 76*(2), 105.

Ueno, T., Fastrich, G. M., & Murayama, K. (2016). Meta-analysis to integrate effect sizes within an article: Possible misuse and Type I error inflation. *Journal of Experimental Psychology: General, 145*, 643-654.

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., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20*(3), 293.

Vazire, S. (2015). Editorial. *Social Psychological and Personality Science, 7*(1), 3-7.

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology, 3*(1).

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.

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., . . . Rennison, D. J. (2013). The availability of research data declines rapidly with article age. *Current Biology, 24*(1), 94-97.

Vogeli, C., Yucel, R., Bendavid, E., Jones, L. M., Anderson, M. S., Louis, K. S., & Campbell, E. G. (2006). Data withholding and the next generation of scientists: Results of a national survey. *Academic Medicine, 81*(2), 128-136.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values *Psychonomic Bulletin & Review, 14*, 779-804.

Wagenmakers, E. J., & Dutilh, G. (2016). Seven selfish reasons for preregistration. *APS Observer, 29*(9).

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

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632-638.

Watson, D. L. (1938). *Scientists Are Human*. London: Watts.

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PLoS One, 8*(7), e66212.

Weston, J., Dwan, K., Altman, D., Clarke, M., Gamble, C., Schroter, S., . . . Kirkham, J. (2016). Feasibility study to examine discrepancy rates in prespecified and reported outcomes in articles submitted to The BMJ. *BMJ Open, 6*(4), e010075.

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.

Wicherts, J. M. (2013). Science revolves around the data. *Journal of Open Psychology Data, 1*(1), e1.

Wicherts, J. M. (2017). Data re-analysis and open data. In J. Plucker & M. Makel (Eds.), *Toward a More Perfect Psychology: Improving Trust, Accuracy, and Transparency in Research*. Wahington, DC: American Psychological Association.

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence*.

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.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*, 726-728.

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

Wiegman, D. A., & Shappell, S. A. (2003). *A human error approach to aviation accident analysis. The human factors analysis and classification system*. Aldershot, England: Ashgate Publishing.

Wigboldus, D. H., & Dotsch, R. (2016). Encourage Playing with Data and Discourage Questionable Reporting Practices. *Psychometrika, 81*(1), 27-32.

Wilcox, R. (2012). *Modern statistics for the social and behavioral sciences: A practical introduction*: CRC Press.

Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences, 112*(17), 5360-5365.

Zimmer, C. (April 16, 2012). A sharp rise in retractions prompts calls for reform. *The New York Times*.

Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary Issues in the Analysis of Data- a Survey of 551 Psychologists. *Psychological Science, 4*(1), 49-53.

APPENDICES

Appendix A:
Supplementary materials of Chapter 2

Appendix B:
Supplementary materials of Chapter 3

Appendix C:
Supplementary materials of Chapter 6

APPENDIX A: SUPPLEMENTARY MATERIALS OF CHAPTER 2

STATISTICAL ANALYSIS AND SAMPLE SIZE DETERMINATION

Study 1

We planned to conduct a total of six ANOVAs (one for each ideal scientist feature) and to therefore use a Bonferroni-corrected family-wise alpha of 0.0083333 in each analysis. A priori power computations* yielded a required total sample size of 531 respondents (n = 133 per group) to obtain a power of .80 to detect a small to medium effect (f = .175). To be on the safe side, we aimed for 150 people per condition.

Study 2

We planned to conduct two different sets of analyses: one where we compared the pooled means of the non-scientist professions to the means of the scientist profession (2 x 2 mixed design), and one where we compared the means of all different professions. We decided to only carry out the former set of analyses, but to include graphs with the means for the separate professions in the supplemental materials. The reasons for this decision were that the first set of analyses would yield more informative results with respect to our research question, and that the second set of analyses would have required a very large number of contrasted to be tested while these contrasts were not meaningful with respect to our research question itself. The nine non-scientists professions together formed a reliable scale on each of the six characteristics (Chonbach's alphas ranging from .81 to .88, see Table S3), providing support for the assumption that these professions together measure the construct 'highly-educated professions'. With our Bonferroni-corrected alpha of 0.0083333, we needed 124* respondents (n = 62 per group) to obtain a power of .80 to detect a small to moderate effect (f= .175). To be on the safe side, we aimed for 75 participants per group.

Study 3

We planned to conduct a total of six ANOVAs (one for each ideal scientist feature) and to therefore use a Bonferroni-corrected family-wise alpha of 0.0083333 in each analysis. A priori power computations[1]* yielded a required total sample size of 762 respondents (n = 85 per group) to obtain a power of .80 to detect a small to medium effect (f = .175).

Study 4

We planned to conduct a total of six ANOVAs (one for each ideal scientist feature) and to therefore use a Bonferroni-corrected family-wise alpha of 0.0083333 in each analysis. A priori power computations* yielded a required total sample size of 531 respondents (n = 133 per group) to obtain a power of .80 to detect a small to medium effect (f = .175).

* Power analysis was carried out in G*Power 1.3.6;

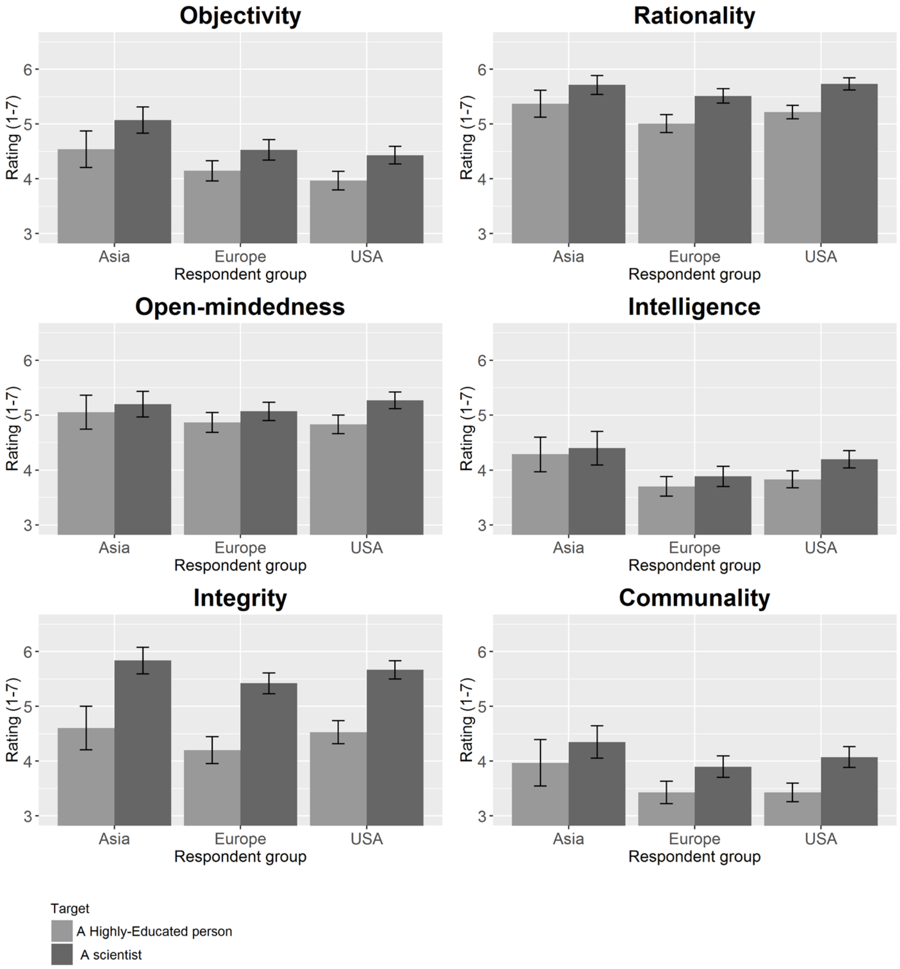
STUDY REGISTRATION AND OUTLIER HANDLING

We registered our studies at the Open Science Framework. The registered studies are described in this article in the following order: ‘Study A’ (= Study 1), ‘Study D’ (= Study 2), ‘Study B’ (= Study 3), ‘Study C’ (= Study 4).   The registration of this series of studies can be found through https://osf.io/z3xt6/.

In line with Bakker and Wicherts (Bakker & Wicherts, 2014b) and Tukey (1977) we regarded data-points that lie 2 Inter Quartile Ranges (IQR) outside the lower and upper quartiles as outliers. The scripts provided on the Open Science Framework can easily be adapted to conduct the analyses without the removal of any outliers or the removal of outliers that lie 1.5 Inter Quartile Ranges (IQR) outside the lower and upper quartiles.

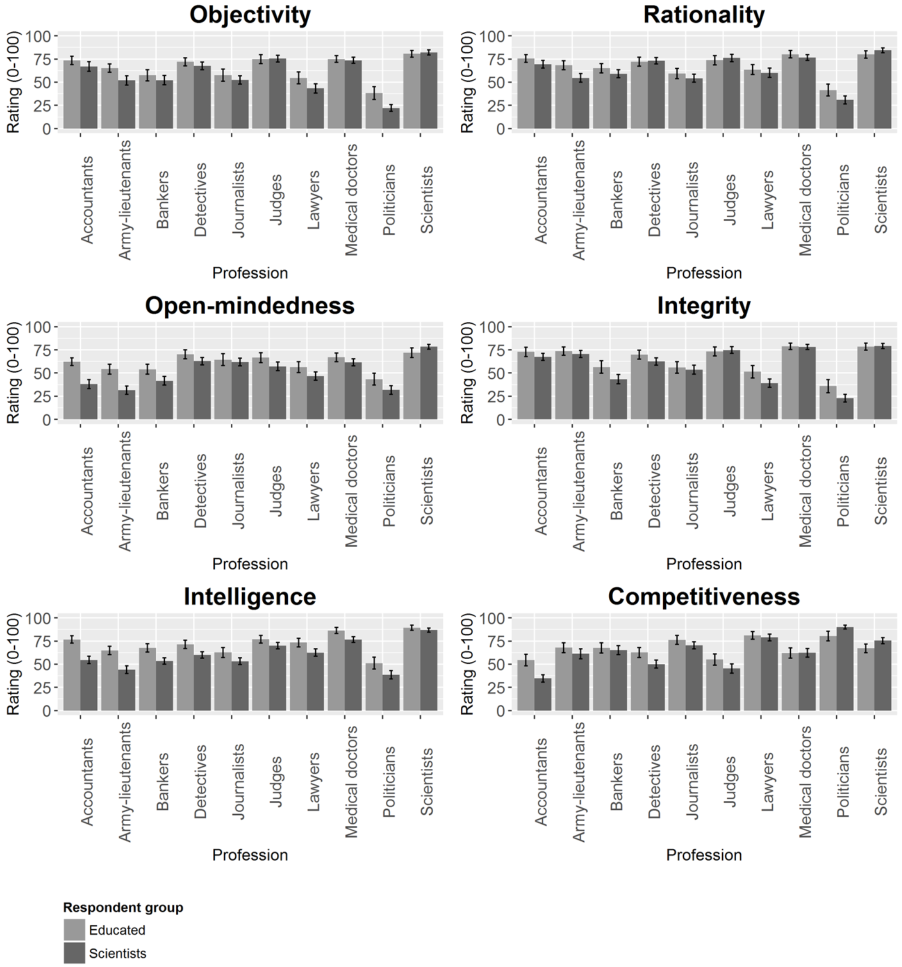
SUPPLEMENTARY FIGURES

Figure S1 *Attributions of Objectivity, Rationality, Open-mindedness, Intelligence, Integrity, and Communality to the typical highly-educated person versus the typical scientist by world part.*



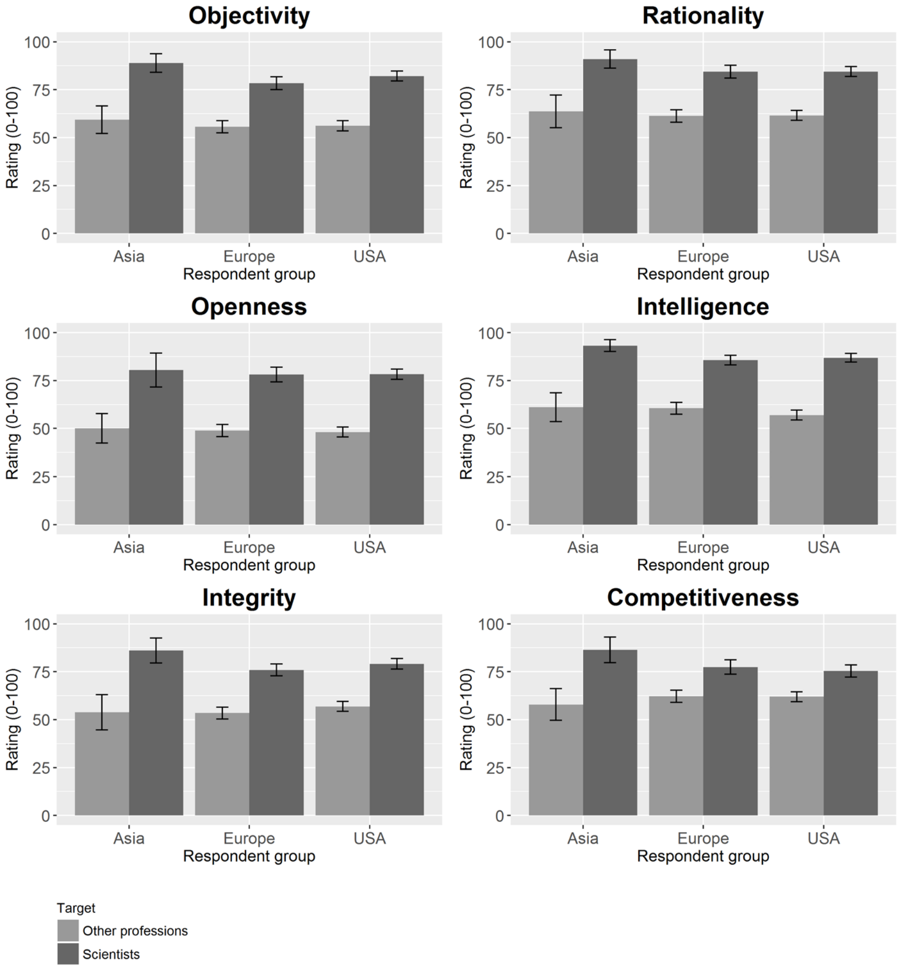
Note: Results are presented by respondent group: Asian scientists, European scientists, and American scientists.

Figure S2 *Attributions of Objectivity, Rationality, Open-mindedness, Intelligence, Integrity, and Communality to people with various professions.*



Note: Results are presented by respondent group.

Figure S3 *Attributions of Objectivity, Rationality, Open-mindedness, Intelligence, Integrity, and Communality to people with highly-educated profession versus people with the profession of scientist by world part.*



Note: Results are presented by respondent group: Asian scientists, European scientists, and American scientists.

SUPPLEMENTARY TABLES

Table S1 *Sample details Study 1.*

Respondent group	N	Mean Age (years)	SD Age (years)	Range Age (years)	Female (%)	Response rate (%)	Response rate after cleaning (%)
American Educated	312	49.2	13.8	23- 84	46	100*	99.37
American Scientists	331	49.0	11.4	26- 77	34	**	**
Asian Scientists	117	41.8	9.3	27- 66	17	**	**
European Scientists	304	43.6	10.5	26- 75	29	**	**
Total Scientists	752					10.58	8.50
Total	1064						
Nobel Prize Laureates***	34	75.3	12.7	45- 93	0	18.95	17.89

Note: *Qualtrics sample: paid survey panel members. **response rates cannot be computed for the world parts separately because we did not know scientists' location beforehand. The response rate is based on the total number of responses from scientists from all over the world, divided by the total number of e-mails sent to scientists from all over the world (for details see https://osf.io/3nepx/). Nobel Prize Laureates were not included in the analyses.

Table S2 *Scale reliabilities Study 1.*

Scale	Cronbach's alpha	95% CI
Objectivity	.73	.66 ; .81
Rationality	.76	.68 ; .83
Open-mindedness	.77	.70 ; .85
Intelligence	.73	.65 ; .81
Integrity	.87	.81 ; .93
Communality	.79	.72 ; .86

Note: Based on the data of the American Educated and American scientist respondents only. 95% CI = 95% confidence interval.

Table S3 *Correlation tables Study 1: correlations between the characteristics of the ideal scientist, by Target.*

A highly-educated person						
Feature	1	2	3	4	5	6
1. Objectivity	1					
2. Rationality	.57***	1				
3. Open-mindedness	.70***	.62***	1			
4. Intelligence	.41***	.40***	.34***	1		
5. Integrity	.63***	.52***	.60***	.40***	1	
6. Communality	.67***	.42***	.56***	.42***	.68***	1
A scientist						
1. Objectivity	1					
2. Rationality	.50***	1				
3. Open-mindedness	.67***	.60***	1			
4. Intelligence	.30***	.27***	.23***	1		
5. Integrity	.57***	.54***	.68***	.10	1	
6. Communality	.69***	.43***	.63***	.27***	.61***	1
Overall						
1. Objectivity	1					
2. Rationality	.56***	1				
3. Open-mindedness	.70***	.62***	1			
4. Intelligence	.39***	.38***	.31***	1		
5. Integrity	.63***	.58***	.64***	.31***	1	
6. Communality	.70***	.46***	.61***	.37***	.67***	1

Note: Based on the data of the American Educated and American scientist respondents only.*significant at α = .05, **significant at α = .01, ***significant at α = .001. All ***p***-values are adjusted for multiple testing.

Table S4 *Statistical analyses Study 1.*

Feature	Respondent group	Target	N	Mean	SD	Interaction	Main effect Target (main effects only model)	t-test main effect Target	Mean diff	Cohen's d [95% CI]	Main effect Resp. Group (main effects only model)	t-test main effect Resp. Group	Mean diff	Cohen's d [95% CI]
Objectivity	Scientists	Scientist	165	4.43	1.05	F(1, 639) = 0.55, p = .459	F(1, 640) = 37.83, p < .001	t(641) = 5.95, p < .001	0.52	0.47 [0.31 ; 0.63]	F(1, 640) = 34.33, p < .001	t(641) = -5.65, p < .001	-0.47	-0.45 [-0.60 ;-0.29]
		Educated	166	3.97	1.09									
	Educated	Scientist	153	4.99	1.14									
		Educated	159	4.40	1.04									
Rationality	Scientists	Scientist	165	5.73	0.71	F(1, 639) = 0.26, p = .613	F(1, 640) = 64.68, p < .001	t(641) = 8.04, p < .001	0.55	0.63 [0.48 ; 0.79]	F(1, 640) = 4.02, p = .045	t(641) = 1.97, p = .049	0.14	0.16 [0.00 ; 0.31]
		Educated	166	5.22	0.79									
	Educated	Scientist	153	5.63	0.88									
		Educated	159	5.05	1.05									
Openness	Scientists	Scientist	165	5.27	0.98	F(1, 639) = 0.58, p = .445	F(1, 640) = 19.28, p < .001	t(641) = 4.40, p < .001	0.38	0.35 [0.19 ; 0.50]	F(1, 640) = 0.89, p = .347	t(641) = 0.96, p = .335	0.08	0.08 [-0.08 ; 0.23]
		Educated	166	4.83	1.12									
	Educated	Scientist	153	5.12	1.23									
		Educated	159	4.81	1.00									
Intelligence	Scientists	Scientist	165	4.19	1.02	F(1, 639) = 2.99, p = .084	F(1, 640) = 32.93, p < .001	t(641) = 5.61, p < .001	0.51	0.44 [0.29 ; 0.60]	F(1, 640) = 22.53, p < .001	t(641) = -4.59, p < .001	-0.42	-0.36 [-0.52 ;-0.21]
		Educated	166	3.83	1.02									
	Educated	Scientist	153	4.77	1.27									
		Educated	159	4.10	1.19									
Integrity	Scientists	Scientist	165	5.66	1.07	F(1, 639) = 2.84, p = .092	F(1, 640) = 96.57, p < .001	t(641) = 9.79, p < .001	0.98	0.77 [0.61 ; 0.93]	F(1, 640) = 9.21, p = .003	t(641) = 2.91, p = .004	0.31	0.23 [0.07 ; 0.38]
		Educated	166	4.53	1.37									
	Educated	Scientist	153	5.19	1.28									
		Educated	159	4.39	1.29									
Communality	Scientists	Scientist	165	4.07	1.23	F(1, 639) = 0.30, p = .582	F(1, 640) = 39.27, p < .001	t(641) = 6.05, p < .001	0.59	0.48 [0.32 ; 0.63]	F(1, 640) = 37.55, p < .001	t(641) = -5.90, p < .001	-0.58	-0.47 [-0.62 ;-0.31]
		Educated	166	3.43	1.10									
	Educated	Scientist	153	4.60	1.29									
		Educated	159	4.06	1.20									

Note: Based on data of American educated and American scientist respondents only. For interactions and main effects, α = .008333; for subsequent tests of simple effects, α = 0.05. Text in grey represents non-significant results.

Table S5 *Sample details Study 2.*

Respondent group	N	Mean Age (years)	SD Age (years)	Range Age (years)	Female (%)	Response rate (%)	Response rate after cleaning (%)
American Educated	75	46.3	14.7	22-83	47%	100*	75.70*
American Scientists	111	49.9	12.4	27-85	20%	**	**
Asian Scientists	20	45.5	12.3	26- 69	15%	**	**
European Scientists	67	44.6	10.6	28- 75	25%	**	**
Total Scientists	198					10.97	6.76
Total	273						

Note: *Qualtrics sample: paid survey panel members. **response rates cannot be computed for the world parts separately because we did not know scientists' location beforehand. The response rate is based on the total number of responses from scientists from all over the world, divided by the total number of e-mails sent to scientists from all over the world (for details see https://osf.io/3nepx/).

Table S6 *Scale reliabilities Study 2.*

Scale	Cronbach's alpha	95% CI
Objectivity	.81	.74 ; .87
Rationality	.83	.77 ; .89
Open-mindedness	.83	.77 ; .89
Intelligence	.88	.83 ; .93
Integrity	.84	.78 ; .90
Competitiveness	.81	.74 ; .87

Note: Based on the data of the American Educated and American scientist respondents only. 95% CI = 95% confidence interval.

Table S7 *Correlation tables Study 2: correlations between the characteristics of the ideal scientist, by profession category.*

Highly-educated professions						
Feature	1	2	3	4	5	6
1. Objectivity	1					
2. Rationality	.73***	1				
3. Open-mindedness	.75***	.68***	1			
4. Intelligence	.75***	.72***	.74***	1		
5. Integrity	.71***	.71***	.68***	.72***	1	
6. Communality	.48***	.49***	.49***	.54***	.38***	1
Profession of scientist						
1. Objectivity	1					
2. Rationality	.60***	1				
3. Open-mindedness	.47***	.39***	1			
4. Intelligence	.48***	.47***	.21*	1		
5. Integrity	.57***	.49***	.33***	.46***	1	
6. Communality	.10	.12	.28***	.13	.09	1

Note: Based on the data of the American Educated and American scientist respondents only.*significant at α = .05, **significant at α = .01, ***significant at α = .001. All *p*-values are adjusted for multiple testing.

Table S8 *Statistical analyses Study 2.*

Feature	Respondent group	Target	N	Mean	SD	Interaction	Simple effects	Mean diff.	Correlation	Cohen's d [95% CI]	Diff. d
Objectivity	Scientists	Scientist	111	82.14	14.02	t(184) = 3.61, p < .001	t(110) = 18.50, p < .001	25.97	.87	1.76 [1.57 ; 1.94]	0.73
		Educated	111	56.18	14.11						
	Educated	Scientist	75	80.64	16.05		t(74) = 8.87, p < .001	17.48	.72	1.02 [0.79 ; 1.25]	
		Educated	75	63.16	15.22						
Rationality	Scientists	Scientist	111	84.48	13.77	t(184) = 4.04, p < .001	t(110) = 15.80, p < .001	22.91	.83	1.50 [1.31 ; 1.69]	0.71
		Educated	111	61.57	13.95						
	Educated	Scientist	75	79.89	16.97		t(74) = 6.83, p < .001	13.30	.62	0.79 [0.56 ; 1.02]	
		Educated	75	66.60	14.80						
Open-mindedness	Scientists	Scientist	111	78.31	14.39	t(184) = 6.62, p < .001	t(110) = 17.99, p < .001	30.17	.86	1.71 [1.52 ; 1.90]	1.08
		Educated	111	48.14	13.83						
	Educated	Scientist	75	71.85	22.58		t(74) = 5.43, p < .001	12.05	.53	0.63 [0.40 ; 0.86]	
		Educated	75	59.80	16.08						
Intelligence	Scientists	Scientist	111	86.87	11.83	t(184) = 4.80, p < .001	t(110) = 19.82, p < .001	29.87	.88	1.88 [1.69 ; 2.07]	0.45
		Educated	111	57.00	13.46						
	Educated	Scientist	75	89.25	12.05		t(74) = 12.43, p < .001	19.17	.82	1.44 [1.21 ; 1.67]	
		Educated	75	70.08	13.97						
Integrity	Scientists	Scientist	111	79.09	14.73	t(184) = 2.89, p = .004	t(110) = 15.86, p < .001	22.18	.83	1.51 [1.32 ; 1.69]	0.64
		Educated	111	56.91	13.46						
	Educated	Scientist	75	78.44	16.21		t(74) = 7.53, p < .001	15.29	.66	0.87 [0.64 ; 1.10]	
		Educated	75	63.15	17.28						
Competitiveness	Scientists	Scientist	111	75.39	16.85	t(184) = 5.44, p < .001	t(110) = 7.87, p < .001	13.4	.60	0.75 [0.56 ; 0.93]	0.77
		Educated	111	61.99	13.78						
	Educated	Scientist	75	67.08	20.74		t(74) =-.22, p = .82	-0.40	.03	-0.03 [-0.26 ; 0.20]	
		Educated	75	67.48	15.47						

Note: Based on data of American educated and American scientist respondents only. For interactions and main effects, α = .008333; for subsequent tests of simple effects, α = 0.05. Text in grey represents non-significant results.

Table S9 *Sample details Study 3.*

Respondent group	N	Mean Age (years)	SD Age (years)	Range Age (years)	Female (%)	Response rate (%)	Response rate after cleaning (%)
Early-career scientists	515	35.2	5.8	26- 94[1]	33%	*	*
Established scientists	903	51.9	9.2	35- 90	22%	*	*
Total	1418					10.55	5.97

Note: [1]Probably erroneous maximum age: one person selected the first answer option on the list, which translates to age = 94. *Response rates cannot be computed for the two respondent groups separately because we did not know scientists' career level beforehand. The response rate is based on the total number of responses divided by the total number of e-mails sent (for details see https://osf.io/3nepx/).

Table S10 *Scale reliabilities Study 3.*

Scale	Cronbach's alpha	95% CI
Objectivity	.63	.57 ; .69
Rationality	.74	.69 ; .79
Open-mindedness	.67	.61 ; .73
Intelligence	.70	.65 ; .75
Integrity	.82	.77 ; .86
Communality	.63	.57 ; .69

Note: 95% CI = 95% confidence interval.

Table S11 *Correlation tables Study 3: correlations between the characteristics of the ideal scientist, by respondent group.*

Early-career scientists						
Feature	1	2	3	4	5	6
1. Objectivity	1					
2. Rationality	.52***	1				
3. Open-mindedness	.61***	.49***	1			
4. Intelligence	.24***	.31***	.16***	1		
5. Integrity	.57***	.53***	.58***	.24***	1	
6. Communality	.54***	.29***	.46***	.18***	.55***	1
Established scientists						
1. Objectivity	1					
2. Rationality	.52***	1				
3. Open-mindedness	.61***	.57***	1			
4. Intelligence	.35***	.44***	.29***	1		
5. Integrity	.54***	.54***	.58***	.21***	1	
6. Communality	.55***	.36***	.49***	.25***	.53***	1
Overall						
1. Objectivity	1					
2. Rationality	.52***	1				
3. Open-mindedness	.61***	.54***	1			
4. Intelligence	.32***	.39***	.25***	1		
5. Integrity	.56***	.54***	.58***	.23***	1	
6. Communality	.55***	.33***	.48***	.23***	.55***	1

Note: *significant at α = .05, **significant at α = .01, ***significant at α = .001. All ***p***-values are adjusted for multiple testing.

Table S12 *Statistical analyses Study 3.*

Feature	Respondent group	Target	N	Mean	SD	Interaction	Simple effects: effect of Target in each respondent group separately	Comparisons	t-tests	Mean diff.	Cohen's D [95% CI]
Objectivity	Early-career	Early-career	179	4.35	0.95	F(2, 1412) = 6.50, p = .002 *	F(2, 512) = 3.83, p = .022	Established- Early	t(344) = -0.06, p = .953	-0.01	-0.01 [-0.22 ; 0.21]
		Established	167	4.34	1.20			Early- PhD	t(346) = 2.65, p = .008	0.28	0.28 [0.07 ; 0.50]
		PhD-students	169	4.07	0.98			Established- PhD	t(334) = 2.25, p = .025	0.27	0.25 [0.03 ; 0.46]
	Established	Early-career	290	4.27	0.99		F(2, 900) = 14.45, p < .001	Established- Early	t(604) = 5.06, p < .001	0.42	0.41 [0.25 ; 0.57]
		Established	316	4.70	1.07			Early- PhD	t(585) =-1.71, p = .088	-0.14	-0.17 [-0.30 ; 0.02]
		PhD-students	297	4.41	0.92			Established- PhD	t(611) = 3.57, p < .001	0.29	0.30 [0.13 ; 0.45]
Rationality	Early-career	Early-career	179	5.25	0.92	F(2, 1412) = 5.07, p = .006	F(2, 512) = 21.04, p < .001	Established- Early	t(344) = 3.14, p = .002	0.31	0.34 [0.12 ; 0.55]
		Established	167	5.55	0.89			Early- PhD	t(346) = 3.40, p < .001	0.32	0.36 [0.15 ; 0.58]
		PhD-students	169	4.92	0.86			Established- PhD	t(334) = 6.59, p < .001	0.63	0.72 [0.50 ; 0.94]
	Established	Early-career	290	5.03	0.99		F(2, 900) = 37.90, p < .001	Established- Early	t(604) = 7.92, p < .001	0.59	0.64 [0.48 ; 0.81]
		Established	316	5.62	0.84			Early- PhD	t(585) =-0.66, p = .509	-0.05	-0.05 [-0.22 ; 0.11]
		PhD-students	297	5.08	0.96			Established- PhD	t(611) = 7.39, p < .001	0.54	0.60 [0.44 ; 0.76]
Openness	Early-career	Early-career	179	5.04	0.91	F(2, 1412) = 11.53, p < .001 *	F(2, 512) = 2.48, p = .085				
		Established	167	5.03	1.10						
		PhD-students	169	4.84	0.82						
	Established	Early-career	290	4.80	1.02		F(2, 900) = 32.22, p < .001	Established- Early	t(604) = 7.68, p < .001	0.62	0.62 [0.46 ; 0.79]
		Established	316	5.42	0.98			Early- PhD	t(585) =-2.80, p = .005	-0.22	-0.23 [-0.49 ;-0.07]
		PhD-students	297	5.02	0.91			Established- PhD	t(611) = 5.22, p < .001	0.40	0.42 [0.26 ; 0.58]

Table S12 *Continued*

Feature	Respondent group	Target	N	Mean	SD	Interaction	Simple effects: effect of Target in each respondent group separately	Comparisons	t-tests	Mean diff.	Cohen's D [95% CI]
Intelligence	Early-career	Early-career	179	3.66	1.04	F(2, 1412) = 1.46, p = .234	Main effect of Respondent Group: F(1, 1414) = 44.08, p < .001. No main effect of Target: F(2, 1414) = 2.20, p = .111 (Model with main effects only)	Established- Early	t(1416) = 6.70, p < .001	0.43	0.37 [0.26 ; 0.48]
		Established	167	3.73	1.12						
		PhD-students	169	3.55	1.09						
	Established	Early-career	290	3.95	1.17						
		Established	316	4.15	1.18						
		PhD-students	297	4.11	1.15						
Integrity	Early-career	Early-career	179	5.24	1.16	F(2, 1412) = 8.62, p < .001 *	F(2, 512) = 3.56, p = .029	Established- Early	t(344) = 1.59, p =.113	0.20	0.17 [-0.04 ; 0.38]
		Established	167	5.45	1.22			Early- PhD	t(346) = 1.08, p =.281	0.13	0.12 [-0.10 ; 0.33]
		PhD-students	169	5.11	1.07			Established- PhD	t(334) = 2.66, p =.008	0.33	0.29 [0.07 ; 0.51]
	Established	Early-career	290	5.05	1.15		F(2, 900) = 31.67, p < .001	Established- Early	t(604) = 7.49, p < .001	0.69	0.61 [0.45 ; 0.77]
		Established	316	5.74	1.12			Early- PhD	t(585) =-5.27, p = .001	-0.46	-0.44 [-0.60 ;-0.27]
		PhD-students	297	5.51	0.97			Established- PhD	t(611) = 2.70, p = .007	0.23	0.22 [0.06 ; 0.38]
Communality	Early-career	Early-career	179	3.88	1.07	F(2, 1412) = 4.36, p = .013 *	Main effect of Target: F(2, 1414) = 11.17, p < .001. No main effect of Resp. group: F(1, 1414) = 2.95, p = .086 (Model with main effects only)	Established- Early	t(928.38) = 3.32, p < .001*	0.25	0.22 [0.09 ; 0.34]
		Established	167	3.89	1.27						
		PhD-students	169	3.97	1.08			Early- PhD	t(931.09) = 3.79, p < .001*	-0.33	-0.32 [-0.45 ;-0.20]
	Established	Early-career	290	3.74	1.03						
		Established	316	4.12	1.24			Established- PhD	t(911.09) =-1.12, p = .263*	-0.08	-0.07 [-0.20 ; 0.05]
		PhD-students	297	4.21	0.93						

Note: For interactions and main effects, α = .008333; for subsequent tests of simple effects, α = 0.05. Text in grey represents non-significant results. *Welch-correction for unequal variances applied (when Levene's test for unequal variances was significant and largest group was more than 1.5 times as large as smallest group).

Table S13 *Sample details Study 4.*

Respondent group	N	Mean Age (years)	SD Age (years)	Range Age (years)	Response rate (%)	Response rate after cleaning (%)
Male scientists	711	45.1	11.9	25- 86	*	*
Female scientists	286	41.8	10.3	24- 77	*	*
Total	1418				11.99	7.62

Note: *Response rates cannot be computed for the two respondent groups separately because we did not know scientists' gender beforehand. The response rate is based on the total number of responses by the total number of e-mails sent (for details see https://osf.io/3nepx/).

Table S14 *Scale reliabilities Study 4.*

Scale	Cronbach's Alpha	95% CI
Objectivity	.62	.55 ; .69
Rationality	.80	.75 ; .86
Open-mindedness	.70	.63 ; .76
Intelligence	.64	.57 ; .71
Integrity	.80	.75 ; .86
Communality	.61	.54 ; .68

Note: 95% CI = 95% confidence interval.

Table S15 *Correlation tables Study 4: correlations between the characteristics of the ideal scientist, by Respondent group.*

Male scientists						
Feature	1	2	3	4	5	6
1. Objectivity	1					
2. Rationality	.65***	1				
3. Open-mindedness	.65***	.65***	1			
4. Intelligence	.37***	.44***	.35***	1		
5. Integrity	.60***	.65***	.58***	.36***	1	
6. Communality	.55***	.40***	.47***	.29***	.57***	1
Female scientists						
1. Objectivity	1					
2. Rationality	.59***	1				
3. Open-mindedness	.62***	.60***	1			
4. Intelligence	.43***	.49***	.44***	1		
5. Integrity	.66***	.65***	.60***	.57***	1	
6. Communality	.50***	.41***	.50***	.43***	.60***	1
Overall						
1. Objectivity	1					
2. Rationality	.63***	1				
3. Open-mindedness	.64***	.63***	1			
4. Intelligence	.39***	.45***	.37***	1		
5. Integrity	.62***	.65***	.58***	.42***	1	
6. Communality	.54***	.40***	.48***	.33***	.58***	1

Note: *significant at α = .05, **significant at α = .01, ***significant at α = .001. All *p*-values are adjusted for multiple testing.

Table S16 *Statistical analyses Study 4.*

Feature	Respondent group	Target	N	Mean	SD	Interaction	Simple effects: effect of Target in each respondent group separately	t-test	Mean Diff.	Cohen's d [95% CI]	Diff. d
Objectivity	Female	Female	153	4.73	1.02	F(1, 993) = 3.94, p = .047	*Model with main effects only* Main effect of Target: F(1, 994) = 26.41, p < .001. No main effect of Respondent Group: F(1, 994) = 0.06, p = .813	t(995) = 5.14, p < .001	0.33	0.33 [0.20 ; 0.45]	
		Male	133	4.19	0.86						
	Male	Female	349	4.61	1.05						
		Male	362	4.36	1.03						
Rationality	Female	Female	153	5.55	0.95	F(1, 993) = 27.68, p < .001 *	F(1, 284) = 47.48, p < .001	t(284) = 6.89, p < .001	0.80	0.82 [0.57 ; 1.06]	0.77
		Male	133	4.75	1.00						
	Male	Female	349	5.09	1.16		F(1, 709) = 0.38, p = .535	t(709) = 0.62, p = .535	0.05	0.05 [-0.01 ; 0.19]	
		Male	362	5.04	0.99						
Openness	Female	Female	153	5.13	1.04	F(1, 993) = 43.71, p < .001 *	F(1, 284) = 70.32, p < .001	t(284) = 8.39, p < .001	0.98	0.99 [0.75 ; 1.24]	0.96
		Male	133	4.15	0.92						
	Male	Female	349	4.78	1.07		F(709) = 0.20, p = .657	t(709) = 0.44, p = .657	0.03	0.03 [-0.11 ; 0.18]	
		Male	362	4.75	0.99						
Intelligence	Female	Female	153	4.65	1.14	F(1, 993) = 6.26 p = .012	*Model with main effects only* Main effect of Target: F(1, 994) = 30.73, p < .001. No main effect of Respondent Group: F(1, 994) = 2.59, p = .108	t(995) = 5.61, p < .001	0.40	0.36 [0.23 ; 0.48]	
		Male	133	3.97	1.01						
	Male	Female	349	4.34	1.08						
		Male	362	4.05	1.17						
Integrity	Female	Female	153	5.10	1.27	F(1, 993) = 18.08, p < .001 *	F(1, 284) = 33.64, p < .001	t(284) = 5.80, p < .001	0.82	0.69 [0.45 ; 0.93]	0.60
		Male	133	4.28	1.08						
	Male	Female	349	4.80	1.17		F(1, 709) = 1.16, p = .219	t(709) = 1.23, p = .219	0.11	0.09 [-0.05 ; 0.24]	
		Male	362	4.69	1.22						
Communality	Female	Female	153	4.32	1.03	F(1, 993) = 25.34, p < .001	F(1, 284) = 90.46, p < .001	t(284) = 9.51, p < .001	1.11	1.13 [0.88 ; 1.38]	0.78
		Male	133	3.21	0.92						
	Male	Female	349	3.96	1.06		F(1, 709) = 22.14, p < .001	t(709) = 4.70, p < .001	0.37	0.35 [0.20 ; 0.50]	
		Male	362	3.59	1.06						

Note: For interactions and main effects, α = .008333; for subsequent tests of simple effects, α = 0.05. Text in grey represents non-significant results. *Welch-correction for unequal variances applied (when Levene's test for unequal variances was significant and largest group was more than 1.5 times as large as smallest group).

MATERIALS

In studies 1, 3 and 4, the following answering options were provided:
□ Strongly Disagree
□ Disagree
□ Somewhat Disagree
□ Neither Agree nor Disagree
□ Somewhat Agree
□ Agree
□ Strongly Agree

Statements used in studies 1, 3, and 4

Study 1

Scientist condition
Below, you will read a series of statements about the typical scientist. By 'scientist', we mean a person who is trained in a science and whose job involves doing scientific research or solving scientific problems. For each statement, please indicate to what extent you agree or disagree. Important: please base your answers on how true you believe each statement is, so the statements do not refer to how you think scientists should behave.

1. A scientist is capable of suppressing personal biases in the interest of objective inquiry *(Objectivity).*
2. A scientist assesses relevant information without prejudicial distortions *(Objectivity).*
3. A scientist exhibits little emotionality with respect to his/her beliefs *(Objectivity).*
4. A scientist has excellent problem-solving skills *(Rationality)*.
5. A scientist can readily discriminate between illogical and logical reasoning *(Rationality)*.
6. A scientist is logical in his/her professional problem solving *(Rationality)*.
7. A scientist suspends judgment when faced with insufficient or ambiguous information *(Open-mindedness)*.
8. A scientist is generally willing to acknowledge evidence that goes against his/her beliefs *(Open-mindedness)*.
9. A scientist is willing to change his/her beliefs when confronted with contrary evidence *(Open-mindedness)*.
10. Standard measures of intelligence are a good predictor of the performance of a scientist *(Intelligence)*.

11. Superior intelligence is a prerequisite for a successful career of a scientist *(Intelligence)*.
12. A scientist has a very high IQ score *(Intelligence)*.
13. A scientist conducts his/her work with integrity *(Integrity)*.
14. A scientist does not engage in unethical behavior to advance his/her career *(Integrity)*.
15. A scientist does not commit fraud in his/her work *(Integrity)*.
16. A scientist does not withhold information from his/her colleagues to protect his/her own interests *(Communality)*.
17. A scientist exhibits cooperative rather than competitive behavior *(Communality)*
18. A scientist is not interested in personal fame or recognition *(Communality)*.

Highly-educated condition

Below, you will read a series of statements about the typical highly-educated person. By ‘a highly-educated person’, we mean a person who obtained a Bachelor’s Degree or a Master’s Degree or a Professional Degree and whose job requires this high level of education. For each statement, please indicate to what extent you agree or disagree. Important: please base your answers on how true you believe each statement is, so the statements do not refer to how you think highly-educated people should behave.

1. A highly-educated person is capable of suppressing personal biases in the interest of objective inquiry *(Objectivity).*
2. A highly-educated person assesses relevant information without prejudicial distortions *(Objectivity).*
3. A highly-educated person exhibits little emotionality with respect to his/her beliefs *(Objectivity).*
4. A highly-educated person has excellent problem-solving skills *(Rationality)*.
5. A highly-educated person can readily discriminate between illogical and logical reasoning *(Rationality)*.
6. A highly-educated person is logical in his/her professional problem solving *(Rationality)*.
7. A highly-educated person suspends judgment when faced with insufficient or ambiguous information *(Open-mindedness)*.
8. A highly-educated person is generally willing to acknowledge evidence that goes against his/her beliefs *(Open-mindedness)*.
9. A highly-educated person is willing to change his/her beliefs when confronted with contrary evidence *(Open-mindedness)*.
10. Standard measures of intelligence are a good predictor of the performance of a highly-educated person *(Intelligence)*.

11. Superior intelligence is a prerequisite for a successful career of a highly-educated person *(Intelligence)*.
12. A highly-educated person has a very high IQ score *(Intelligence)*.
13. A highly-educated person conducts his/her work with integrity *(Integrity)*.
14. A highly-educated person does not engage in unethical behavior to advance his/her career *(Integrity)*.
15. A highly-educated person does not commit fraud in his/her work *(Integrity)*.
16. A highly-educated person does not withhold information from his/her colleagues to protect his/her own interests *(Communality)*.
17. A highly-educated person exhibits cooperative rather than competitive behavior *(Communality)*.
18. A highly-educated person is not interested in personal fame or recognition *(Communality)*.

Study 3

PhD-student condition

Below, you will read a series of statements about the typical PhD-student. By 'PhD-student', we mean a graduate student at an academic institution who is conducting scientific research for his/her doctoral dissertation. For each statement, please indicate to what extent you agree or disagree. Important: please base your answers on how true you believe each statement is, so the statements do not refer to how you think PhD-students should behave.

1. A PhD-student is capable of suppressing personal biases in the interest of objective inquiry (Objectivity).
2. A PhD-student assesses relevant information without prejudicial distortions (Objectivity).
3. A PhD-student exhibits little emotionality with respect to his/her beliefs (Objectivity).
4. A PhD-student has excellent problem-solving skills (Rationality).
5. A PhD-student can readily discriminate between illogical and logical reasoning (Rationality).
6. A PhD-student is logical in his/her professional problem solving (Rationality).
7. A PhD-student suspends judgment when faced with insufficient or ambiguous information (Open-mindedness).
8. A PhD-student is generally willing to acknowledge evidence that goes against his/her beliefs (Open-mindedness).
9. A PhD-student is willing to change his/her beliefs when confronted with contrary evidence (Open-mindedness).

10. Standard measures of intelligence are a good predictor of the performance of a PhD-student (Intelligence).
11. Superior intelligence is a prerequisite for a successful career of a PhD-student (Intelligence).
12. A PhD-student has a very high IQ score (Intelligence).
13. A PhD-student conducts his/her work with integrity (Integrity).
14. A PhD-student does not engage in unethical behavior to advance his/her career (Integrity).
15. A PhD-student does not commit fraud in his/her work (Integrity).
16. A PhD-student does not withhold information from his/her colleagues to protect his/her own interests (Communality).
17. A PhD-student exhibits cooperative rather than competitive behavior (Communality).
18. A PhD-student is not interested in personal fame or recognition (Communality).

Early-career scientist condition

Below, you will read a series of statements about the typical early-career scientist. By 'early-career scientist', we mean a post-doctoral academic who obtained their PhD less than 10 years ago, and does not yet have tenure at a university or other academic institution. For each statement, please indicate to what extent you agree or disagree. Important: please base your answers on how true you believe each statement is, so the statements do not refer to how you think early-career scientists should behave.

1. An early-career scientist is capable of suppressing personal biases in the interest of objective inquiry *(Objectivity).*
2. An early-career scientist assesses relevant information without prejudicial distortions *(Objectivity).*
3. An early-career scientist exhibits little emotionality with respect to his/her beliefs *(Objectivity).*
4. An early-career scientist has excellent problem-solving skills *(Rationality)*.
5. An early-career scientist can readily discriminate between illogical and logical reasoning *(Rationality)*.
6. An early-career scientist is logical in his/her professional problem solving *(Rationality)*
7. An early-career scientist suspends judgment when faced with insufficient or ambiguous information *(Open-mindedness)*.
8. An early-career scientist is generally willing to acknowledge evidence that goes against his/her beliefs *(Open-mindedness)*.

9. An early-career scientist is willing to change his/her beliefs when confronted with contrary evidence *(Open-mindedness)*.
10. Standard measures of intelligence are a good predictor of the performance of an early-career scientist *(Intelligence)*.
11. Superior intelligence is a prerequisite for a successful career of an early-career scientist *(Intelligence)*.
12. An early-career scientist has a very high IQ score *(Intelligence)*.
13. An early-career scientist conducts his/her work with integrity *(Integrity)*.
14. An early-career scientist does not engage in unethical behavior to advance his/her career *(Integrity)*.
15. An early-career scientist does not commit fraud in his/her work *(Integrity)*.
16. An early-career scientist does not withhold information from his/her colleagues to protect his/her own interests *(Communality)*.
17. An early-career scientist exhibits cooperative rather than competitive behavior *(Communality)*.
18. An early-career scientist is not interested in personal fame or recognition *(Communality)*.

Established scientist condition

Below, you will read a series of statements about the typical established scientist. By ‘established scientist’, we mean a scientist who obtained their PhD more than 10 years ago, and has tenure at a university or other academic institution. For each statement, please indicate to what extent you agree or disagree. Important: please base your answers on how true you believe each statement is, so the statements do not refer to how you think established scientists should behave.

1. An established scientist is capable of suppressing personal biases in the interest of objective inquiry *(Objectivity).*
2. An established scientist assesses relevant information without prejudicial distortions *(Objectivity).*
3. An established scientist exhibits little emotionality with respect to his/her beliefs *(Objectivity).*
4. An established scientist has excellent problem-solving skills *(Rationality)*.
5. An established scientist can readily discriminate between illogical and logical reasoning *(Rationality)*.
6. An established scientist is logical in his/her professional problem solving *(Rationality)*.
7. An established scientist suspends judgment when faced with insufficient or ambiguous information *(Open-mindedness)*.

8. An established scientist is generally willing to acknowledge evidence that goes against his/her beliefs *(Open-mindedness)*.
9. An established scientist is willing to change his/her beliefs when confronted with contrary evidence *(Open-mindedness)*.
10. Standard measures of intelligence are a good predictor of the performance of an established scientist *(Intelligence)*.
11. Superior intelligence is a prerequisite for a successful career of an established scientist *(Intelligence)*.
12. An established scientist has a very high IQ score *(Intelligence)*.
13. An established scientist conducts his/her work with integrity *(Integrity)*.
14. An established scientist does not engage in unethical behavior to advance his/her career *(Integrity)*.
15. An established scientist does not commit fraud in his/her work *(Integrity)*.
16. An established scientist does not withhold information from his/her colleagues to protect his/her own interests *(Communality)*.
17. An established scientist exhibits cooperative rather than competitive behavior *(Communality)*.
18. An established scientist is not interested in personal fame or recognition *(Communality)*.

Study 4

Male scientist condition
Below, you will read a series of statements about the typical male scientist. For each statement, please indicate to what extent you agree or disagree. Important: please base your answers on how true you believe each statement is, so the statements do not refer to how you think male scientists should behave.

1. A male scientist is capable of suppressing personal biases in the interest of objective inquiry *(Objectivity).*
2. A male scientist assesses relevant information without prejudicial distortions *(Objectivity).*
3. A male scientist exhibits little emotionality with respect to his beliefs *(Objectivity).*
4. A male scientist has excellent problem-solving skills *(Rationality)*.
5. A male scientist can readily discriminate between illogical and logical reasoning *(Rationality)*.
6. A male scientist is logical in his professional problem solving *(Rationality)*.
7. A male scientist suspends judgment when faced with insufficient or ambiguous information *(Open-mindedness)*.

8. A male scientist is generally willing to acknowledge evidence that goes against his beliefs *(Open-mindedness)*.
9. A male scientist is willing to change his beliefs when confronted with contrary evidence *(Open-mindedness)*.
10. Standard measures of intelligence are a good predictor of the performance of a male scientist *(Intelligence)*.
11. Superior intelligence is a prerequisite for a successful career of a male scientist *(Intelligence)*.
12. A male scientist has a very high IQ score *(Intelligence)*.
13. A male scientist conducts his work with integrity *(Integrity)*.
14. A male scientist does not engage in unethical behavior to advance his career *(Integrity)*.
15. A male scientist does not commit fraud in his work *(Integrity)*.
16. A male scientist does not withhold information from his/her colleagues to protect his own interests *(Communality)*.
17. A male scientist exhibits cooperative rather than competitive behavior *(Communality)*.
18. A male scientist is not interested in personal fame or recognition *(Communality)*.

Female scientist condition

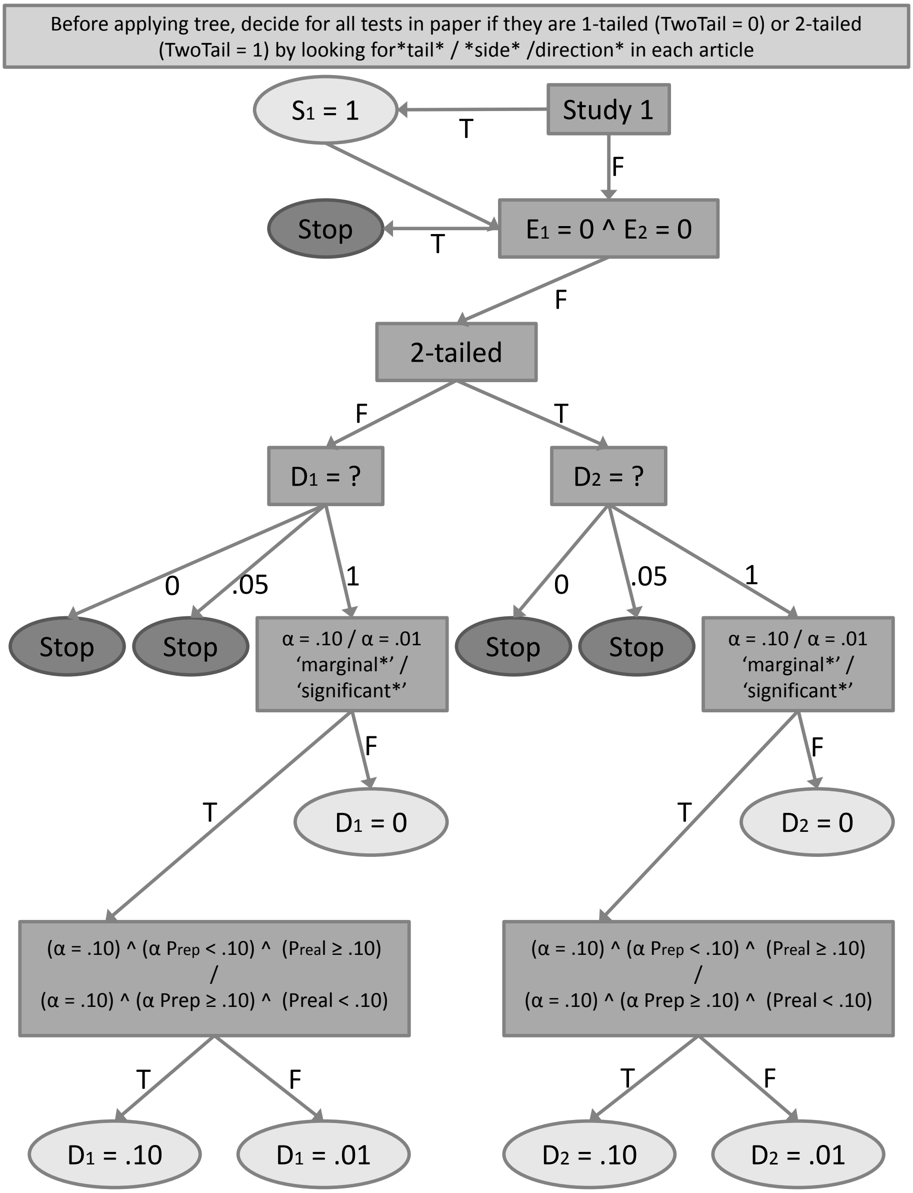
Below, you will read a series of statements about the typical female scientist. For each statement, please indicate to what extent you agree or disagree. Important: please base your answers on how true you believe each statement is, so the statements do not refer to how you think female scientists should behave.

1. A female scientist is capable of suppressing personal biases in the interest of objective inquiry *(Objectivity).*
2. A female scientist assesses relevant information without prejudicial distortions *(Objectivity).*
3. A female scientist exhibits little emotionality with respect to her beliefs *(Objectivity).*
4. A female scientist has excellent problem-solving skills *(Rationality)*.
5. A female scientist can readily discriminate between illogical and logical reasoning *(Rationality)*.
6. A female scientist is logical in her professional problem solving *(Rationality)*.
7. A female scientist suspends judgment when faced with insufficient or ambiguous information *(Open-mindedness)*.
8. A female scientist is generally willing to acknowledge evidence that goes against her beliefs *(Open-mindedness)*.

9. A female scientist is willing to change her beliefs when confronted with contrary evidence *(Open-mindedness)*.
10. Standard measures of intelligence are a good predictor of the performance of a female scientist *(Intelligence)*.
11. Superior intelligence is a prerequisite for a successful career of a female scientist *(Intelligence)*.
12. A female scientist has a very high IQ score *(Intelligence)*.
13. A female scientist conducts her work with integrity *(Integrity)*.
14. A female scientist does not engage in unethical behavior to advance her career *(Integrity)*.
15. A female scientist does not commit fraud in her work *(Integrity)*.
16. A female scientist does not withhold information from her colleagues to protect her own interests *(Communality)*.
17. A female scientist exhibits cooperative rather than competitive behavior *(Communality)*.
18. A female scientist is not interested in personal fame or recognition *(Communality)*.

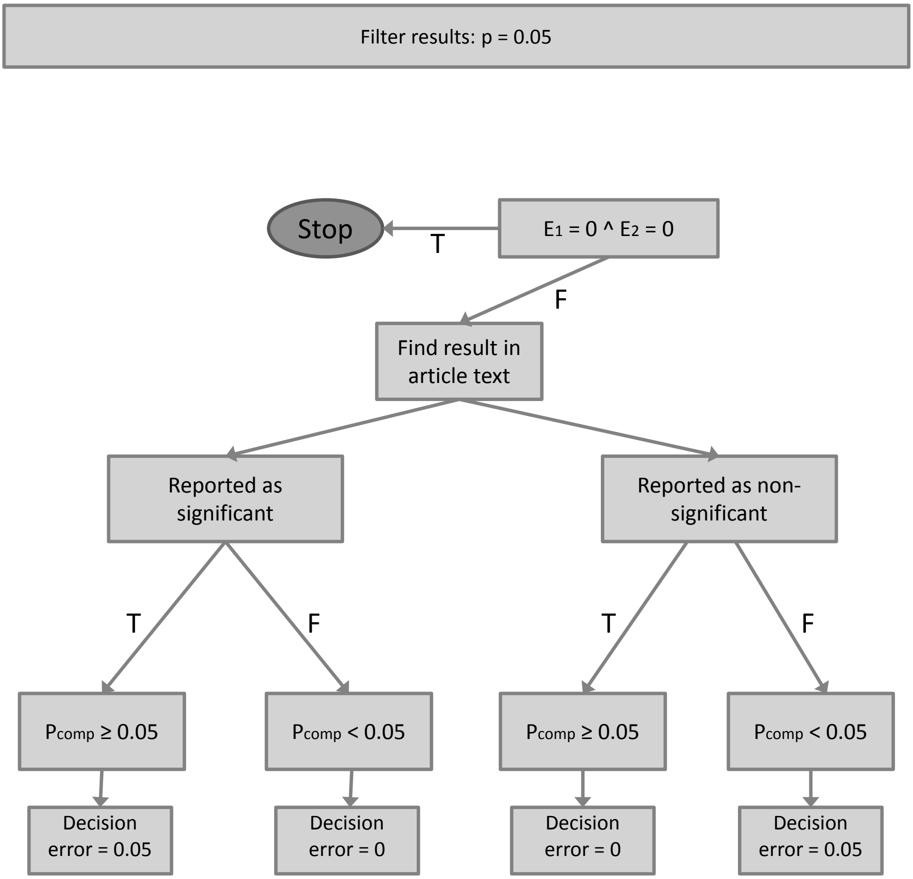
APPENDIX B: SUPPLEMENTARY MATERIALS OF CHAPTER 3

Figure S1 *Manual check of decision errors part 1.*


Before applying tree, decide for all tests in paper if they are 1-tailed (TwoTail = 0) or 2-tailed (TwoTail = 1) by looking for*tail* / *side* /direction* in each article
S1 = 1
Study 1
T
F
Stop
E1 = 0 ^ E2 = 0
T
F
2-tailed
F
T
D1 = ?
D2 = ?
0
.05
1
0
.05
1
Stop
Stop
α = .10 / α = .01 'marginal*' / 'significant*'
Stop
Stop
α = .10 / α = .01 'marginal*' / 'significant*'
F
T
D1 = 0
F
T
D2 = 0
(α = .10) ^ (α Prep < .10) ^ (Preal ≥ .10) / (α = .10) ^ (α Prep ≥ .10) ^ (Preal < .10)
(α = .10) ^ (α Prep < .10) ^ (Preal ≥ .10) / (α = .10) ^ (α Prep ≥ .10) ^ (Preal < .10)
T
F
T
F
D1 = .10
D1 = .01
D2 = .10
D2 = .01

Note: S_1 = study 1, E_1 = one-sided error, E_2 = two-sided error, D_1 = one-tailed decision error, D_2 = two-tailed decision error, P_{rep} = reported *p*-value, P_{real} = computed/real *p*-value.

Figure S2 *Manual check of decision errors part 2.*



Note: S_1 = study 1, E_1 = one-sided error, E_2 = two-sided error, P_{rep} = reported ***p***-value, P_{comp} = computed/real *p*-value

Table S1 *Coding protocol*

Variable	How entered	Coding	Explanation
Source	statcheck	String	Filename as coded by student assistant. This includes journal, author, year, volume, issue, page number of first page, and title.
Statistic	statcheck	t, F, r, X2, Z, Wald	Test 'statistic'. In this case, r and Wald are also counted as test statistics.
df1	statcheck	Numerical	Reported degrees of freedom.
df2	statcheck	Numerical	Reported degrees of freedom.
Test.Comparison	statcheck	<, >, =	How is the test statistic reported? E.g. t(10) < 1, t(10) > 1 or t(10) = 1.
Value	statcheck	Numerical	Value of the test statistic.
Reported.Comparison	statcheck	<, >, =, ns	How is the p value reported? P < .05, p > .05, p = .05, or ns for not significant.
Reported.P.Value	statcheck	Numerical	Reported *p*-value.
Computed	statcheck	Numerical	Computed *p*-value.
Raw	statcheck	String	The raw result as read by statcheck.
Error	statcheck	Logical (1, 0)	1 if the reported p value is incongruent with the computed p value (else 0).
DecisionError	statcheck/manually	0, 0.01, 0.05,0.1, 1	A decision error occurs when a reported significant result is not significant after recomputation or when a reported non-significant result is actually significant Statcheck: 0=no decision error, .05= decision error at alpha is .05, 1=decision error at alpha is .10 or .01 Manually: when DecisionError=1, check the actual level of significance in paper. If it's not mentioned, fill in 0. If it is, fill in the reported level of significance.
CopyPaste	statcheck	Logical (1, 0)	1 if the exact string of the extracted raw results also occurs somewhere else in the article.
Error_OneTail	statcheck	Logical (1, 0)	1 if the reported p value is incongruent with the one sided computed p value (else 0).
DecisionError_OneTail	statcheck/manually	0, 0.01, 0.05,0.1, 1	A decision error occurs when a reported significant result is not significant after recomputation or when a reported non-significant result is actually significant Statcheck: 0=no one tailed decision error, .05= one tailed decision error at alpha is .05, 1= one tailed decision error at alpha is .10 or .01 Manually: when DecisionError_OneTail=1, check the actual level of significance in paper. If it's not mentioned, fill in 0. If it is, fill in the reported level of significance.

Table S1 *Continued*

Variable	How entered	Coding	Explanation
TwoTailed	statcheck/manually	Logical (1, 0)	1 if the test is two tailed (default). Manually change it to 0 if the test is one tailed.
Study 1	Manually	Logical (1, 0)	1 if the result belongs to study 1 (0=default). Manually change it to 1 if the result belongs to study 1.
Coder 1	manually	Initials	Who coded the article the first time? Fill in your initials (e.g., MN for Michèle Nuijten)
Coder 2	manually	Initials	Who checked the article the second time? Fill in your initials (e.g., CV for Coosje Veldkamp)
Check error coder 1	manually	Logical (1, 0)	If statcheck reported an error, check this manually. 1 if the result really is an error. 0 if the result is correct and statcheck wrongly classified it as error.
Check error coder 2	manually	Logical (1, 0)	If statcheck reported an error, check this manually. 1 if the result really is an error. 0 if the result is correct and statcheck wrongly classified it as error.

APPENDIX C: SUPPLEMENTARY MATERIALS OF CHAPTER 6

PROTOCOL SCORING PREREGISTRATIONS

Hypotheses

Q1. Is at least one hypothesis specified such that it is clear what are the independent and dependent variables? From the text it should be clear what will be tested (*at the level of the preregistration*).

- **NO** → **T1 = all DFs = 0**
- **YES** → **T1 = R6 = 2**
- **YES**, and the text specifies that this is/ these are the only hypotheses tested in the confirmatory part of the analyses → **T1 = R6 = 3**

If the answer was **YES**, fill out the **number of specified hypotheses** in the column '**Nr_hyp_C_1**' or '**Nr_hyp_C_2**' (depending on whether you're coder 1 or coder 2 for this pre-registration). Please look at columns C and D to check this!

For *all hypotheses*, answer the following questions for each analysis mentioned in the pre-registration. **<u>If</u>** multiple hypotheses are described in the pre-registration, then your evaluation of a researcher df will be the <u>minimum</u> of the evaluations for each hypothesis in that pre-registration.

Q2. Is the direction of the hypothesis specified?

- **NO** → **T2 = 0**
- **YES** → **T2 = 2**
- **YES**, and the text specifies the sidedness of the statistical test (if one-tailed, then the direction needs to be specified) → **T2 = 3**

Design

Q3a. Does the design include manipulated variables? Note: independent variables in within-subjects designs are also considered manipulated variables)

- **NO →** **D1=A8=A9 = 99**(NA) **+ go to Q5**
- **YES →** **Go to Q3b**

Q3b. Does the text explicitly exclude the possibility that at least one of the manipulated variables specified in the hypothesis will be omitted in the test of the hypothesis reported in the confirmatory analysis section?

- **NO →** **D1 = A8 = 0**
- **YES →** **D1 = A8 = 3**

Q4. Does the text specify exactly how the manipulated variable will be used in the analysis to test the hypothesis (is it reproducible how each of the variables will be treated in the analysis, i.e. what are the values of the independent variable in the analysis)?

- **NO →** **D1 = A9 = 0**
- **YES →** **D1 = min(A8, 2); A9 = 2**
- **YES**, and the text specifies the manipulated variable will not be used in another way (combination or splitting of conditions) in analyses to test the hypothesis in the confirmatory analysis section → **D1 = min(A8, 3); A9 = 3**

Q5. Does the text explicitly exclude the possibility that at least one other variable (e.g. a covariate) is included in the analysis testing the hypothesis reported in the confirmatory analysis section?

- **NO →** **D2 = A10 = 0**
- **YES →** **D2 = A10 = 3**

Q6. Does the text specify exactly which measurement instrument (test, scale, question set, physical measurement) will be used as the main outcome variable?

- **NO**, it is not specified how the outcome variable is measured → **D3=A5=A6=0**
- **YES** the measurement instrument is specified → **D3=A5=2**

- **YES,** and states that this is the only measurement instrument to be used in the analyses→ **D3=A5=3**

Q7. Does the text explicitly specify that the confirmatory analysis section of the paper will not include another dependent variable than the ones specified in all hypotheses of the preregistration? (answer at the level of the paper)

- **NO** → **D4 = A7 = R6 = 0**

- **YES** → **D4 = A7 = 3; R6 = min(T1, 3)**

Q8. Does the pre-registration indicate inclusion and exclusion criteria in selecting data points used in the analysis?

- **NO** not described at all → **D5 = A12 = 0**

- **PARTIAL** mentioned, but inclusion and exclusion criteria are insufficiently reproducible because of a lack of objective criteria or operationalization of the inclusion and exclusion of data (e.g., not specific on what it means for a participant to not participate seriously, what "awareness of the study purposes" means, what scores on the selection variable are used to (de)select participants or data points, which clinical criteria are used) → **D5 = A12 = 1**

- **YES** inclusion and exclusion criteria are objective and reproducible → **D5 = A12 = 2**

- **YES** like 2 & explicitly excluding other reasons for inclusion and exclusion ("we will only use" or including statement "we will use data from all participants") → **D5 = A12 = 3**

Q9. Is a power analysis reported?

- **NO** → **D6 = 0**

- **YES** but power level used for the power analysis < .8 → **D6 = 1**

- **YES** the effect size estimate used for the power analysis is based on ((a representative preliminary study or meta-analytical results OR (set at medium or smaller)) AND (at the same time the power analysis is used to make a sample size decision) → **D6 = 2**

- **YES** like previous AND the text indicates no other power analysis will be included in the paper than this one → **D6 = 3**

Q10. (1) Is the exact number or range of participants given (not a minimum) AND at the same time (2) is the protocol of sampling described in all its details [i.e. (2a) exact number of people that will be approached is given, AND (2b) how (exact time frame and situation in which participants will be invited, e.g., all visitors of shop X in week Z are invited to participate in our experiment), AND (2c) inclusion and exclusion criteria for selecting participants or data points, AND (2d) how many and how additional participants or data points are sampled when pre-set sample size is not reached?

- **NO** → **D7 = C4 = 0**
- **YES** partly but at most 1 of 2 (exact number/range or protocol) → **D7 = C4 = 1**
- **YES** → **D7 = C4 = 2**
- **YES** and text states paper will not deviate from the sampling plan →**D7 = C4 = 3**

Collecting data

Q11a. Is the design of the study experimental, i.e. does it involve randomization of participants across conditions, or, in a within subjects design, does it involve randomization of task or question order?

- **NO** → **C1= 99 (= NA)**
  + go to Q12a
- **YES** → **Go to Q11b**

Q11b. Is specified how randomization is implemented?

- **NO** → **C1 = 0**
- **YES** text describes randomization procedure → **C1 = 1**
- **YES** text describes randomization procedure, AND at the same time the randomization procedure seems to work (i.e. randomization procedure cannot result in dependencies in the data, e.g. all participants who arrive early are in one condition and the participants who arrive later are in the other condition) → **C1 = 2**

- **YES** like previous and the text indicates the implemented randomization for the paper will not be different from the preregistration → **C1 = 3**

Q12a. Is blinding of participants and/or experimenters mentioned?

- **NO** → **C2= 99** (NA) **+ go to Q13**
- **YES** → **Go to Q12b**

Q12b. Does the pre-registration describe procedures to blind participants to and/or experimenters to conditions?

- **NO** → **C2 = 0**
- **YES** describes procedures to blind participants and/or blind experimenters, but not in a detailed and reproducible manner → **C2 = 1**
- **YES** provides detailed AND reproducible protocol for blinding participants and/or experimenters → **C2 = 2**
- **YES** like previous AND assures no other blinding procedures are used (i.e. study will not deviate from the protocols of the experiment concerning knowledge of subjects and experimenters on participation of subjects in conditions, and (possible) contact between subjects and experimenter) → **C2 = 3**

Q13. Does the pre-registration include protocols concerning coding of data, discarding of cases, or correction of scores during data collection?

- **NO** → **C3 = 0**
- **PARTIAL** → text provides protocol but not for all three issues → **C3 = 1**
- **YES** → text provides reproducible protocol for all three issues → **C3 = 2**
- **YES** like previous AND text indicates that experiment will not deviate from the protocol concerning these three issues → **C3 = 3**

Analyses

Answer all remaining questions for each analysis mentioned in the pre-registration concerning the hypothesis. **<u>If</u>** multiple analyses are run to test the same hypothesis, then your evaluation of a researcher df will be the <u>minimum</u> of the evaluations for each analysis of that same hypothesis.

Q14. Does the pre-registration indicate how the study deals with incomplete or missing data?

- **NO** not described at all → **A1 = 0**
- **PARTIAL** described but not entirely reproducible on at least one aspect: criterion to drop cases because of missingness (definition of missing case), procedure (pairwise deletion, listwise deletion, imputation, full information methods, intention to treat), or method to check for randomness of missingness (or selective missingness) → **A1 = 1**
- **YES** reproducible on all three aspects (criterion, procedure, check for randomness) → **A1 = 2**
- **YES** like previous AND explicitly excluding other ways of dealing with incomplete or missing data ("we will only use") → **A1 = 3**

Q15a. Does the study involve data collection methods such as EEG, MRI, MEG, (molecular) genetic measures, physiological measures, hormonal measures, blood readings, coded behaviour of participants of observational studies, or other data-intensive measures requiring pre-processing of the data?

- **NO** → **A2= 99** (NA) **+ go to Q16**
- **YES** → **Go to Q15b**

Q15b. Does the pre-registration offer a protocol for pre-processing the data (e.g., cleaned, normalized, smoothed, corrected for motion and other artifacts)?

- **NO** not described at all → **A2 = 0**
- **YES** detailed protocol offered → **A2 = 2**
- **YES** like previous AND explicitly excluding other methods of pre-processing ("we will only use") → **A2 = 3**

Q16. Does the pre-registration indicate how to test for violations of statistical assumptions and what to do with possible violations?

- **NO** not described at all → **A3 = 0**

- **PARTIAL** described but not reproducible on at least one of the following three aspects: which assumptions are checked (e.g., normality, homoscedascity, linearity, homogeneity of variances, sphericity), how these assumptions are checked (e.g., type of test like Levene's test, alpha level etc.), and what is to be done in cases of violations (e.g., transformations, non-parametric tests, etc.) → **A3 = 1**

- **YES** reproducible on all three aspects (type of assumptions, checks of assumptions, dealing with violations) → **A3 = 2**

- **YES** like previous AND explicitly excluding other methods of dealing with model violations ("we will only use") → **A3 = 3**

Q17. Does the pre-registration indicate how to detect outliers and how they should be dealt with?

- **NO** not described at all → **A4 = 0**

- **PARTIAL** described but not reproducible on at least one of the following two aspects: what objectively defines an outlier (e.g., particular Z value, values for median absolute deviation statistic (MAD), interquartile range (IQR), Mahalanobis distance) and how they are dealt with (e.g., exclusion, method of Winsorisation, type of non-parametric test, type of robust method, bootstrapping) → **A4 = 1**

- **YES** reproducible on both aspects (objective definition of outliers & method of dealing with outliers) → **A4 = 2**

- **YES** like previous AND explicitly excluding other methods of dealing with outliers ("we will only use") → **A4 = 3**

In case the text clearly specifies how the outcome variable is measured, we distinguish between a non-composite (one measurement Y, go to Q18a) and a composite (several measurements or items are combined to one scale or measurement [using a sum or linear combination, or SEM] for the outcome variable Y, go to Q18b).

Q18a. *Non-composite.* Is protocol Z to measure outcome variable Y using instrument A described (i.e., the exact procedure of measurement, including a list of conditions that are controlled while measuring, list and range of potential values)?

- **NO** description provided→ **A6 = 0**
- **PARTIAL** description given but on one of the aspects not reproducible (procedure, conditions, list/range of values) → **A6 = 1**
- **YES** reproducible on all aspects → **A6 = 2**
- **YES** reproducible on all aspects, and promise not to deviate → **A6 = 3**

Q18b. *Composite.* The text describes protocol Z to measure each element of the composite [protocol], the procedure how to construct the composite from its elements (arithmetic mean, weighted mean, sum, other) [dealing index], and how is dealt with
(i) possible deviating individual items [dealing items];
(ii) possibly changing / combining values of individual items [dealing values];
(iii) scores of individuals who have at least one missing [dealing missings].

- **NO** description provided (e.g., only scale is mentioned) → **A6 = 0**
- **PARTIAL**, at least one, but not all of the following aspects is missing: [dealing values], [dealing items], [dealing index], [protocol], [dealing missings] → **A6 = 1**
- **YES** incudes all 5 aspects and is thereby reproducible → **A6 = 2**
- **YES** like previous AND promise not to use other scores → **A6 = 3**

*Answer Q19 for each IV in each analysis of the same hypothesis. **If** multiple IVs are involved in the analysis/es of the same hypothesis, then your answer to Q19 will be the minimum of the answers for each IV involved in the analysis/es of that same hypothesis.*

Q19a. Does the design include non-manipulated independent variables?

- **NO** → **A11= 99** (NA) **+ go to Q20**
- **YES** → **Go to Q19b**

Q19b. Does the text clearly specify which measurement instrument (test, scale, question set, physical measurement) will be used as the non-manipulated independent variable?

- **NO**, it is not specified how the non-manipulated independent variable(s) is or are measured → **A11 = 0** **+ go to Q20**

In case the text clearly specifies how the IV is measured, we distinguish between a non-composite (one measurement X, go to Q19c) and a composite (several measurements or items are combined to one scale or measurement [using a sum or linear combination, or SEM] for the IV, go to Q19d).

Q19c. *Non-composite.* Is protocol Z to measure IV X using instrument A described (i.e., the exact procedure of measurement, including a list of conditions that are controlled while measuring, list and range of potential values)?

- **NO** description provided→ **A11 = 0**

- **PARTIAL** description given but on one of the aspects not reproducible (procedure, conditions, list/range of values) → **A11 = 1**

- **YES** description reproducible on all aspects → **A11 = 2**

- **YES** like previous AND promise not to deviate → **A11 = 3**

Q19d. *Composite.* The text describes protocol Z to measure each element of the composite [protocol], the procedure how to construct the composite from its elements (arithmetic mean, weighted mean, sum, other) [dealing index], and how is dealt with
(i) possible deviating individual items [dealing items];
(ii) possibly changing / combining values of individual items [dealing values];
(iii) scores of individuals who have at least one missing [dealing missings].

- **NO** description provided (e.g., only scale is mentioned) → **A11 = 0**

- **PARTIAL**, at least one of the following aspects is missing: [dealing values], [dealing items], [dealing index], [dealing missings] → **A11 = 1**

- **YES** incudes all 5 aspects and is thereby reproducible → **A11 = 2**

YES like previous AND promise not to use other scores → **A11 = 3**

Q20. Does the pre-registration specify the statistical model(s) that will be used to test the hypothesis (e.g., MANOVA, logistic regression, linear regression, multilevel regression, loglinear analysis, SEM)?

- **NO** not described at all → **A13 = 0**
- **PARTIAL** type of model is mentioned but descriptions fails to explicate all relevant predictors/factors (including covariates and interaction terms) and the manner in which these are used in analysis (e.g., mean centered, SEM model specification including potential residual covariances) → **A13 = 1**
- **YES** full statistical model is presented and thereby reproducible → **A13 = 2**
- **YES** like previous AND explicitly excluding other models to be used ("we will only use") → **A13 = 3**

Q21a. Does the pre-registration indicate details of the estimation technique used to estimate the statistical model and to compute standard errors?

- **NO** not described at all → **A14 = 0**
- **PARTIAL** estimation technique is mentioned in general terms (e.g., maximum likelihood) and no mention is made of potential use of robust SEs or any other correction to the model fit measures or SEs (e.g., Satorra Bentler correction) → **A14 = 1**
- **YES** script is provided and/or specific estimation technique is described in detail (e.g., restricted maximum likelihood, specification of generalized linear model estimation, weighted least squares, mean and variance adjusted weighted least squares, partial least squares, robust standard errors) including the manner in which standard errors are computed, thereby the estimation is reproducible → **A14 = 2**
- **YES** like previous AND explicitly excluding other corrections to be used ("we will only use") → **A14 = 3**

Q21b. Does the pre-registration specify which statistical software package and version is used for running the analyses?

- **NO** not described at all → **A14 = 0**

- **PARTIAL** mentioned in general terms (e.g., analyses to be run in SPSS, R, SAS, LISREL) but without mentioning version number and/or without specific package (in R) or add-on syntax (if applicable) → **min(A14 =1, Q21a)**

- **YES** mentions software package, version code, and specific package/syntax (if applicable), thereby the power-analysis is reproducible → **min(A14 =2, Q21a)**

- **YES** like previous AND explicitly excluding use of other software or packages ("we will only use") → **min(A14 =3, Q21a)**

Q22. Does the pre-registration indicate the inference criteria (e.g., Bayes factors, Alpha level, sidedness of the test, corrections for multiple testing)?

- **NO** not described at all → **A15 = 0**

- **PARTIAL** mentions one of the criteria (e.g., overall Alpha level), but not all. For instance, it fails to report the sidedness of the test or possible corrections for multiple testing → **A15 = 1**

- **YES** mentions all inference criteria, including Alpha level/Bayes factor thresholds, corrections for multiple testing, sidedness of the test, thereby the inference criteria are reproducible → **A15 = 2**

- **YES** like previous AND explicitly excluding use of additional inference criteria ("we will only use") → **A15 = 3**

ADDENDUM

Author publications

Dankwoord

AUTHOR PUBLICATIONS

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), e0172792.

Sijtsma, K., **Veldkamp, C. L. S.,** & Wicherts, J. M. (2015). Improving the conduct and reporting of statistical analysis in psychology. *Psychometrika*, *81*(1), 33-38.

Tijdink, J. K., Bouter, L. M., **Veldkamp, C. L. S.**, van de Ven, P. M., Wicherts, J. M., & Smulders, Y. M. (2016). Personality traits are associated with research misbehavior in Dutch scientists: a cross-sectional study. *PLOS ONE, 11*(9), e0163251.

Veldkamp, C. L. S., Hartgerink, C. H. J., van Assen, M. A. L. M., & Wicherts, J. M. (2017). Who believes in the storybook image of the scientist? *Accountability in research, 24*(3), 127-151.

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.

Wicherts, J. M., **Veldkamp, C. L. S.**, Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid *p*-hacking. *Frontiers in Psychology, 7*(1832).

Nuijten, M. B., van Assen, M. A., **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.

DANKWOORD

Toen ik vijf jaar geleden aan mijn promotietraject begon had ik niet voor mogelijk gehouden dat dit zo'n bijzondere tijd zou worden. Zonder de mensen die ik hier wil bedanken was mijn proefschrift niet tot stand gekomen en had ik niet zulke mooie herinneringen aan mijn periode als promovendus gehad.

Allereerst wil ik mijn promotoren Prof. Jelte Wicherts en Prof. Marcel van Assen bedanken. Jullie waren ieder op jullie eigen manier fantastische begeleiders die mij hebben laten groeien als onderzoeker en als persoon. Jelte, meteen vanaf het begin gaf je mij de vrijheid waar ik behoefte aan had, terwijl je er tegelijkertijd ook direct was als ik je nodig had. Hoe druk je het ook had, je stelde altijd mijn belangen voorop en kwam onmiddellijk als een soort ridder voor mij op als er ook maar iets was dat mij in de weg stond. Met je ongelooflijke drive, kennis, enthousiasme, en humor gaf je me vleugels. Wanneer er echter iets te veel ideeën uit mijn hoofd kwamen vliegen, trok je me weer met beide benen op de grond. Zeer veel dank voor alles!

Marcel, als een vader heb je altijd over mijn welzijn gewaakt. In het begin moest ik even wennen aan je feedback-stijl en je directheid, maar al gauw ging dat goed. Je was altijd buitengewoon behulpzaam. Als ik dacht met een ingewikkeld statistisch probleem te zitten, begreep je al na drie woorden wat ik niet begreep en tekende je de oplossing meteen helder op papier. Sowieso geloof jij niet in problemen; je joeg elke beer die ik op de weg zag direct het bos in. Dankzij jouw relativeringsvermogen heb ik geleerd onderscheid te maken tussen wat meer en minder belangrijk is, en dat niet alleen in de wetenschap, maar ook in het leven in de breedste zin. Ook had ik nooit gedacht dat er in de wetenschap zo ontzettend hard en veel te lachen viel! Enorm veel dank voor alles.

Ook zou ik graag de leden van mijn leescommissie willen bedanken. Prof. Lex Bouter, Prof. Klaas Sijtsma, Prof. Eric-Jan Wagenmakers, en Dr. Rink Hoekstra, veel dank voor de tijd en moeite die jullie hebben genomen om mijn proefschrift te lezen en om naar Tilburg te komen om deel te nemen aan de discussie van mijn proefschrift. Dr. Simine Vazire, thank you for the time and effort you invested in reading my dissertation, and for traveling all the way from the USA to join the discussion of my dissertation. Prof. Franca Agnoli, La ringrazio per il tempo e lo sforzo che ha investito nel leggere la mia tesi, e per viaggiare dall'Italia a partecipare alla discussione della mia tesi.

Dan onze onderzoeksgroep! Vijf jaar geleden begonnen we met z'n vieren, maar na een paar jaar waren we al met z'n negenen en hadden we zelfs een officiële naam: the Meta-Research Center at Tilburg University. Michèle Nuijten, Paulette Flore, Robbie van Aert, Chris Hartgerink, Hilde Augusteijn, en Marjan Bakker, samen waren wij het (t)error team, (on)geleid door Jelte en Marcel. Heel veel dank voor jullie bijdragen aan mijn artikelen en het schieten op de conceptversies ervan, voor het sparren over onze ideeën, en voor onze serieuze en hilarische discussies. Ik zal onze tweewekelijkse leesclub enorm missen: de bijpraatronde (Marcel's anekdotes!), en het (op de energie van een heel pak koekjes) bekritiseren van alle artikelen die er nu weer waren verschenen. Wat hebben we veel meegemaakt en veel gelachen. Een hoogtepunt was de APS conferentie in San Francisco in 2014 (Bob's party!), en de road trip die Michèle, Paulette, Robbie en ik na afloop door het westen van de Verenigde Staten maakten. Watervallen en confidence intervallen in Yosemite! Een ander hoogtepunt was Chicago 2016, waar we allemaal bij waren. We waren onderdeel van de meta-research invasie bij APS, en we ontdekten het beste sushi-restaurant van de wereld (toch, Robbie?). Het feit dat ik me zo thuis voelde in onze groep heeft zeer veel bijgedragen aan het plezier dat ik heb beleefd aan mijn promotietraject. Dank hiervoor, en voor alle mooie herinneringen.

Ook buiten onze onderzoeksgroep zijn er mensen die een belangrijke bijdrage aan mijn proefschrift hebben geleverd. Ik wil de reviewers van mijn artikelen bedanken voor hun constructieve commentaren, en de student-assistenten voor hun harde werk. Linda Dominguez-Alvarez, Elise Crompvoets, en How Hwee Ong, jullie hebben geweldig werk verricht, veel dank hiervoor.

Ook mijn kamergenoten wil ik in het bijzonder bedanken. Robert Hillen, Paulette Flore, en Michèle Nuijten, ruim vier jaar zaten we samen. Vier jaar waarin we alles deelden, van publicatie-extase tot liefdesverdriet tot bruiloft. Onze kamer in P toverden we om tot een lounge room met pooltafel, basketbalspel, een verstopte nespresso-tap, en een mooi uitzicht op George. Elke maandagochtend brachten we elkaar op de hoogte van elkaars leven, om daarna weer hard te gaan 'knallen'. Bedankt voor een geweldige tijd! Ik mis jullie nu al.

Onze afdelingssecretaresse Marieke Timmermans mag ook niet ontbreken in mijn dankwoord. Marieke, je was altijd bereid om te helpen met alles, zelfs nadat ik al weg was uit Tilburg. Bedankt voor alles! Ik blijf graag over je vakanties naar het Verenigd Koninkrijk horen. Ook wil ik hier alle andere collega's en oud-collega's van MTO noemen: veel dank voor de fijne sfeer in het departement!

Dan mijn paranimfen Marlous Agterberg en Anna Hoogenboom. Lieve Lous en Annina, het is al meer dan 22 jaar geleden dat we daar samen bij Latijn naar meneer Prins zaten te luisteren, en sinds die tijd horen wij gewoon bij elkaar. Dank dat jullie er altijd zijn, en altijd zijn gebleven. Jullie hebben me altijd in alles gesteund, en ook nu weer staan jullie (letterlijk) achter mij. Zonder jullie steun en begrip, en zonder de energie die ik kreeg van de gesprekken, etentjes, en koffietjes met jullie, was mijn boekje er nooit gekomen.

Mijn alleroudste en dierbare vriendin Marjolein van Zoelen wil ik ook in het bijzonder bedanken. Lieve Mar, onze jeugd in de Badhuislaan heeft mij meer dan wat dan ook gevormd. Je bent een essentieel onderdeel van mijn leven, en mijn grote voorbeeld in wilskracht en doorzettingsvermogen. Met jouw humor en levenshouding laat je zien dat alles mogelijk is.

Mijn psychologievriendinnen Maaike Weber, Astrid Jehle, Floor de Groot, en Anna van der Horst moeten hier uiteraard ook genoemd worden. Lieve Maaike, toen ik jou leerde kennen tijdens de Bachelor begon ik mijn studie pas echt gezellig te vinden. Jij hebt mij het Amsterdamse studentenleven in getrokken, en hebt ervoor gezorgd dat ik ook in Amsterdam kon komen wonen. Ik ben zo blij dat we nog steeds veel contact hebben! Lieve Anna, Astrid, en Floor, wie had kunnen denken dat er uit een programma als Lisrel zo'n mooie vriendschap zou komen? Samen hebben we ons door de Research Master heen geslagen. Ik weet niet of ik het zonder jullie volgehouden zou hebben! Heel veel dank dat jullie er zijn.

Aan mijn ouders heb ik wat betreft het behalen van mijn doctoraat misschien wel het meest te danken. Mijn lieve papa en mama, zonder jullie onuitputtelijke steun en geduld was dit proefschrift er nooit gekomen. Altijd hebben jullie mij de vrijheid gegeven om geheel mijn eigen keuzes te maken. Dat ik deze vrijheid zo optimaal zou gebruiken hadden jullie misschien niet helemaal voorzien, maar ik ben er ontzettend dankbaar voor. Jullie gunden mij mijn wilde jaren in Australië, en steunden daarna mijn keuze om paarden te gaan bestuderen in Wales. Toen ik uiteindelijk tot inzicht kwam dat dit het toch niet zou worden, bleven jullie mij herkansingen bieden. Mijn hart bleek uiteindelijk bij de psychologie te liggen, iets wat jullie misschien stiekem altijd al wisten. Papa, vroeger vertelde ik stoer aan mijn vriendinnen dat mijn vader 'statisticus' was, ook al had ik geen idee wat dat was. Wie had ooit gedacht dat ik uiteindelijk zou promoveren in de Methodologie en Statistiek?

De andere twee leden van ons gezin hebben ook indirect veel bijgedragen. Lieve broer en zus, wat is het toch heerlijk om jullie in mijn leven te hebben. Met jullie kan ik over alles praten, we begrijpen elkaar compleet. Juist in drukke periodes geniet ik er altijd extra van om tijd met jullie door te brengen; bij jullie kom ik meteen tot rust en kan ik daarna weer de hele wereld aan!

Als laatste wil ik mijn fantastische echtgenoot en dochter bedanken. Lieve Sven, jouw steun is cruciaal geweest voor de totstandkoming van mijn proefschrift. Als eindeloos geduldige R-goeroe heb je mij door de steile learning curve van het programmeren gesleept, terwijl ik daar niet altijd even lief bij bleef. Het was echter niet alleen je praktische steun die zo waardevol was. Gedurende de hele vijf jaar wist je me altijd op de juiste momenten te motiveren, en op de juiste momenten te helpen bepaalde zaken te relativeren. Op de momenten dat ik twijfelde of ik het allemaal wel kon, deed jij me weer in mezelf geloven. De uiteindelijke afronding van mijn project was niet mogelijk geweest als jij niet op zo'n geweldige manier ons gezinnetje en huishouden draaiende had gehouden in de weken waarin ik dag en nacht aan het werk was. Ik kan niet uitdrukken hoe dankbaar ik voor je ben. Lieve Nina, wat ben ik blij dat je er bent. Jij hebt mijn PhD-periode tot de meeste bijzondere tijd van mijn leven gemaakt, en mij geleerd waar het leven om draait.