Preregistration in psychology:

Past, present, and prospects

A PhD-dissertation by
Olmo R. van den Akker (Tilburg University)

ISBN: 978-94-6469-776-6

Cover design: Jan Postma

Preregistration in psychology:

Past, present, and prospects

Proefschrift

ter verkrijging van de graad van doctor aan Tilburg University, op gezag van de rector magnificus, prof. dr. W. B. H. J. van de Donk, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de Aula van de Universiteit op

vrijdag 22 maart om 10:00 uur

door Olmo Robin van den Akker

geboren te Amsterdam

Promotores:	prof. dr. J. M. Wicherts (Tilburg University)
	prof. dr. M. A. L. M. van Assen (Tilburg University)
Copromotor:	dr. M. Bakker (Tilburg University)
Leden promotiecommissie:	prof. dr. K. Sijtsma (Tilburg University)
	prof. dr. C. D. Chambers (Cardiff University)
	dr. R. Hoekstra (University of Groningen)
	dr. I. A. Cristea (University of Padova)

This work was supported by a Consolidator Grant (IMPROVE) from the European Research Council (ERC; grant no. 726361).

Table of Contents

CHAPTER 1

Introduction

This dissertation revolves around preregistration, the practice where researchers publish their hypotheses, study design and/or analysis plan before collecting or analyzing their data. While this practice has been suggested as a useful tool for researchers as early as the 1950s and 1960s (Bakan, 1966; De Groot, 1956/2014; 1969), it has only started to become common in the 2000s in biomedicine and in the 2010s in psychology. The main trigger in biomedicine was the awareness that the results of clinical trials were often not reported, resulting in a biased literature (Dickersin & Rennie, 2003). To prevent this so-called publication bias, public registries for clinical trials (most notably https://clinicaltrials.gov) were set up, and a registration mandate (DeAngelis et al., 2004) was initiated. Preregistration (or registration, as it is known in biomedicine, see Rice and Moher, 2019) in public registries is important because it provides a paper trail of all clinical trials being conducted, making it possible to assess whether the set of published clinical trials is a good representation of the set of conducted clinical trials (Serghiou et al., 2023).

In psychology, the main trigger for the advent of preregistration was that many important findings could not be found in newer studies that used the same research designs (i.e., could not be replicated, e.g., Hagger et al., 2016; Klein et al., 2014; Open Science Collaboration, 2015). This led to the so-called replication crisis, or crisis of confidence, a state of uncertainty about what findings in the field were true and what findings were not (Pashler & Wagenmakers, 2012; Baker, 2016). This state of uncertainty induced many researchers in psychology to reflect on the scientific practices in the field, which helped identify causal factors for the replication crisis. Broadly speaking, these factors include the widespread use of suboptimal research designs (e.g., with a small number of observations), statistical misinterpretations, questionable research practices in analyzing data and presenting results, and perverse incentive structures in academia that aggravate these problems (Spellman, 2015). Among many suggested solutions (for an overview, see Munafo et al., 2017; Nosek, Spies, & Motyl, 2012) was preregistration (Nosek et al., 2018; Wagenmakers et al., 2012). According to Hardwicke and Wagenmakers (2023) preregistration mainly aims to (1) increase transparency and (2) reduce bias. The second aim mainly functions through preventing questionable research practices, or QRPs.

Two of the most prominent QRPs that preregistration aims to prevent are Hypothesizing After the Results are Known (HARKing, Kerr, 1998) and *p*-hacking, as discussed in Chapter 6. HARKing occurs when researchers attribute their research results to a specific hypothesis *after* analyzing the data. This practice is problematic because researchers will often find coincidentally statistically significant associations in datasets with many variables. On the other hand, *p*-hacking involves researchers making decisions contingent on the data to achieve a *p*-value below 0.05, artificially generating positive results. However, conclusions arrived at through HARKing or QRPs may not be warranted by the

data itself, as the association between variables might not truly exist or may be smaller than before *p*-hacking (Murphy & Aguinis, 2019; Simmons, Nelson, & Simonsohn, 2011).

Preregistration serves as a solution to counter HARKing since it requires researchers to publish their study's hypotheses before collecting or analyzing data. This approach makes it impossible for researchers to pretend that they theorized the study results beforehand. Similarly, preregistration helps prevent *p*-hacking by demanding researchers to explicitly specify research decisions before data collection. This explication curtails their freedom to make these decisions contingent on the data, ensuring greater transparency and minimizing the likelihood of biased results. The freedom researchers have to make decisions contingent on the data is often captured in the concept of "researcher degrees of freedom" (Simmons, Nelson, Simonsohn, 2011). The more decisions a researcher needs to make from the start of a project to its conclusion, the more researcher degrees of freedom a study is said to have and the more room there is for HARKing and *p*-hacking.

To what extent does preregistration prevent HARKing and *p*-hacking in practice? As discussed in Chapter 3, the effectiveness of preregistration in achieving these goals depends on at least two aspects. First, the *producibility* of the preregistration (i.e., the extent to which the study can be properly conducted based on the information in the preregistration) plays a vital role. A high level of producibility is desirable because it refers to the extent to which the provided information is sufficiently comprehensive helps prevent researchers from opportunistically exploiting their degrees of freedom. Second, it is crucial that the preregistration and the published study are consistent and hence that the study was conducted as outlined in the preregistration. When a preregistration only contains limited information or when researchers deviate significantly from the preregistered plan, the effectiveness of preregistration diminishes. In such cases, fewer degrees of freedom for researchers are restricted, creating more room for practices like *p*-hacking and HARKing. The *effectiveness* of preregistration thus requires both producibility *and* consistency. To maximize the benefits of preregistration, it is important to ensure that a preregistration contains comprehensive information and that researchers adhere closely to the preregistered plan throughout the study.

In Chapters 2 and 3, we investigated the effectiveness of preregistrations published in psychology between 2015 and 2019. To that end, we collected a sample of 459 preregistered studies that either won a Preregistration Challenge Prize or earned a Preregistration Badge. The Preregistration Challenge was an initiative by the Center of Open Science in which researchers could earn a monetary award if they published a study that was preregistered. Preregistration Badges are digital markers that are being used by scientific journals to show that a given paper within the journal involves at least one

preregistered study. This sample of 459 preregistered studies is used as the basis for the studies in Chapters 2, 3, and 4.

In Chapter 2, we assessed the consistency of more than 2,100 preregistered hypotheses. Specifically, we checked whether the hypotheses that were presented in preregistrations could be retrieved in the same form in the corresponding papers, or whether hypotheses were omitted, added, promoted from secondary to primary, demoted from primary to secondary, or changed in direction. If these forms of selective hypothesis reporting were not transparently disclosed, we categorized them as selectively reported. Based on our findings, we also discussed whether the hypotheses presented in preregistrations were reported sufficiently clearly (i.e., whether the preregistered hypotheses were producible enough).

In Chapter 3, we assessed both the producibility and the consistency of other study parts than hypotheses, most importantly the operationalization of the main variables, the data collection procedure, the statistical model, and the inference criteria. For this assessment, we only selected the 300 studies with consistent hypotheses in preregistration and paper. We determined whether these study parts were described in sufficient detail in the preregistration (producibility) and whether this description was consistent with the description in the corresponding paper (consistency).

Taken together, the results presented in Chapters 2 and 3 provide insight into whether preregistration can diminish the potential for QRPs in hypothesis-testing studies, as good scores on producibility and preregistration-paper consistency would mean there is less room for *p*-hacking or HARKing. However, as noted in Chapter 4, these results do not necessarily prove whether *p*-hacking took place in practice because the research process largely takes place behind closed doors. To make a reasonable judgment on whether preregistration prevents QRPs, we are therefore reliant on proxies of *p*-hacking. One of such proxies is the proportion of statistically significant, or positive, results, as such results are desirable and, therefore, often the end goal of *p*-hacking. As such, we expected a lower proportion of statistically significant results in preregistered studies than in non-preregistered studies.

In Chapter 4, we compared 193 preregistered studies to 193 non-preregistered studies to see whether the proportion of positive results is lower for preregistered studies. Of the original sample of 459 studies from 259 papers, we only included one study per paper, and we excluded all studies for which we could not find at least one statistical result associated with a preregistered hypothesis. The non-preregistered studies were selected to match the preregistered studies on topic and year of publication. We also tested the hypotheses that effect sizes would be smaller, sample sizes would be larger,

statistical inconsistencies would be less common, and power analyses would be more common in preregistered studies compared to non-preregistered studies. The results shed light on whether the potential benefits of preregistration are accrued in the practice of psychological science.

In Chapter 5, preregistration takes a less prominent role. The primary research goal was to assess how results of replication studies are interpreted by researchers. We did so by asking more than 1,800 psychology researchers to assess a range of fictional research scenarios in which they conducted four studies all testing a given theory. These scenarios varied (1) in the number of statistically significant results in the set of four studies, (2) in whether the studies were direct or conceptual replications, and (3) in whether the studies were preregistered or not. Participants then had to provide their belief in the theory for each scenario. Aside from assessing whether our variations influenced the researchers' belief in the theory, we also assessed how accurate researchers were when compared to a Bayesian inferential baseline.

Finally, in Chapters 6 and 7, we took a more practical approach. Many researchers find preregistration challenging, and we believe that we can facilitate the process by handing researchers tools to make their preregistrations more producible and, with that, more effective. To that end, we developed two preregistration templates: one for the preregistration of *secondary data analyses* (the use of existing data to answer a different question than the data was originally collected for), and one for the preregistration of *systematic reviews* of studies in the scientific literature. We developed the secondary data analysis template because the existing templates only focused on primary data analyses, and we developed the systematic review template because we felt the existing templates were too narrow in scope (e.g., by only focusing on interventional studies or health-related studies). The two templates are presented in these chapters, including guidance on how to best make use of them.

To summarize all the research outlined above, Chapter 8 discusses the current state of preregistration in psychological science. I will discuss whether the empirical data I found show that the potential of preregistration in preventing QRPs is achieved or whether there is room for improvement. I also embed my findings into the growing meta-research on preregistration and discuss whether several concerns that have been levied against the practice of preregistration are rooted in empirical data. In the end, I formulate the start of an answer to the question of whether preregistration is a valuable tool to potentially increase the trustworthiness and replicability of psychological science.

References

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

Baker, M. (2016). Reproducibility crisis. *Nature*, *533*(26), 353-66.

De Groot, A. D. (1969). Methodology: Foundations of inference and research in the behavioral sciences. The Hague: Mouton.

De Groot, A. D. (2014). The meaning of "significance" for different types of research (E. J. Wagenmakers, D. Borsboom, J. Verhagen, R. Kievit, M. Bakker, A. Cramer, D. Matzke, D. Mellenbergh, & H. L. J. Van der Maas, Trans.]. *Acta Psychologica*, *148*, 188-194. https://doi.org/10.1016/j.actpsy.2014.02.001 (Original work published in 1956)

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., ... & Van Der Weyden, M. B. (2004). Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *The Lancet*, *364*(9438), 911-912.

Dickersin, K., & Rennie, D. (2003). Registering clinical trials. *JAMA*, *290*(4), 516-523.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... & Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546-573.

Hardwicke, T. E., & Wagenmakers, E. J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour*, *7*(1), 15-26.

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

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology, 45*(3), 142-152.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., ... & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1-9.

Murphy, K. R., & Aguinis, H. (2022). HARKing: How badly can cherry-picking and question trolling produce bias in published results? In *Key Topics in Psychological Methods* (pp. 93-109). Cham: Springer Nature Switzerland.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615-631.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600-2606.

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

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.

Rice, D. B., & Moher, D. (2019). Curtailing the use of preregistration: A misused term. *Perspectives onPsychological Science*, 14(6), 1105-1108.

Serghiou, S., Axfors, C., & Ioannidis, J. P. (2023). Lessons learnt from registration of biomedical research. *Nature Human Behaviour*, 1-4.

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

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*(6), 886-899.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632-638.

CHAPTER 2

2

Selective hypothesis reporting in psychology: comparing preregistrations and corresponding publications

Olmo R. van den Akker[1], Marcel A. L. M. van Assen[1, 2], Manon Enting[1], Myrthe de Jonge[1], How Hwee Ong[3], Franziska Rüffer[1], Martijn Schoenmakers[1], Andrea H. Stoevenbelt[1, 4], Jelte M. Wicherts[1], Marjan Bakker[1]

[1] Department of Methodology and Statistics, Tilburg University, The Netherlands

[2] Department of Sociology, Utrecht University, The Netherlands

[3] Department of Social Psychology, Tilburg University, The Netherlands

[4] Department of Educational Science, University of Groningen, The Netherlands

This paper is published in Advances and Methods in Psychological Science.
See http://doi.org/10.1177/25152459231187988

Abstract

This study assesses the extent of selective hypothesis reporting in psychological research by comparing the hypotheses found in a set of 459 preregistrations to the hypotheses found in the corresponding papers. We found that more than half of the preregistered studies we assessed contained omitted hypotheses (N_s = 224; 52%) or added hypotheses (N_s = 227; 57%), and about one-fifth of studies contained hypotheses with a direction change (N_s = 79; 18%). We found only a small number of studies with hypotheses that were demoted from primary to secondary importance (N_s = 2; 1%) and no studies with hypotheses that were promoted from secondary to primary importance. In all, 60% of studies included at least one hypothesis in one or more of these categories, indicating a substantial bias in presenting and selecting hypotheses by researchers and/or reviewers/editors. Contrary to our expectations, we did not find sufficient evidence that added hypotheses and changed hypotheses were more likely to be statistically significant than non-selectively reported hypotheses. For the other types of selective hypothesis reporting, we likely did not have sufficient statistical power to test for a relationship with statistical significance. Finally, we found that replication studies were less likely to include selectively reported hypotheses than original studies. In all, selective hypothesis reporting is problematically common in psychological research. We urge researchers, reviewers, and editors to ensure that hypotheses outlined in preregistrations are clearly formulated and accurately presented in the corresponding papers.

Keywords: *hypotheses, bias, selective reporting, statistical significance, preregistration*

Introduction

Scientists should be open-minded and consider all new evidence, hypotheses, theories, and innovations when doing research, even those that challenge or contradict their own interests and beliefs (Anderson, 2000; Merton, 1973). However, scientists do not always abide by this Mertonian norm. Studies have shown that scientists regularly add, drop, or alter study elements when preparing reports for publication (Dwan et al., 2014; Dwan, Gamble, Williamson, & Kirkham, 2013; Mazzola & Deuling, 2013; O'Boyle, Banks, & Gonzalez-Mulé, 2017), a practice known as selective reporting (Cairo, Green, Forsyth, Behler, & Raldiris, 2020). For example, researchers may fail to report study results that are not statistically significant and thus 'not interesting' for publication (Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2004) or they may alter hypotheses after seeing the data to make their paper's narrative cleaner and more convincing (Giner-Sorolla, 2012; Kerr, 1998).

Selective reporting seems to be driven at least partly by a desire to publish work in prestigious selective journals (Van der Steen et al., 2018) and biases the scientific literature toward papers with publishable (often statistically significant) results. Indeed, statistically significant results are so abundant in the scientific literature that it is unlikely that the literature represents all research that has been conducted (Scheel, Schrijen, & Lakens, 2021; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995).

Selective reporting practices have been identified in many scientific fields, but studies on this issue have been especially prevalent in biomedicine (see DeVito, Bacon, & Goldacre, 2020; Thibault et al., 2021; Vinkers, et al., 2020). The reason for this is that clinical trials in this field are generally required to be registered in a formal and publicly accessible registry (DeAngelis et al., 2005; European Commission, 2012; Food and Drug Administration Amendments Act of 2007, 2018). This requirement enables comparing the registered protocol and the actual scientific publication to assess whether the authors of the publication changed, omitted, or added results, outcomes, or hypotheses. A systematic review of dozens of such meta-studies by Thibault et al. found that between 10% to 68% (95% prediction interval) of articles contain at least one primary outcome discrepancy.

The social sciences do not have an extensive registration infrastructure, so selective reporting has mainly been studied by comparing publications to dissertations (Cairo et al., 2020; Mazzola & Deuling, 2013; O'Boyle, et al., 2017) and archived research proposals (Franco, Malhotra, & Simonovits, 2016). Only a handful of studies compared publications to their corresponding preregistration, and all of them found that these publications often contained undisclosed deviations (psychology: Claesen, Gomes, Tuerlinckx, & Vanpaemel (2021); gambling: Heirene, et al., 2021; economics and political science: Ofosu

& Posner, 2021). In our study, we make use of the increased popularity of preregistration in psychological research in recent years (Hardwicke et al., 2022; Nosek & Lindsay, 2018) and check a large sample of preregistered psychology publications to assess the prevalence of one form of selective reporting: the selective reporting of hypotheses.

Selective hypothesis reporting can take on different types. We derived the terminology for these types from the biomedical literature, more specifically from Chan et al. (2004) and Thibault et al. (2021). One major difference between our study and earlier biomedical studies, though, is that we focus on hypotheses while biomedical studies typically focus on outcomes (i.e., dependent variables). This may be because outcomes take a prominent place in the clinicaltrials.gov registration template used for many clinical trials. In the current study, we distinguish five types of selective hypothesis reporting.

First, the number of hypotheses can change from the preregistration to the publication, which includes hypotheses that were present in the preregistration but did not appear in the publication (*omitted hypotheses*), and hypotheses that were not present in the preregistration but did appear in the publication (*added hypotheses*). Second, the status of hypotheses can change between the preregistration and the publication, which includes hypotheses that were labeled as primary in the publication but as secondary in the preregistration (*promoted hypotheses*), and hypotheses that were labeled as secondary in the publication but as primary in the preregistration (*demoted hypotheses*). Third, the direction of hypotheses (i.e., a positive, negative, null, or non-directional effect; or $A>B$, $A<B$, $A=B$, or $A\neq B$ when comparing groups) can change between the preregistration and the publication (*changed hypotheses*). Note that hypotheses can also differ in other ways between preregistration and paper. For example, sometimes authors alter the names of certain variables in the paper compared to the preregistration, or sometimes authors change the hypothesis from passive to active tense or vice versa. We do not consider such changes in this study because they only change a hypothesis superficially, rather than structurally. We thus use the adjective *changed* only for hypotheses with a direction change.

Note that the presence of statistical results related to added hypotheses in a publication is fine as long as they are labeled as exploratory (Logg & Dorison, 2021; Nosek, Ebersole, DeHaven, & Mellor, 2018). This is exemplified by the fact that both the CONSORT 2010 reporting guideline (Schulz, Altman, & Moher, 2010) and the JARS reporting guideline (Appelbaum et al., 2018) explicitly encourage the reporting of exploratory analyses. Readers will then know that the hypotheses were drawn up *a posteriori* and that using hypothesis tests to make statistical inferences may be invalid (Wagenmakers et al., 2012). However, if the results of added hypotheses are labeled as confirmatory or not labeled at all, readers are unaware of the exploratory nature of the hypotheses and may

inappropriately interpret the results using a hypothesis testing framework. In these instances, undisclosed and statistically uncontrolled explorations could unjustly be perceived as solid confirmatory evidence. In this study, we will therefore use the term *added hypotheses* only for non-preregistered hypotheses with statistical results that are labeled as confirmatory or not labeled at all.

We investigate the different forms of selective hypothesis reporting in psychological research by identifying hypotheses in our sample of preregistrations and the accompanying publications. We distinguish between hypotheses that are part of direct replications and hypotheses that are part of original studies because we believe selective hypothesis reporting to be less of an issue for the former than for the latter (Hypothesis 1). We also assess whether forms of selective hypothesis reporting are related to statistically significant results (Hypotheses 2a-2d). Our specific hypotheses and our rationale for these hypotheses are outlined below.

Hypotheses

We had no hypotheses about the exact proportion of studies involving selective hypothesis reporting, but we did expect that forms of selective hypothesis reporting would be less common among direct replication hypotheses than original hypotheses because direct replication hypotheses need to adhere (both in the preregistration and the publication) to the hypotheses outlined in the original study. We also expected some forms of selective hypothesis reporting to be associated with statistical significance because results that are statistically significant are more likely to be published than results that are not statistically significant (Kerr, 1998; Scheel, et al., 2021). Our hypotheses are listed more formally below and can also be found in our preregistration at https://osf.io/z4awv. Note that we originally uploaded our preregistration on OSF on 21 January 2021, before data collection. However, we formally entered it into the registry on 5 March 2023 to increase the findability of our preregistration (see https://osf.io/nxgtv). Aside from correcting the erroneous statement listed in footnote 3 we did not make any changes.

1) A hypothesis that is part of a direct replication is less likely to be selectively reported (omitted, promoted, demoted, or changed) than an original hypothesis
2a) The test result of an added hypothesis is more likely to be statistically significant than the test result of a preregistered hypothesis that is appropriately reported
2b) The test result of a promoted hypothesis is more likely to be statistically significant than the test result of a preregistered hypothesis that is appropriately reported
2c) The test result of a demoted hypothesis is less likely to be statistically significant than the test result of a preregistered hypothesis that is appropriately reported
2d) The test result of a changed hypothesis is more likely to be statistically significant than the test result of a preregistered hypothesis that is appropriately reported

Because the statistical significance of omitted hypotheses is unknown, we did not formulate a hypothesis on the association between omitted hypotheses and statistical significance.

Method

Sample

We used two main sources to find published preregistrations. First, we looked at published papers that earned a Preregistration Challenge prize. The Preregistration Challenge was an educational campaign organized by the Center for Open Science (COS) in 2017 and 2018 where researchers could earn $1,000 if they published a study that was preregistered using a specific preregistration template (see https://web.archive.org/web/20230305173237/https://www.cos.io/initiatives/prereg-more-information for more information). A full list of Preregistration Challenge prize winning papers ($N_p = 180$) can be found in the OSF Zotero Library at https://web.archive.org/web/20230305173614/https://www.zotero.org/groups/479248/osf/collections/D77RMN4N.

Second, we looked at published papers that earned a Preregistration Badge in 2019 or before as part of the COS' Open Science Badges initiative (see https://web.archive.org/web/20230418120332/https://www.cos.io/initiatives/badges). Papers can earn a Preregistration Badge if the authors provide the URL, DOI, or other permanent paths to the preregistration in a public, open access repository. We extracted 244 papers that earned a Preregistration Badge from a database with all papers that earned an Open Science Badge up until 21 February 2020 (Kambouris et al., 2020). After deleting these duplicate papers, the total number of papers in our sample was 180 + 193 + 51 − 26 = 398.

To assess whether these papers were from the field of psychology we looked up their Research Areas as listed in the Web of Science Core Collection. If the paper was not listed in that database, we categorized the Research Area ourselves based on the publishing journal or the departmental affiliation of the authors. The papers in our sample often contained multiple preregistered studies. We considered a study separate from other studies in a paper when that study was based on a different sample of participants. Each of these studies was coded separately. For the 329 psychology papers we included we derived 613 preregistered studies.

Of these 613 preregistered studies, we omitted 48 studies because they were conducted in a registered report framework (where the studies are peer-reviewed before data col-

lection), 52 studies because they were part of a multi-lab paper that did not focus on the individual studies but only on the bigger picture (e.g., Many Labs 2, Klein et al., 2018), five studies using non-human subjects, 14 studies because we were unable to locate a preregistration, and 14 studies because it was unclear which study was described in which (part of the) preregistration. Finally, we excluded 21 studies with preregistrations of secondary data analyses (i.e., data that already existed and were gathered to answer another research question from the one in the study), because such preregistrations qualitatively differ from those using primary data (Weston et al., 2019; Van den Akker et al., 2021) and would therefore have required different coding procedures. All exclusions left us with a final sample of 459 studies from 259 papers, yielding an average of 1.8 studies per paper. Screening for eligible studies was done by the first author before coding started, although 25 exclusions (5% of the total) were made during coding. These later exclusions were made by the first author following advice from coders who noticed that a certain study did not match the inclusion criteria after all. A PRISMA flow diagram (Moher, Liberati, Tetzlaff, Altman, & PRISMA Group, 2009) outlining the full sample selection procedure (including exclusions during coding) can be found in Figure 1.

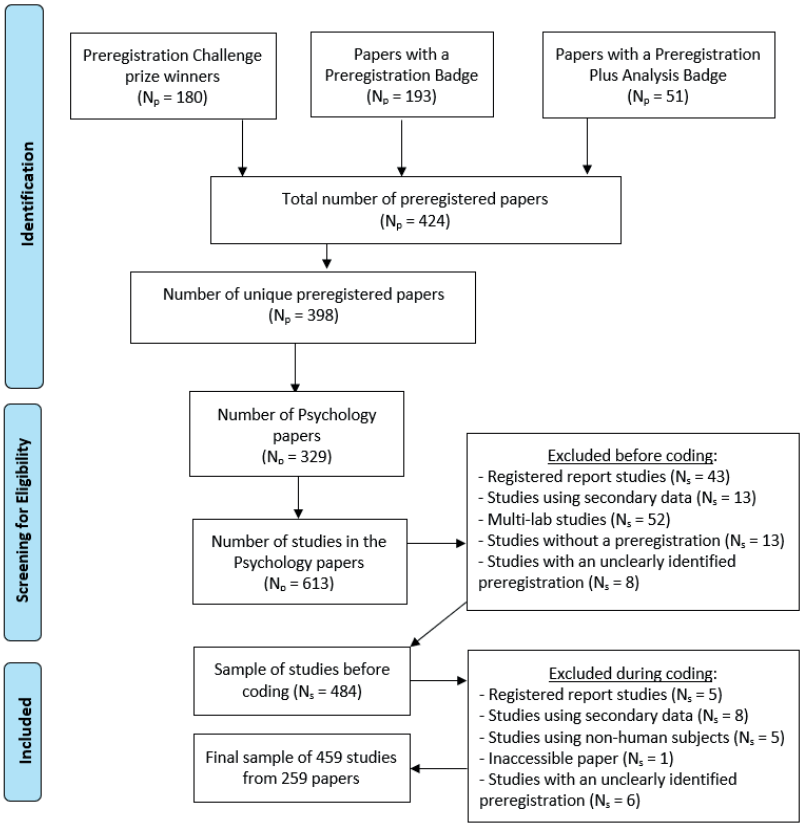


Figure 1. PRISMA flow diagram outlining the full sample selection procedure

Identifying hypotheses

Because we could not find a validated procedure to systematically and manually extract hypotheses from scientific papers[1], we developed two new Qualtrics protocols: one for preregistrations (https://osf.io/fdmx4), and one for their accompanying publications (https://osf.io/uyrds). These protocols were created after a series of meetings (involving OvdA, MvA, MB, and JW) and a series of pilots using papers not included in the eventual sample (involving all authors except for JW). The protocols were preregistered before data collection. Coding was carried out by all authors except JW and consisted of four phases: (1) two coders independently identified hypotheses in the preregistration, (2) the coders discussed any inconsistencies in their coding and resolved these together, (3) the same two coders independently identified hypotheses in the publication, (4) the coders discussed any inconsistencies in their coding and resolved these together. Coders were trained before coding by the first author who instructed them about the protocol and assessed how they coded a trial run. The first author provided guidance throughout this trial run until both the first author and the coder were satisfied about the coders' grasp of the protocol.

We identified hypotheses in preregistrations and publications by first checking if any hypotheses were listed in a separate section. If not, we searched the running text for the following keywords (chosen based on Scheel et al. (2021)'s analysis of hypothesis introduction phrases): "replicat", "hypothes", "investigat", "test", "predict", "examin", and "expect". We included a hypothesis if the authors hypothesized a relationship between two or more variables using any of these keywords.

If we found a hypothesis that was phrased in a conceptual way (e.g., "we expect an association between extraversion and IQ") as well as an operational way (e.g., "we expect an association between scores on the Multidimensional Introversion-Extraversion Scale and scores on the Wechsler Intelligence Scale for Children") we only counted the more specific operational hypothesis because we did not want to count equivalent hypotheses twice. Moreover, we reasoned that it would be easier to identify operational hypotheses in scientific papers than it would be to identify conceptual hypotheses. If we found multiple operational hypotheses (e.g., one using the Wechsler Intelligence Scale for Children *and* one using the RAKIT Intelligence Test) we counted each one as a different hypothesis. Because there could be additional measures in other sections than the section in which we found the hypothesis (e.g., in the methods/measures/variables section), we checked the entire preregistration for additional measures. The same principle holds for additional control variables (e.g., in the variables/analysis section) so we checked the other sections for control variables as well.

1 For procedures to extract outcomes from biomedical papers, see Chan et al. (2004) and Thibault et al. (2021).

To investigate whether hypotheses were omitted, we used two approaches. In the first approach, we checked whether the preregistered hypothesis was referred to as a hypothesis in the *introduction or methods section* of the paper and, if so, concluded that the hypothesis was not omitted. In the second approach, we checked whether we could find a statistical result related to the preregistered hypothesis in the *results section* of the paper and if so, concluded that the hypothesis was not omitted. In this second approach, the result should have been reported in the main text, not tucked away in a Table, Figure, or Appendix. We decided to be strict in this regard because we believe that testing the preregistered hypotheses is the reason for conducting the confirmatory study in the first place, and as such we believe all of them should be mentioned in the main body of the paper. We include both the first and second approach when we present statistics about the prevalence of selective hypothesis reporting, but we only use hypotheses omitted from the results sections for our hypothesis tests.

Of the preregistered hypotheses identified as hypotheses somewhere in the paper, we checked whether they were labeled as equally important as in the preregistration. To this end, we used the keywords "key", "leading", "main", "major", "primary", and "principle" for *primary hypotheses*, and "additional", "auxiliary", "minor", and "secondary" for *secondary hypotheses*. If none of these words could be associated with the hypothesis, we categorized its importance as *non-specified*. We had to rely on these keywords because, unlike study outcomes in biomedicine, hypotheses are typically not labeled as primary or secondary in psychology. We assessed the directionality of hypotheses in the preregistrations and papers by giving coders both a concrete indication (*directional*: "men will score higher on the Verbal Aggression Scale than women"; *non-directional*: "men and women score differently on the Verbal Aggression Scale"; *null*: "men and women will not differ in their scores on the Verbal Aggression Scale") and an abstract indication (directional: "M > W"; non-directional: "M=/= W"; null: "M = W") of what to look for. We also assessed whether these categorizations were consistent between the paper and the preregistration. These assessments gave us the necessary information to establish the prevalence of promoted hypotheses (H2b), demoted hypotheses (H2c), and changed hypotheses (H2d).

Because several coders indicated they were unsure about their responses related to the directionality of the hypotheses, the first author manually checked (and corrected) all hypotheses for which the directionality was originally coded as inconsistent between preregistration and paper. The corrections can be found in the Excel-file with the data, see the ManualChanges columns in https://osf.io/8y2dv, and were discussed with and accepted by the original coders.

We also assessed how many statistical results were presented in the paper that were not related to a preregistered hypothesis and not explicitly stated as exploratory or non-preregistered. Such added hypotheses (H2a) should involve a different relationship between the variables than in a preregistered hypothesis or involve a different variable or measure altogether, and should be reported in the main text, rather than being tucked away in a Table, Figure, or Appendix. In our assessment of added hypotheses, we only included studies with at least one preregistered hypothesis because we inadvertently failed to present the coders with questions about added hypotheses for studies with zero hypotheses in Qualtrics.

Because of time constraints, we only assessed selective hypothesis reporting for the first sixteen preregistered hypotheses of a study, even if more than sixteen preregistered hypotheses were identified. In those instances, we also did not check for added hypotheses. Finally, note that the categories omitted, added, promoted, and demoted hypotheses are mutually exclusive but not exhaustive categories. In the present paper, we state that a study does not include selective hypothesis reporting when it does not include any omitted, added, promoted, demoted, or changed hypotheses.

Assessing whether a hypothesis is part of a direct replication

We operationalized the replication status of hypotheses (see Hypothesis 1) in three ways. In line with Scheel et al. (2021), we assessed whether a hypothesis was part of a replication study or an original study by searching the preregistration and paper for the string "replic" and assessing whether the authors referred to the hypothesis as being part of a replication attempt. If they did, in either the preregistration or the paper, we coded the hypothesis as a *replication hypothesis*. If they did not, we coded the hypothesis as an *original hypothesis*.

Second, we checked the papers to see whether hypotheses were part of a *direct replication* or *conceptual replication*. We coded hypotheses as part of a direct replication when the authors used the same methods (materials *and* procedure) to test the hypothesis as in a prior study. The methods had to be truly identical except that the replication study used a different sample and except for any translations of study materials. If the methods were not identical in this way, we coded the hypothesis as part of a conceptual replication.

Third, we logged the way the authors themselves labeled the hypotheses in the papers and coded hypotheses as part of a direct replication if the authors referred to them using any of the following words: "direct", "directly", "exact", "exactly", "identical", or "direct & very close" and as part of a conceptual replication if the authors referred to them using other words (e.g., "conceptual", "similar, except", "close").

We preregistered (see https://osf.io/z4awv) that we would use the second operationalization of replication status to test Hypothesis 1 if more than 20% of the replication hypotheses found in the papers were categorized as direct as opposed to conceptual. However, direct replication hypotheses constituted only 19.4% of the replication hypotheses. As preregistered, we therefore used the first operationalization of replication status for the main test of Hypothesis 1 and used the second and third operationalizations as robustness checks.

Assessing whether a hypothesis is supported

For every preregistered hypothesis for which we found a statistical result, we coded whether the result was statistically significant or not (see Hypotheses 2a-2d). We did this by comparing the reported *p*-value to .05 unless the authors specifically mentioned that they used a significance level lower than .05 (e.g., because they used a Bonferroni correction). In case of the latter, we concluded that the result was significant if the *p*-value was smaller than the authors' significance level. If the authors reported a Bayes Factor instead of a *p*-value, we concluded that the hypothesis was supported if the Bayes Factor was larger than 3. We used a threshold value of 3 because it has long been used as the value above which evidence for a hypothesis is deemed substantial (Jeffreys, 1961)[2]. If authors specifically mentioned that they used another Bayes Factor threshold than 3, we concluded that the hypothesis was supported if the Bayes Factor was larger than the authors' Bayes Factor. In light of our hypothesis tests, we consider a supported Bayesian hypothesis as equivalent to a statistically significant result.

Results

Descriptive statistics

We identified 2,119 hypotheses in 459 preregistered studies from 259 papers. The number of hypotheses per study (paper) is thus 4.6 (8.2), with 30 studies with zero hypotheses and 29 studies with more than 16 hypotheses. When two coders counted and coded the number of hypotheses in a preregistration, they agreed about the number of hypotheses in only 53.7% of the cases. With regard to assessing study difficulty, we found medium consistency between coders: Kendall's tau = 0.21, $z = 5.03$, $p < .001$.

Of all hypotheses identified in the preregistrations, we categorized 455 (21.5%) as part of a replication and 1,664 (78.5%) as 'original'. Of all hypotheses identified in the papers, we categorized 143 (6.7%) as part of a direct replication, 595 (28.1%) as part of a conceptual

2 Technically, the threshold value proposed by Jeffreys was $10^{1/2} \approx 3.16$, but it was later rounded to 3 to make statistical inference easier (see Jarosz & Wiley, 2014; Wetzels et al., 2011).

replication, and 1,381 (65.2%) as 'original'. The proportion of direct replications we found for preregistered studies (6.7%) is higher than estimates for non-preregistered psychology studies, which range from 1.1% (Makel, Plucker, & Hegarty, 2012) to 2.6% (Scheel, et al., 2021). At the same time, our estimate is substantially lower than the 57.8% estimate for registered reports (Scheel, et al., 2021). In all, it appears that preregistered studies are more likely to be replications than non-preregistered studies are.

The vast majority of hypotheses (N_h = 1,475; 69.6%) concerned associations/effects between two variables. The other hypothesis types were less common: interaction/moderation (N_h = 326; 15.4%), mediation (N_h = 87; 4.1%), univariate (N_h = 57; 2.7%), and other (N_h = 174; 8.2%). In the 'other' category we placed hypotheses that did not fit any of the types, like predictions indicating atypical or complex relationships between variables (e.g., curvilinear associations or three-way interactions). Comparing hypothesis types between independent samples of preregistered and non-preregistered studies could be an interesting follow-up project, especially if one would focus on the complexity or riskiness of hypotheses. Pham and Oh (2021) argued that the prestige premium of preregistration may result in "a bias toward studies that are easy to preregister […] and a preference for research hypotheses that are obvious." In contrast, Scheel et al. (2021) proposed that researchers may deliberately preregister risky hypotheses because the negative effects of getting a small or negative result may be compensated by the credence received from preregistration.

Using our two approaches to assess omitted hypotheses, we were able to retrieve 1,143 of 2,119 preregistered hypotheses (53.9%) in the introduction or methods sections and 1,132 results of 2,119 preregistered hypotheses (53.4%) in the results section. Consequently, 976 hypotheses were missing from the introduction and methods sections (46.1%) and 987 hypothesis results were missing from the results section (46.6%). Of the 1,132 results we found in the results section, 743 (65.6%) were statistically significant, and 389 (34.4%) were not. The number of omitted hypotheses per study and per paper can be found in Table 1, as is the case for the other forms of selective hypothesis reporting we discuss below. The proportion of omitted hypotheses in the results (46.6%) is somewhat higher than earlier estimates by Ofosu and Posner (2021) and Heirene et al. (2021) who found that a little over one-third of studies included omitted hypotheses. Based on a meta-analysis of 89 studies from mainly biomedicine, Thibault et al. (2021) estimated that 6-16% (95% CI) of studies contain at least one omitted primary outcome, and 14-62% (95% CI) of studies contain at least one omitted secondary outcome. The results from this meta-analysis, which we believe is the most recent and comprehensive assessment of selective outcome reporting in biomedicine to date, are comparable to our results as shown in Table 1.

Of the 401 studies with at least one and at most sixteen hypotheses we counted the number of added hypotheses (i.e., non-preregistered statistical results). In studies with at least one added hypothesis (N_s = 227; 56.8%) the total number of added hypotheses was 1,634. The mean number of added hypotheses per study was 4.09 (see also Table 1) and the median number of added hypotheses per study was 1. The maximum number of added hypotheses in a single study was 48. Ofosu and Posner (2021) found that 18% of the studies in their sample included added hypotheses, of which 82% failed to mention that they were non-preregistered, possibly suggesting that adding hypotheses is more common in psychology than in economics and political science. In their meta-analysis, Thibault et al. (2021) found that the number of studies with added primary outcomes in biomedicine (95% CI: 7-14%) was somewhat lower than Ofosu and Posner's estimate of 18%. The number of studies with added secondary outcomes was found to be 8-80% (95% CI), which is consistent with the estimate of both Ofosu and Posner as well as our estimate of 56.8%.

From all preregistered primary hypotheses that were not omitted in the paper (N_h = 329), we found that 52 (15.8%) were primary in both the preregistration and the paper, 14 (4.3%) were demoted from primary to secondary, while the primacy of 263 hypotheses (80.0%) was not specified in the paper. From all preregistered secondary hypotheses that were not omitted in the paper (N_h = 54), we found that 21 (38.9%) were secondary in both the preregistration and the paper, none were promoted from secondary to primary, and the importance of 33 (61.1%) was not specified in the paper. A visual depiction of the hypotheses with a change in importance between preregistration and paper can be found in Figure 2. Allocating the label of 'primary' to one or more hypotheses was done in 151 out of 429 studies (35.2%). This practice appears to be less common in psychological research than in biomedical research (78.7% of studies; Thibault et al., 2021), where the prevalence of promoted (95% CI: 3-9%) and demoted (95% CI: 7-18%) hypotheses seems to be higher. Psychological researchers may do well to take up the distinction between primary and secondary hypotheses as Ofosu and Posner (2021) posit that this distinction may help to prevent researchers from determining a hypothesis' importance post-hoc based on statistical significance.

Finally, we assessed the number of changed hypotheses. Of the 958 preregistered directional hypotheses that were not omitted in the paper, 882 had the same direction in the paper (92.1%), 4 had a different direction (0.4%), 69 (7.2%) became nondirectional, and three became null hypotheses (0.3%). Of the 151 preregistered nondirectional hypotheses not omitted in the paper, 131 remained nondirectional in the paper (86.8%), 20 became directional (13.2%), while 0 became null hypotheses. Of the 65 preregistered null hypotheses not omitted in the paper, 49 remained null in the paper (75.4%), 14 became directional (21.5%), and 2 became nondirectional (3.1%). A visual depiction of

the hypotheses with a change in directionality between preregistration and paper can be found in Figure 2. In sum, the vast majority of hypotheses did not involve a change in direction from preregistration to paper, a result mimicked by Cairo et al. (2020) who found that the direction of only 3.4% of social psychology hypotheses changed from dissertations to published papers.

When we excluded, per our preregistration, studies that were classified as 'very difficult' by the coders (N_s = 73; 17.09%), the degree of selective hypothesis reporting decreased slightly compared to our results from the whole sample. However, it is still substantial (i.e., around 50% of the studies have omitted hypotheses and added hypotheses, and around 20% of studies have changed hypotheses, see https://osf.io/geuxv for the full results excluding very difficult studies).

Table 1

An overview of the prevalence of the different forms of selective hypothesis reporting

	Percentage of studies (N_s=429*)	Percentage of papers (N_p=259*)	Average number per study	Average number per paper
Selective hypothesis reporting	.60	67	2.48	4.12
Omitted hypotheses (introduction)	56	62	2.28	3.77
Omitted hypotheses (results)	52	61	2.30	3.81
Added hypotheses**	57	69	4.09	6.92
Promoted hypotheses***	0	0	0	0
Demoted hypotheses****	1	2	0.09	0.16
Changed hypotheses	18	12	0.26	0.43

Notes. * indicates the number of studies/papers with at least one preregistered hypothesis. ** indicates that the proportions are calculated using a denominator with the number of studies (N_s = 400) and the number of papers (N_p= 236) with at least one preregistered and at most 16 preregistered hypotheses). *** indicates that the proportions are calculated using a denominator with the number of studies (N_s = 61) and the number of papers (N_p = 44) with at least one secondary hypothesis. **** indicates that the proportions are calculated using a denominator with the number of studies (N_s = 151) and the number of papers (N_p = 87) with at least one primary hypothesis.

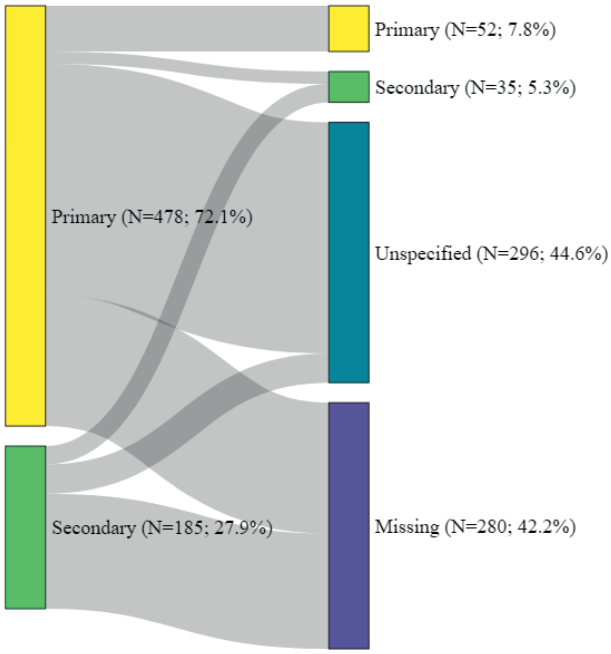


Figure 2. Sankey diagram indicating how primary and secondary hypotheses changed from preregistration (left) to paper (right)

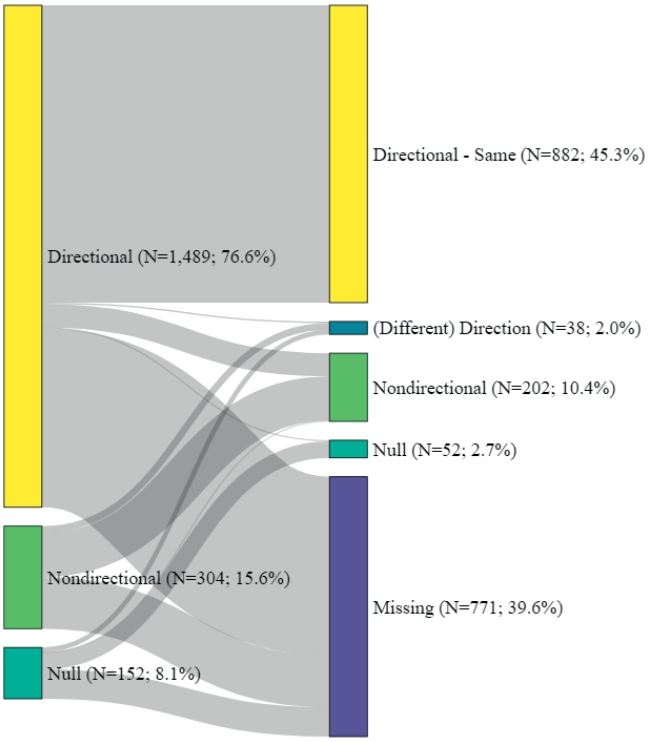


Figure 3. Sankey diagram indicating how the directionality of hypotheses changed from preregistration (left) to paper (right)

Selective hypothesis reporting and replication status (H1)

To test whether selective hypothesis reporting is more common for replication hypotheses than for original hypotheses (H1), we employed a multilevel logistic regression with hypothesis as Level 1, and study as Level 2. The regression includes a binary dependent variable indicating whether a hypothesis is selectively reported in the publication (i.e., omitted, promoted, demoted, and/or changed), and a binary independent variable on Level 1 indicating whether a hypothesis is part of a replication. We tested Hypothesis 1 against alpha = .05, as preregistered. The results indicate that hypotheses that are not part of a replication were more than twice less likely to be selectively reported than hypotheses that were part of a replication, β_1 = -0.92, z = -4.08, OR = 0.40, 95% CI = [0.25, 0.62], p = .00005 (see Model 1 in Table 2 for the complete regression output). This supports our preregistered Hypothesis 1[3].

As preregistered robustness checks, we ran two additional models. In the first model, we coded a hypothesis as part of a replication if the coders identified the hypothesis as a part of a direct replication based on the information in the paper only (6.7% of the 2,119 hypotheses described in the paper: β_1 = -0.92, z = -1.94, OR = 0.40, 95% CI = [0.16, 1.01], p = .052). In the second model, we coded a hypothesis as part of a direct replication if the authors themselves labeled the hypothesis as part of a "direct", "exact", "identical", or "(very) close" replication (4.0% of the 2,119 hypotheses: β_1 = -1.26, z = -2.46, OR = 0.30, 95% CI = [0.10, 0.77], p = .014). The robustness checks showed mixed results when strictly looking at statistical significance , but the odds ratios were similar to or more extreme than the odds ratio from our main preregistered hypothesis. We therefore give precedence to the main analysis and conclude that hypotheses that are part of a replication are less often selectively reported than hypotheses that are not part of a replication. This constitutes new knowledge as earlier studies assessing selective hypothesis reporting in the social sciences did not consider replication status.

Exploratively, we also compared whether studies in our sample that won a Preregistration Challenge prize (N = 141) and studies in our sample that earned a Preregistration Badge (N = 305) differed in the degree of selective hypothesis reporting. For this analysis we excluded studies with both a Preregistration Challenge prize and a Preregistration Badge. We ran a multilevel model with study type (Challenge vs. Badge) as the independent variable, and selective hypothesis reporting (the same variable as used in Model 1) as the dependent variable. We found that studies with a Preregistration Challenge prize less often involved selective hypothesis reporting (42%) than studies that earned a Preregistration Badge (54%), β_1 = -0.97, z = -3.32, OR = 0.38, 95% CI = [0.21, 0.67], p = .001.

3 Note that we omitted paper as Level 3, as preregistered, because the model including that level did not converge. Moreover, the preregistration incorrectly stated that an odds ratio < 1 indicates more selective hypothesis reporting instead of less (see Version 3 at https://osf.io/z4awv).

This difference may have come about because the Preregistration Challenge required researchers to fill out a detailed preregistration template whereas there was no such requirement to earn a Preregistration Badge. A detailed template could have prompted researchers to more clearly lay out their hypotheses, which could in turn have increased researcher's sense of urgency in being consistent with their hypotheses. Alternatively, it could be that the researchers who participated in the Preregistration Challenge differed from those who earned a Preregistration Badge. For example, perhaps because of the added effort of filling out the template they could have been more motivated to preregister well and subsequently adhere to their preregistration. These speculative explanations would need to be tested in a confirmatory study.

Table 2

Results of the Multilevel Regression Models Testing Hypothesis 1 (Model 1) and Hypothesis 2a and 2d (Model 2)

Parameters	Model 1	Model 2
Regression coefficients (fixed effects)		
Intercept	-0.15 (0.17)	0.53 (0.18) **
Level 1		
Replication	-0.92 (0.23) **	-
Added	-	0.75 (0.23) **
Changed	-	-0.23 (0.38)
Variance components (random effects)		
Study-level	4.87 (2.21)	1.73 (1.32)

Notes. Standard errors are in parentheses. Replication is a binary variable that takes on the value of 1 if the hypothesis was scored as part of a replication in either the preregistration or the paper, and 0 otherwise. Added is a binary variable that takes on the value of 1 if the study including the preregistered hypothesis had added hypotheses, and 0 if not. Changed is a binary variable that takes on the value of 1 if the preregistered hypothesis had a direction change from preregistration to paper, and 0 if not. * $p < .05$, ** $p < .01$

Selective hypothesis reporting and statistical significance (H2a, H2b, H2c, H2d)

As tests of our Hypotheses 2a, 2b, 2c, and 2d we preregistered a multilevel logistic regression with hypothesis as Level 1, study as Level 2, and paper as Level 3. The regression would include a binary dependent variable indicating whether the result is statistically significant, and four Level 1 binary variables, each indicating whether a hypothesis is selectively reported in a certain way: added hypotheses (H2a), promoted hypotheses (H2b), demoted hypotheses (H2c), and changed hypotheses (H2d). We had to omit 'promoted hypotheses' from our model as we did not encounter these. The remaining model did not converge when we included Level 3 or when we included demoted hypotheses. Therefore, we adjusted our model to a 2-level model that could only test H2a

and H2d. We had preregistered the conditional move to a 2-level model but dropping the promoted and demoted hypotheses was unforeseen and thus non-preregistered. We tested Hypotheses 2a and 2d against alpha = .01, as was preregistered. We found that preregistered hypotheses in studies with added hypotheses were more likely to be statistically significant than preregistered hypotheses in studies without added hypotheses, $\beta_1 = 0.75$, $z = 3.18$, OR = 2.11, 99% CI = [1.15, 3.86], $p = .001$ (Hypothesis 2a; Model 2 in Table 2), but we did not find that changed hypotheses were more likely to be statistically significant than unchanged hypotheses[4], $\beta_2 = -0.23$, $z = -0.60$, OR = 0.77, 99% CI = [0.29, 2.05], $p = .547$ (Hypothesis 2d; Model 2 in Table 2).

In hindsight, we realized that our preregistered test regarding added hypotheses was not entirely in line with our Hypothesis 2a. While our test showed that studies with added hypotheses included more statistically significant preregistered hypotheses, we were more interested in whether added hypotheses themselves were more likely to be statistically significant than preregistered hypotheses. Therefore, we also tested this at the level of hypotheses rather than at the level of studies. Each study has a proportion of statistically significant preregistered hypotheses, *p*, and a proportion of statistically significant added hypotheses, *a*. We compared the means of these two sets of proportions using a non-preregistered dependent *t*-test. We found a statistically significant difference using an alpha of .05 but no statistically significant difference when using an alpha of .01, $M_{p\text{-}a} = -0.08$, $t(191), = -2.52$, $p = .013$, Cohen's $d = -0.18$. A non-parametric Wilcoxon rank sum test corroborated this result, $V = 3844$, $p = .017$. When comparing dissertations and journal articles, Cairo et al. (2020) found that supported hypotheses were not more likely to be added than unsupported hypotheses. For biomedicine, the Thibault et al. (2021) meta-analysis indicated that 49-66% (95% CI) of outcome discrepancies involved a statistically significant result. Taken together, the results are not clear cut about whether researchers in psychology and biomedicine add hypotheses primarily based on statistical significance. If there is an effect, it is most likely small.

As preregistered, we also ran our analyses without studies that we labeled as 'very difficult' to code ($N_s = 73$; 15.9%). We still found support for our Hypothesis 1 that hypotheses that are part of a replication are less likely to be selectively reported (omitted, promoted, demoted, or changed) than original hypotheses ($\beta_1 = -0.76$, $z = -2.95$, OR = 0.47, 95% CI = [0.28, 0.78], $p = .003$). The robustness analysis for Hypothesis 2a was not in line with the original analysis: preregistered hypotheses in studies with added hypotheses were not more likely to be statistically significant ($\beta_1 = 0.56$, $z = 2.35$, OR = 1.76, 99% CI = [0.95, 3.26], $p = .019$). The robustness analysis for Hypothesis 2d was in line with the

4 One way we deviated from our preregistration was by scoring hypotheses that changed from directional to null and from non-directional to null as 0 instead of 1 for the variable 'changed' because we would expect less significant results for such changes, not more.

original analysis: preregistered hypotheses that were changed were not more likely to be statistically significant (β_2 = -0.32, z = -0.82, OR = 0.72, 99% CI = [0.26, 1.99], p = .410). We conclude that there is inconclusive evidence with regard to Hypothesis 2a, and a robust lack of evidence in favor of Hypothesis 2d. For an overview of the results without studies that were very difficult to code, see https://osf.io/geuxv.

General Discussion

In this project, we assessed the prevalence of omitted, added, promoted, demoted, and changed hypotheses in psychological research. Moreover, we tested whether replication studies were more or less likely to involve these types of selective hypothesis reporting and whether these types of selective hypothesis reporting were associated with statistically significant results. We found that more than half of the preregistered studies we assessed contained omitted hypotheses (N_s = 224; 52%) or added hypotheses (N_s = 227; 57%), and about one fifth of studies contained hypotheses with a direction change (N_s = 79; 18%). Additionally, we found only a small number of studies with demoted hypotheses (N_s = 2; 1%) and no promoted hypotheses. Replication studies were less likely to include selectively reported hypotheses than original studies, but we did not find that added and changed hypotheses were more likely to be statistically significant. We were not able to test whether promoted and demoted hypotheses were associated with statistical significance because of the low prevalence of such hypotheses.

When interpreting these results, it is important to consider the particularities of the coding protocol we used. One consideration is that we limited our sample to studies from the Preregistration Challenge and studies that earned a Preregistration Badge. This selection could have negatively impacted the representativeness of our results, but we feel that our sample is sufficiently in line with the wider population of preregistered studies. The Preregistration Challenge and the Preregistration Badge initiatives are very well-known in the psychological science community and have fundamentally changed the preregistration infrastructure. Preregistration badges are handed out by a large variety of psychology journals, including important journals in the field like Psychological Science, Advances in Methods and Practices of Psychological Science, Psychological Methods, and the Journal of Experimental Social Psychology. Similarly, the Preregistration Challenge winners included papers published in a wide range of scientific journals and was paramount in the increased popularity of preregistration we see now (Pennington, 2023). Moreover, our sample of 459 studies is the largest to date with regard to both quantity and time range. Consequently, our conclusions about (the quality of) preregistrations relate to most of the population of preregistrations in psychology.

That being said, there are undoubtedly preregistrations that we overlooked by selecting our sample based on these two sources. How this could have influenced our results is hard to say, but we contend that these 'hidden' preregistrations might be of lower quality than the preregistrations we did select. The reason for that is that there were strict requirements for Preregistration Challenge prizes and Preregistration Badges. For example, to take part in the Preregistration Challenge, researchers were required to base their preregistrations on a detailed preregistration template. Similarly, Preregistration Badges were only handed out if several conditions were met, including that "the preregistered study design corresponds to the actual study design" and that "papers include a full disclosure of the results in accordance with the preregistration". We believe these quality checks may have filtered out preregistrations of lower quality or publications with more selective reporting. The consequence of this is that the problems we identified with selective hypothesis reporting in this study may be an underestimate of issues in the wider psychological literature.

While developing the protocol, we had to make many decisions to balance coding comprehensiveness and coding practicality. For example, to avoid spending a disproportionate time on single preregistration-study pairs we chose to assess selective hypothesis reporting for only the first sixteen hypotheses we identified in a preregistration even though a preregistration could include more. Another example is that we selected the operational hypothesis when both a conceptual and an operational hypothesis were present in a preregistration. We did so because we believed that the specific nature of operational hypotheses would make them easier to retrieve in the paper. Yet another example is that we tried to retrieve preregistered hypotheses only in the published paper itself, not in any supplementary materials, because we believe all preregistered hypotheses should be correctly presented in the main text. Even though some of these decisions may appear arbitrary, they could have substantively influenced our results and may make comparison with other studies on this topic difficult. Importantly, our coding protocols (https://osf.io/fdmx4, https://osf.io/uyrds), data (https://osf.io/8y2dv), and code (https://osf.io/xjzre) are openly available for everyone to scrutinize, and we strongly encourage readers to do this. Moreover, our protocols for identifying hypotheses as well as our dataset could be valuable resources for meta-researchers that have research questions about hypotheses in preregistrations and/or papers, or research questions about meta-research projects like ours.

Despite our extensive protocol, the coders in our project often indicated that they struggled with identifying hypotheses in preregistrations and subsequently retrieving these hypotheses from published papers. These difficulties may be due to authors consciously or subconsciously omitting or changing hypotheses from preregistration to paper. What could help to prevent this is a stricter adherence to existing reporting

guidelines like CONSORT for biomedicine studies (Schulz, Altman, & Moher, 2010) and JARS for psychology studies (Appelbaum et al., 2018). These guidelines typically emphasize that the results of all hypotheses should be reported and labeled as either primary or secondary, and either exploratory or confirmatory. An alternative explanation is that hypotheses were phrased so vaguely in preregistrations, papers, or both, that they could not effectively be identified or matched. This could have inflated the number of omitted hypotheses we found. Indeed, when two coders counted and coded the number of hypotheses in a preregistration, they agreed about the number of hypotheses in only 53.7% of the cases. Note that this is substantially higher than an earlier study by Bakker et al. (2020) who found agreement about the number of hypotheses in only 14.3% of cases. This difference may have come about because our more expansive protocol left less room for the coders' own interpretations.

Based on the results and the experience of the coders in this project, we believe that authors can improve the way they formulate their hypotheses. One simple recommendation would be to systematically put the hypotheses in a separate 'hypotheses' section in both the preregistration and the eventual study, and number all of them (possibly using letters to indicate hypotheses that are clustered together as we did in our hypothesis section). This will help readers to quickly delineate what the hypotheses are in a (proposed) study and quickly assess whether they are selectively reported in the paper. Maintaining consistency between preregistration and paper is also important regarding variable names. Like Claesen et al. (2021) we frequently encountered cases in which the names of one or more of the variables in a hypothesis differed between paper and preregistration, making our assessment of selective reporting challenging. Finally, it would help if all hypotheses were machine readable (Lakens and DeBruine, 2021). This would increase the reproducibility of research even more, and with that the ability to trail a hypothesis' progress from preregistration to publication.

A more structural solution to improve the way hypotheses are phrased would be to push more strongly for the registered reports format championed by, among others, Chris Chambers (Chambers & Tzavella, 2022; Nosek & Lakens, 2014). In the registered report format peer review takes place in two stages. In the first stage, the preregistration is peer-reviewed, which has the advantage that ambiguously phrased or overly complex hypotheses can be identified and corrected before the study is actually carried out. In the second stage, the resulting paper is peer-reviewed, where reviewers explicitly compare the preregistration and the paper. This explicit check might decrease the prevalence of selective hypothesis reporting in the final papers. Indeed, the first studies on the effectiveness of registered reports found that the proportion of positive results in registered report studies was substantially lower than in non-preregistered studies, indicating less selective reporting (Allen & Mehler, 2019; Scheel et al., 2021).

Because registered reports are not yet commonplace in research, an intermediate solution could be for editors to explicitly encourage reviewers to compare the preregistration and the paper. However, finding reviewers is already challenging as it is and requiring them to do additional tasks would not make this any easier. An increase in workload may be prevented if reviewers only need to verify if authors do what they promised in the preregistration, such as for registered reports (Chambers & Tzavella, 2022), next to checking the appropriateness of additional (so-called exploratory) analyses. At the very least, reviewers should be required to check whether a preregistration exists and can be easily accessed if the authors mention one. A pilot of a so-called discrepancy review, in which reviewers are explicitly assigned to check for discrepancies between preregistration and paper, found that this is practice is effective and could feasibly be introduced without many obstacles (TARG Meta-Research Group and Collaborators, 2022). What may also help is if reviews would have a more prominent place in the reward structure of academia, for example by making reviews public and assigning them DOIs. This would publicly show researchers' review, which could elevate reviews to be units of prestige besides regular peer-reviewed publications. Although there are some concerns (Rodríguez-Bravo, et al., 2017), this development could even be beneficial to early career researchers (Van den Akker, 2019).

In all, we need efforts on multiple fronts to arrive at a situation with clearer hypotheses and less selective hypothesis reporting. On an individual level, researchers, editors, and reviewers can bundle forces to make comparisons between preregistrations and papers more feasible. On a more structural level, journals can implement the registered reports format, and employers and funders can create more effective incentives for thorough reviews. This multi-faceted approach could lead to clearer and more consistent hypotheses, and with that more certainty about the validity of results in the scientific literature.

References

Allen, C., & Mehler, D. M. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biology*, *17*(5), e3000246.

Anderson, M. S. (2000). Normative orientations of university faculty and doctoral students. *Science and Engineering Ethics*, *6*(4), 443-461.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191

Bakker, M., Veldkamp, C. L., van Assen, M. A., Crompvoets, E. A., Ong, H. H., Nosek, B. A., ... & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*(12), e3000937.

Cairo, A. H., Green, J. D., Forsyth, D. R., Behler, A. M. C., & Raldiris, T. L. (2020). Gray (Literature) Matters: Evidence of Selective Hypothesis Reporting in Social Psychological Research. *Personality and Social Psychology Bulletin*, 0146167220903896.

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*(1), 29-42.

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.

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies. *Royal Society Open science*, *8*(10), 211037. https://doi.org/10.1098/rsos.211037

DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., ... & Schroeder, T. V. (2005). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *Archives of Dermatology*, *141*(1), 76-77.

DeVito, N. J., Bacon, S., & Goldacre, B. (2020). Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: A cohort study. *The Lancet*, *395*(10221), 361- 369.

Dwan, K., Altman, D. G., Clarke, M., Gamble, C., Higgins, J. P., Sterne, J. A., ... & Kirkham, J. J. (2014). Evidence for the selective reporting of analyses and discrepancies in clinical trials: A systematic review of cohort studies of clinical trials. *PLOS Medicine*, *11*(6).

Dwan, K., Gamble, C., Williamson, P. R., & Kirkham, J. J. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias: An updated review. *PLOS One*, *8*(7).

European Commission. (2012). Commission guideline: Guidance on posting and publication of result-related information on clinical trials in relation to the implementation of Article 57(2) of Regulation (EC) No 726/2004 and Article 41(2) of Regulation (EC) No 1901/2006. Retrieved from https://web.archive.org/web/20230305174330/https://op.europa.eu/en/publication-detail/-/publication/9a64920e-1134-11e2-8e28-01aa75ed71a1/language-en

Food and Drug Administration Amendments Act of 2007. (2018). Retrieved from https://www.govinfo.gov/content/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf

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.

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.

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, *17*(1), 239- 251.

Heirene, R., LaPlante, D., Louderback, E. R., Keen, B., Bakker, M., Serafimovska, A., & Gainsbury, S. M. (2021). Preregistration specificity & adherence: A review of preregistered gambling studies & cross-disciplinary comparison. *PsyArXiv Preprint*. https://doi.org/10.31234/osf.io/nj4es

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2.

Jeffreys, H. (1961). *Theory of ProbabilityTheory of probability*. Oxford: UK Oxford University Press.

Kambouris, S., Singleton Thorn, F., Van den Akker, O. R., De Jonge, M., Rüffer, F., Head, A., & Fidler, F. (2020). Database of Articles with Open Science Badges: 2020-02-21 Snapshot. https://doi.org/10.17605/osf.io/q46r5

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

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams R. B. Jr., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490.

Lakens, D., & DeBruine, L. M. (2021). Improving transparency, falsifiability, and rigor by making hypothesis tests machine-readable. *Advances in Methods and Practices in Psychological Science*, *4*(2), 2515245920970949.

Logg, J. M., & Dorison, C. A. (2021). Pre-registration: Weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes*, *167*, 18-27.

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.

Mazzola, J. J., & Deuling, J. K. (2013). Forgetting what we learned as graduate students: HARKing and selective outcome reporting in I–O journal articles. *Industrial and Organizational Psychology*, *6*(3), 279-284.

Merton, R. K. (1973). The Normative Structure of Science. In R. K. Merton (ed.), *The Sociology of Science: Theoretical and Empirical Investigations* (pp. 267-180). University of Chicago Press. (Original work published 1942)

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group*. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264-269.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600-2606.

Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer*, *31*(3).

O'Boyle, E. H. Jr., Banks, G. C., & Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, *43*(2), 376-399.

Ofosu, G. K., & Posner, D. N. (2021). Pre-analysis plans: An early stocktaking. *Perspectives on Politics*, 1-17. https://doi.org/10.1017/S1537592721000931.

Pham, M. T., & Oh, T. T. (2021). Preregistration is neither sufficient nor necessary for good science. *Journal of Consumer Psychology*, *31*(1), 163-176.

Rodríguez-Bravo, B., Nicholas, D., Herman, E., Boukacem-Zeghmouri, C., Watkinson, A., Xu, J., ... & Świgoń, M. (2017). Peer review: The experience and views of early career researchers. *Learned Publishing*, *30*(4), 269-277.

Scheel, A. M., Schrijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard Psychology literature with Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007467.

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Journal of Pharmacology and Pharmacotherapeutics*, *1*(2), 100-107.

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. *The American Statistician*, *49*(1), 108-112.

TARG Meta-Research Group and Collaborators. (2022). Discrepancy review: A feasibility study of a novel peer review intervention to reduce undisclosed discrepancies between registrations and publications. *Royal Society Open Science*, *9*(7), 220142.

Thibault, R. T., Clark, R., Pedder, H., Van den Akker, O. R., Westwood, S., Thompson, J., & Munafo, M. (2021). Estimating the prevalence of discrepancies between study registrations and publications: A systematic review and meta-analyses. *medRxiv Preprint*. https://doi.org/10.1101/2021.07.07.21259868

Van den Akker, O. R. (2019, October 10). Why I think open peer review benefits PhD students. *Behavioural and Social Sciences Community*. https://web.archive.org/web/20230305174558/https://socialsciences.nature.com/posts/54659-why-i-think-open-peer-review-benefits-phd-students?channel_id=2140-is-it-publish-or-perish

Van den Akker, O. R., Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P., ... & Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, 5.

Van der Steen, J. T., Van den Bogert, C. A., Van Soest-Poortvliet, M. C., Fazeli Farsani, S., Otten, R. H., Ter Riet, G., & Bouter, L. M. (2018). Determinants of selective reporting: A taxonomy based on content analysis of a random selection of the literature. *PLOS One*, *13*(2), e0188247.

Vinkers, C. H., Lamberink, H. J., Tijdink, J. K., Heus, P., Bouter, L., Glasziou, P., ... & Otte, W. M. (2020). Randomized clinical trial quality has improved over time but is still not good enough: An analysis of 176,620 randomized controlled trials published between 1966 and 2018. *medRxiv*. https://doi.org/10.1101/2020.04.22.20072371

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632-638.

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science*, *2*(3), 214-227.

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.

CHAPTER 3

3

The effectiveness of preregistration in psychology: Assessing preregistration producibility and preregistration-study consistency

Olmo R. van den Akker[1,2], Marjan Bakker[1], Marcel A. L. M. van Assen[1,3], Charlotte R. Pennington[4], Leone Verweij[1,5], Mahmoud M. Elsherif[6], Aline Claesen[7], Stefan D. M. Gaillard[8,9], Siu Kit Yeung[10], Jan-Luca Frankenberger[11], Kai Krautter[12], Jamie P. Cockcroft[13], Katharina S. Kreuer[8], Thomas Rhys Evans[14], Frédérique M. Heppel[15], Sarah F. Schoch[16], Max Korbmacher[17], Yuki Yamada[18], Nihan Albayrak-Aydemir[19,20], Shilaan Alzahawi[21], Alexandra Sarafoglou[15], Maksim M. Sitnikov[22], Filip Děchtěrenko[23], Sophia Wingen[24], Sandra Grinschgl[25], Helena Hartmann[26], Suzanne L. K. Stewart[27], Cátia M. F. de Oliveira[13], Sarah Ashcroft-Jones[28], Bradley J. Baker[29], and Jelte M. Wicherts[1]

[1]Department of Methodology and Statistics, Tilburg University, The Netherlands
[2]QUEST Center for Responsible Research, Berlin Institute of Health, Berlin, Germany
[3]Department of Sociology, Utrecht University, The Netherlands
[4]School of Psychology, Aston University, Birmingham, United Kingdom
[5]Department of Public Administration and Sociology, Erasmus University, Rotterdam, The Netherlands
[6]Department of Neuroscience, Psychology and Behaviour, University of Leicester, United Kingdom
[7]Faculty of Psychology and Educational Sciences, KU Leuven, Belgium
[8]Institute for Science in Society, Radboud University, Nijmegen, The Netherlands
[9]Center of Trial and Error, Utrecht, the Netherlands
[10]Department of Psychology, Chinese University of Hong Kong, Hong Kong SAR, China
[11]Behavioral Science, Radboud University, Nijmegen, The Netherlands
[12]Harvard Business School, United States
[13]Department of Psychology, University of York, United Kingdom
[14]School for Human Sciences and Institute for Lifecourse Development, University of Greenwich, United Kingdom
[15]Department of Psychological Methods, University of Amsterdam, The Netherlands
[16]Donders Institute, Radboud University Medical Center, Nijmegen, The Netherlands
[17]Western Norway University of Applied Sciences
[18]Faculty of Arts and Science, Kyushu University, Fukuoka, Japan
[19]Open Psychology Research Centre, Open University, United Kingdom
[20]Department of Psychological and Behavioural Science, London School of Economics and Political Science, United Kingdom
[21]Graduate School of Business, Stanford University, United States
[22]Department of Organization Studies, Tilburg University, The Netherlands
[23]Department of Cognitive Psychology, Institute of Psychology, Czech Academy of Sciences, Prague, Czechia
[24]Faculty of Management, Economics and Social Sciences, University of Cologne, Germany
[25]Institute for Psychology, University Graz, Austria
[26]Clinical Neurosciences, Department of Neurology and Center for Translational Neuro- and Behavioral Sciences (C-TNBS), University Hospital Essen, Germany
[27]School of Psychology, University of Chester, United Kingdom
[28]Department of Experimental Psychology, University of Oxford, United Kingdom
[29]Department of Sport and Recreation Management, Temple University, Philadelphia, United States

This paper is currently under review. A preprint can be found at https://doi.org/10.31222/osf.io/h8xjw.

Abstract

Study preregistration has become increasingly popular in psychology, but its effectiveness in restricting potentially biasing researcher degrees of freedom remains unclear. We used an extensive protocol to assess the producibility (i.e., the degree to which a study can be properly conducted based on the available information) of preregistrations and the consistency between preregistration and their corresponding papers for 300 psychology studies. We found that preregistrations often lack methodological details and that undisclosed deviations from preregistered plans are frequent. Combining the producibility and consistency results highlights that biases due to researcher degrees of freedom are likely in many preregistered studies. More comprehensive registration templates typically yielded more producible and hence better preregistrations. We did not find that effectiveness of preregistrations differed over time or between original and replication studies. Furthermore, we found that operationalizations of variables were generally more effectively preregistered than other study parts. Inconsistencies between preregistrations and published studies were mainly encountered for data collection procedures, statistical models, and exclusion criteria. Our results indicate that, to unlock the full potential of preregistration, researchers in psychology should aim to write more producible preregistrations, adhere to these preregistrations more faithfully, and more transparently report any deviations from their preregistrations. This could be facilitated by training and education to improve preregistration skills, as well as the development of more comprehensive templates.

Keywords: *preregistration, preregistration producibility, preregistration-study consistency, preregistration deviation, preregistration template, open science, meta-research*

Introduction

Hypothesis testing research involves making a lot of decisions. Such decisions include choosing a statistical model, the construction of outcome measures, and data handling strategies like dealing with missing data and outliers (Wicherts et al., 2016). These decisions are commonly known as 'researcher degrees of freedom' (Simmons, Nelson, & Simonsohn, 2011). The more decisions a researcher needs to make from the start of a project to its conclusion, the more degrees of freedom a study is said to have. In contrast to popular belief, researchers do not always make such decisions in a rational and objective manner (see Veldkamp, Hartgerink, Van Assen, & Wicherts, 2017). One reason for this is that researchers are susceptible to cognitive biases like confirmation bias and motivated reasoning bias (Bishop, 2020; Munafò, Chambers, Collins, Fortunato, & Macleod, 2020). In recent years, these biases have been highlighted as one of the main reasons for the replication crisis, the phenomenon that many studies fail to replicate in psychology and beyond (Malich & Munafò, 2022). One of the most common research biases involves a strong preference for research results that are easier to publish and hence beneficial to one's career because of similarly biased systematic incentives (Nosek, Spies, & Motyl, 2012). Because results involving *p*-values lower than .05 are deemed easier to publish, the label *p*-hacking has been used for the phenomenon of making research decisions to achieve a desired result (Parsons et al., 2022), although these decisions are typically neither explicitly intentional nor malicious (Smaldino & McElreath, 2016).

Following the replication crisis, several solutions have been proposed to combat questionable research practices such as *p*-hacking (see overview by Pennington, 2023). One particularly promising solution is preregistration (Nosek, Ebersole, DeHaven, & Mellor, 2018; Wagenmakers, Wetzels, Borsboom, Van der Maas, & Kievit, 2012), where researchers openly publish their hypotheses, study design, and analysis plan before collecting or analyzing the research data. Because researchers publish their decisions beforehand, preregistration can restrict researcher degrees of freedom and lower the possibility for *p*-hacking (Wicherts, et al., 2016), thereby diminishing the potential for biased outcomes to appear in the literature. The effectiveness of preregistration in achieving this goal depends on at least two aspects: (1) the *producibility* of the preregistration (i.e., whether the information provided in the preregistration is comprehensive enough to properly conduct the study)[5], and (2) the *consistency* between the preregistration and the published study (i.e., whether the study was carried out in line with the preregistered plan). When a preregistration only contains limited information, or when researchers do not largely adhere to the preregistered plan, preregistration is less effective (i.e., fewer researcher degrees of freedom are restricted and there is more room for *p*-hacking and other biased decision-making).

5 We called this aspect 'strictness' in our preregistration but changed this based on a reviewer's comment.

Empirical evidence on the effectiveness of preregistration in the social sciences is limited but the available studies from different fields show that preregistrations do not typically restrict most relevant researcher degrees of freedom (economics and political science: Ofosu & Posner, 2021; gambling studies: Heirene et al., 2021; multiple fields: Bakker et al., 2020). Specifically, Ofosu and Posner noted that independent variables, dependent variables, and statistical models were clearly outlined in most preregistrations, but that only a small proportion of preregistrations specified how missing data and outliers were to be handled. Heirene et al. and Bakker et al. found similar results: decisions relating to study design were relatively well-restricted compared to decisions regarding data collection and statistical analysis. This is problematic because the many decisions in analyzing data could still create sizeable variation in outcomes that researchers could selectively report (Olsson-Collentine, Van Aert, Bakker, & Wicherts, 2023).

In studies examining preregistration-study consistency, estimates of undisclosed deviations range from approximately two-thirds in a sample of gambling studies (Heirene et al., 2021) to about 90% in the journal *Psychological Science* (Claesen, Gomes, Tuerlinckx, & Vanpaemel, 2021). This is in line with earlier studies from biomedicine that also identified many inconsistencies between study registrations and papers (Li et al., 2018; Thibault et al., 2021). In the field of economics and political science, Ofosu and Posner (2021) focused on inconsistencies with regard to hypotheses and found that preregistered hypotheses could be retrieved in only two-thirds of the corresponding papers. Finally, in a sample of psychology studies, Van den Akker et al. (2023) found that about half of preregistered hypotheses could not be identified in the published paper and about one-fifth of preregistered hypotheses involved a change in the hypothesized direction of the effect. Consequently, although preregistrations could theoretically reduce questionable research practices, research suggests their implementation may not be as effective as initially hoped and thought.

It is important to note that deviations from a preregistration need not always be problematic (Nosek et al., 2019). Scientific research can be nonlinear and sometimes things change during the research process that could not have been foreseen. For example, the statistical assumptions of the preregistered model may not hold in practice, a subset of participants may need to be excluded because of a technical error, or the preregistration could simply have included a mistake. In situations like these, deviating from the preregistration may be the most reasonable way to still enable proper tests of the predetermined hypothesis. However, it is crucial to explain in the published work why any deviations were necessary, perhaps through Preregistration Planning and Deviation Documentation (Van 't Veer et al., 2019). Only then can readers assess the rationale behind the deviations and calibrate their confidence in the claims being made.

The current project is the first to simultaneously investigate both the producibility of preregistrations and the consistency between preregistrations and published studies in psychology. We do so in a sample of published preregistrations and papers (N = 300 when assessing producibility and N = 57 when assessing consistency). Aside from this overall assessment of preregistration effectiveness, we also assess how effectively the following specific study parts are preregistered: the operationalizations of the variables, the data collection procedure, the statistical model, the inference criteria, the exclusion criteria, the treatment of missing data, and the treatment of violations of statistical assumptions. For the study parts with the most inconsistencies between preregistration and paper, we also assess the different types of inconsistencies, the frequency with which they occur, and any explanations the authors may have for them. This may help identify areas where preregistration practices require the biggest improvements. Finally, we test several novel hypotheses that illustrate what factors may influence preregistration effectiveness, like replication status, time, and the comprehensiveness of the preregistration template.

We preregistered (see https://osf.io/83ahg) hypotheses about the overall effectiveness of psychology preregistrations, expecting that preregistration effectiveness would vary between different preregistration and study types. Our first hypothesis was that replication studies would be preregistered more effectively than original studies. Preregistration producibility may be better for replication preregistrations because available information about the primary (to-be-replicated) study nudges researchers to specify more study details in the preregistration of the replication study, making such preregistrations more producible. Additionally, preregistration-study consistency might be better for replication preregistrations because the principal goal of a replication study is to mimic the primary study. Given that the details of the primary study are specified in the published replication study, researchers doing replication studies can be expected to adhere more to the preregistration than researchers doing original studies.

Our second hypothesis was that more comprehensive preregistration templates (i.e., those targeting a greater number of research decisions) would yield more effective preregistrations than less comprehensive templates. The reasoning underlying this hypothesis is that comprehensive templates nudge researchers to specify more study details, making the preregistrations more producible than preregistrations based on less comprehensive templates. Moreover, researchers using more comprehensive templates may value restricting researcher degrees of freedom more than researchers using less comprehensive templates and are therefore more likely to adhere to the preregistration. These predictions are in line with the finding that registrations using formats with detailed instructions restricted the opportunistic use of researcher degrees of freedom better than formats with minimal direct guidance (Bakker et al., 2020). A number of

preregistration templates have been developed in recent years, some with a general purpose (e.g., Bowman et al., 2020; Preregistration Task Force, 2021), and some with a specific emphasis (e.g., for replication studies: Brandt et al., 2014; for secondary data analyses: Van den Akker et al., 2021; for systematic reviews: Van den Akker et al., 2022; for qualitative research, Haven & Van Grootel, 2019). In this study, we limited ourselves to general-purpose preregistration templates for hypothesis-testing research.

Our third hypothesis was that preregistration effectiveness has improved over time, something that was previously found by Heirene et al. (2021). We expected this to be likely as researchers are preregistering more and more (Pfeiffer & Call, 2022) and should therefore be getting more familiar and experienced with the practice of preregistration. Intuitively, this would make them more effective at (a) making their preregistrations more producible and (b) ensuring higher preregistration-study consistency.

Overview of preregistered hypotheses

1) Replication studies are more effectively preregistered than original studies
    a. Preregistrations of replication studies are more producible than preregistrations of original studies
    b. Replication studies are more consistent with their preregistration than original studies
2) Studies based on more comprehensive preregistration templates are more effectively preregistered than studies based on less comprehensive preregistration templates
    a. Preregistrations based on more comprehensive templates are more producible than preregistrations based on less comprehensive templates
    b. Studies based on more comprehensive preregistration templates are more consistent with their preregistration than studies based on less comprehensive preregistration templates
3) Preregistration effectiveness has improved over time
    a. Preregistration producibility has improved over time
    b. Preregistration-study consistency has improved over time

Method

Selection of preregistered studies

Our selection of preregistered studies was derived from a population of 459 preregistered psychology studies that had either won a Preregistration Challenge prize via the Center for Open Science initiative (see https://cos.io/our-services/prereg-more-information) or earned a Preregistration Badge before 2020 (see https://cos.io/our-services/

open-science-badges). This set of preregistrations has been previously used to assess whether hypotheses outlined in preregistrations matched those outlined in the corresponding papers (Van den Akker, et al., 2023). To search for hypotheses, Van den Akker et al. used the following keywords: "replicat", "hypothes", "investigat", "test", "predict", "examin", and "expect". Once they determined that the sentence with the keyword was indeed a hypothesis, they copy-pasted the text from the preregistration and separately extracted the variables (independent variables, dependent variables, mediating variables, and control variables). In the second stage of the project, coders were presented with the texts and the variables of all hypotheses and were asked to try to match the hypotheses to the hypotheses in the corresponding papers' introduction or methods sections. A hypothesis was labeled as a 'match' if the hypothesis in the paper involved the same variables and the same relationship between the variables as detailed in the preregistration. The authors ended up with a total of 1,143 matching hypotheses from 346 preregistration-study pairs (PSPs).

For the current project, we randomly selected one hypothesis per PSP. We did this because assessing more than one matching hypothesis in a given study would have led to dependencies in our data. Moreover, we wanted to assess preregistration effectiveness for study elements that are typically constrained to one particular hypothesis (e.g., the operationalization of the variables, and the statistical model). During the selection process, we excluded 46 studies that only involved hypotheses for which we could not clearly determine the type of hypothesis (i.e., an association, effect, moderation, or mediation) or hypotheses involving only one variable. We did so because our method for computing preregistration effectiveness required one clear hypothesis with at least two variables. Note that we did not explicitly preregister these exclusions. As a result, our final sample consisted of 300 hypotheses from 300 PSPs. Other than these exclusions, there were no unanticipated missing data.

An overview of our sample selection procedure can be found in the PRISMA flow diagram (Moher, Liberati, Tetzlaff, Altman, & PRISMA Group, 2009) in Figure 1. The protocols used by Van den Akker et al. (2023) to identify the hypotheses in preregistrations and their accompanying papers can be found at https://osf.io/fdmx4 and https://osf.io/uyrds, respectively.

Measuring preregistration effectiveness

We coded preregistration effectiveness using a protocol (adapted from Bakker et al., 2020) administered via Qualtrics that extracts information from the preregistration and the paper, and then helps assess preregistration producibility as well as preregistration-study consistency. The static version of this protocol can be found at https://osf.io/dpg3v. Filling out the protocol for one PSP typically took between 20 and 80 minutes,

although particularly challenging pairs could take multiple hours. Each PSP was coded by two independent coders, who subsequently resolved any coding inconsistencies

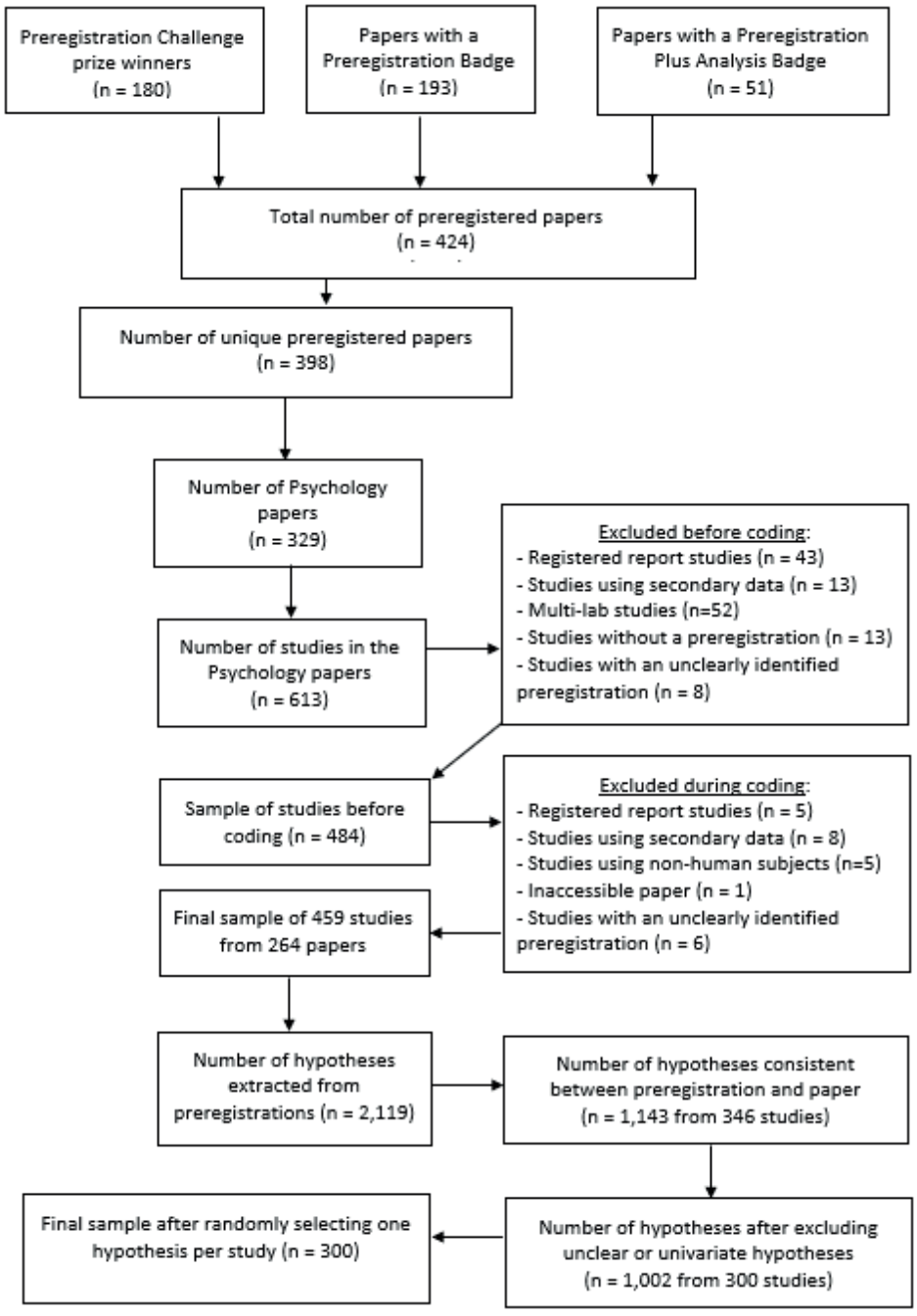


Figure 1. PRISMA flow diagram outlining the full sample selection procedure

among each other. The 28 coders in this project were researchers interested in assessing the field of psychology from a meta-scientific perspective. They were trained using a set of ten example PSPs, and coded on average of 20.9 PSPs (min. = 4, max. = 33).

Assessing five major study parts

We extracted information about the preregistration and the paper by answering questions about five major study parts (denoted by numbers below), some of which we divided into smaller study elements (denoted by letters below):

1. the operationalization of the independent variable (in case the hypothesis implied a directional link between two or more variables) or the first variable (in case the hypothesis did not imply a directional link between two or more variables):[6]
   a. the procedure of measurement;
   b. the potential values;
   c. how the variable was constructed from its components (e.g., a Likert scale based on item responses), if applicable
2. the operationalization of the dependent variable (in case the hypothesis implies a directional link between two or more variables) or the second variable (in case the hypothesis does not imply a directional link between two or more variables):
   a. the procedure of measurement;
   b. the potential values;
   c. how the variable was constructed from its components, if applicable;
3. the data collection procedure:
   a. sample size;
   b. sampling frame (i.e., the author's procedure for sampling participants);
4. the statistical model used:
   a. the model itself;
   b. the specification of the variables (e.g., whether a variable was added or changed);

c. the manner in which the variables were used in the model (e.g., the contrasts or whether they were standardized);

5. the statistical inference criteria used.

We selected these study parts because they represent the whole process of testing a hypothesis - study design (operationalization of the variables), data collection, and

6 Because it proved to be impossible to determine whether authors intended for hypotheses to be directional, we used manipulation status as a demarcation criterion: hypotheses involving at least one manipulated variable were presumed to be directional (i.e., have an independent and dependent variable) whereas all other hypotheses were not presumed to be directional. Manipulated variables were not further divided into study elements, but measured variables were. A result of this change is that we now list five major study parts in both Table 2 and Table 3.

statistical analysis (model and inference) - and are thus crucial to restrict researcher degrees of freedom for.

Measuring preregistration producibility

We scored the five study parts on *preregistration producibility* by assessing whether they were described in a *specific* (all steps that will be taken were described) and *precise* (each of the described steps allowed only one interpretation or implementation) manner (Bakker et al., 2020; Wicherts, et al., 2016) in the preregistration. When any part of a preregistration was described in a specific and precise manner, that part of the preregistration was scored with 2 points for producibility. When some but not all elements related to a part of the preregistration were described specifically and precisely, we awarded 1 point to that part. And, finally, when none of the elements was deemed specific and precise, we awarded 0 points.

An exception was the question about the data collection procedure, for which the protocol asked about two elements: sample size and sampling frame. If *either one of these two elements* was described specifically and precisely, the entire data collection procedure was scored with 2 points. We implemented this exception because researchers can choose to preregister either an exact sample size *or* a specific and precise sampling method, either of which would minimize researcher degrees of freedom. After taking the mean of all scores on the five major parts of the study, the preregistration could score between 0 (not producible at all) and 2 (optimally producible).

Measuring paper reproducibility

To be able to compare study parts between preregistration and paper properly, it is necessary that sufficient information about a study part is available in both the preregistration and the paper. For example, if the preregistration outlines in detail the statistical model that will be used, but the paper mentions the model only indirectly or not at all, it would be impossible to assess whether the model in the paper corresponds to the model in the preregistration. To assess whether sufficient information about a study part was provided in the paper, we also measured *paper reproducibility*. We measured this in exactly the same way as we measured preregistration producibility (see above). The term reproducibility was chosen for papers because studies presented in papers are already carried out and thus can only be *re*produced. Studies planned in preregistrations, on the other hand, need to be carried out (produced) later and are therefore labeled 'producible'. We deemed study parts to be sufficiently comparable if a study part scored either a 1 or 2 on preregistration producibility (specifying the level of detail in the preregistration) *and* paper reproducibility (specifying the level of detail in the paper). For the parts where this was not the case, we did not compute preregistration-study consistency.

Measuring preregistration-study consistency

To assess the *consistency* between a preregistration and the actual study, we scored whether the description of a study part in the preregistration and the corresponding paper were consistent. A preregistration and a study were considered 'consistent' when the researcher adhered to the action described in the preregistration within the published paper. In the preregistration-study consistency part of the protocol, any part could earn 1 point (consistent) or 0 points (inconsistent). This meant that the total consistency score could be between 0 (not consistent at all) and 5 (very consistent).

Combining producibility and consistency

To compute preregistration effectiveness for a given preregistration, we first multiplied the score for preregistration producibility with the score for preregistration-study consistency for each part separately. These multiplied scores signify how effectively each individual study part was preregistered. The highest possible score per part was 2, and could be achieved with a producibility score of 2 and a consistency score of 1. The lowest possible score was 0 and could be achieved if the producibility score and/or the consistency score were 0. We then took the mean of all of these partial effectiveness scores to get a total score that indicates how effectively a given study was preregistered as a whole (with scores varying from 0 to 2, where higher values indicate higher effectiveness). For example, let us suppose a PSP scored on preregistration producibility 1 point for the operationalization of the independent variable, 2 points for the operationalization of the dependent variable, 1 point for the data collection protocol, and 0 points for the statistical model and inference criteria; and on preregistration-study consistency 1 point for the operationalizations of the independent and dependent variable, and 0 points for the data collection protocol, the statistical model and the inference criteria. The preregistration effectiveness score of that study would then be $(1\times1 + 2\times1 + 1\times0 + 0\times0 + 0\times0)/10 = 0.3$. We took the mean of the partial effectiveness scores, instead of the sum like we preregistered, because we believe the resulting score range of [0, 2] is more interpretable than a sum score range of [0, 10].

Assessing minor study parts

Aside from the five 'major' parts of a study outlined above, we also scored four 'minor' study parts. Note that with the term 'minor' we do not mean that these study parts are less important to preregister well, but merely that these study parts may not apply to each study design. For example, if the analysis of a study does not involve a control variable, the first minor study part below is no longer applicable. Similarly, the second minor study part is not applicable if study participants were forced to respond to all items in a questionnaire, thereby circumventing missing data other than from attrition. The minor study parts are listed below using numbers, and the elements that constitute those parts are listed using letters.

1. the operationalization of the control variable:
   a. the procedure of measurement;
   b. the potential values;
   c. how the variable was constructed from its components, if applicable;
2. how missing data was handled:
   a. the definition of missing data;
   b. how missing data were dealt with;
3. how violations of statistical assumptions were handled:
   a. which assumptions were checked;
   b. how the assumptions were checked;
   c. how violations of assumptions were dealt with;
4. exclusion criteria.[7]

We scored the minor parts in the same way as the major parts, but the scores for these parts were not used to calculate a score for the preregistration/study overall. As such, they only provide information about preregistration producibility, preregistration-study consistency, and preregistration effectiveness of the individual study parts.

Assessing whether a hypothesis is part of a replication

Information about the replication status of hypotheses was taken directly from Van den Akker et al. (2023). They assessed whether a hypothesis was part of a replication or an original study by first searching the preregistration and paper for the string "replic" and assessing whether the authors referred to the hypothesis as being part of a replication attempt. If the authors did, in either the preregistration or the paper, Van den Akker et al. coded the hypothesis as a replication hypothesis. If the authors did not, Van den Akker et al. coded the hypothesis as an original hypothesis. The protocols used to assess whether a hypothesis was part of a replication can be found at https://osf.io/fdmx4 (for preregistrations) and https://osf.io/uyrds (for published papers).

Determining the comprehensiveness of preregistration templates

To identify the preregistration template used for a specific study we searched the paper presenting that study for the keyword "regist" to find the link to the preregistration. We then looked at the preregistration link and the surrounding paragraph to identify any

7 In our own preregistration, we divided the exclusion criteria into two elements: the definition of the criteria and the procedure of exclusion. When inspecting the data, however, we noticed that the types of inconsistencies listed for the definition were equivalent to the types of inconsistencies listed for the procedure: they all mentioned that one or more exclusion criteria were not mentioned, added, or changed in the paper compared to the preregistration. After some discussion among coders, we realized that authors typically did not describe the procedure of exclusion (e.g., whether the criteria were determined before or after data collection, or whether exclusion was listwise or pairwise) in a preregistration or paper. We suspect that most authors assumed that listwise exclusion was self-evident and other information was superfluous. Because of this, we decided to disregard the procedure of exclusion as a study element and only regard the definition of the criteria (to assess preregistration producibility).

references to a preregistration template. If there were no such references, we looked at the preregistration itself to identify which template had been used.

We scored the three preregistration templates with the highest frequency on their comprehensiveness (i.e., their potential to restrict researcher degrees of freedom) using a newly developed protocol. Using that protocol, we assessed whether the template included a prompt, additional instructions, and an example for the nine major and minor study parts (see https://osf.io/rtuvb for the filled-out protocol). The maximum possible comprehensiveness score using this protocol was 27 (very comprehensive), which each of the five major and four minor study parts receiving a maximum of 3 points. We gave 1 point if the study part was included in the template without additional instructions and an example, 2 points if it was included with either additional instructions or an example, and 3 points if it was included with both additional instructions and an example. When the study part was not included in the template, 0 points were given. Scoring was done by two independent coders (ORA and CRP) who together resolved three initial coding discrepancies. For one discrepancy, an independent third coder (MB) made the final call. Table 1 provides an overview of the preregistration templates we identified, their frequency and their comprehensiveness score. We observed large differences in comprehensiveness between the templates. While the OSF Prereg template scored almost the maximum number of points (24/27), the AsPredicted template and the Pre-Registration in Social Psychology template scored substantially less well, with 10 and 14 out of 27 points, respectively.

Table 1.
Frequencies and Comprehensive Scores of the Preregistration Templates used to draft the Preregistrations in our Sample.

Template	Freq.	Comprehensiveness
OSF Prereg template (Bowman et al., 2020)	122	24
AsPredicted (https://aspredicted.org)	112	10
Pre-Registration in Social Psychology (Van 't Veer & Giner-Sorolla, 2016)	21	14
OSF's Open Templates (https://osf.io/9j6d7; https://osf.io/haadc)	7	-
Happy Lab Pre-Registration Template (https://osf.io/yvsj8)	7	-
Replication Recipe (Brandt et al., 2014) (https://osf.io/4jd46)	1	-
Unknown	45	-
Total	315	-

Note: The OSF-Standard Pre-Data Collection Registration is combined with OSF's Open-Ended Registration into OSF's Open Templates because they share a minimalistic setup. This minimalistic setup also means they automatically score 0 on comprehensiveness.

Determining registration dates

To assess whether preregistration effectiveness increased over time we coded the date that the preregistration was formally registered. For frozen registrations (i.e., dated registrations that cannot be altered after the registration date) on the Open Science Framework, this information is clearly listed on the right-side of the preregistration document next to the word "registered". For frozen registrations on AsPredicted, this information is clearly listed on the top of the preregistration document next to the word "public". For non-frozen registrations we used the date at which the preregistration was last modified. The registration dates were recoded to the number of months since the date of the first preregistration in the sample, which was 14 April 2014 (Van Zant & Moore, 2015).

Determining the type of deviations and authors' explanations

We used an open-ended question to elicit the deviations between preregistrations and papers. For example, coders could state that the sample size was higher in the paper than in the preregistration, or could state which exclusion criteria differed between preregistration and paper. We also used an open question to elicit the authors' explanations for inconsistencies between the preregistration and the actual study in the published paper, if any. Both questions are listed in the static version of the protocol, which can be found at https://osf.io/dpg3v.

Results

Descriptive statistics

Of the 300 PSPs in our sample, we classified 138 (46%) as replication studies, and the remaining 162 (54%) as original studies. Registration time, as measured by the number of months since the registration date of the first preregistration in our sample, had a mean of 38.7 months (*SD* = 12.4), a median of 40, and a maximum of 67.

The data used in our analyses are publicly available on the Open Science Framework (https://osf.io/vwgak). The R-code we used is also publicly available, at https://osf.io/2yzsr (for analyses regarding producibility and effectiveness) and https://osf.io/g3fra (for analyses regarding consistency).

Table 2 presents the mean scores for preregistration producibility, consistency, and effectiveness for all the separate study parts as well as the total mean scores for the five major study parts. Table 2 also provides the frequency of the individual scores (0, 1, and 2 for producibility and effectiveness; 0, 1, and NA for consistency). The overall mean producibility score of the preregistrations in our sample was 1.33 out of 2 (N = 300, *SD* = 0.41, min. = 0, max. = 2), and the overall mean consistency score was 0.71 out of 1 (N = 57, *SD* = 0.20, min. = 0, max. = 1). The mean effectiveness score per PSP was 0.79 out of 2 (N = 300, *SD* = 0.43, min. = 0, max. = 2).[8] The correlation between the producibility scores and consistency scores was $r = -.11$, $t(55) = -0.82$, $p = .418$).

8 We also assessed the preregistration effectiveness for the current study and arrived at a score of 2.0 for preregistration producibility, a score of 0.8 for preregistration-study consistency (because our sample size was not consistent), and therefore a score of 0.8 for preregistration effectiveness. For the non-essential elements, we scored 2 points for the producibility of the exclusion criteria and the handling of missing data but 0 points for the handling of violations of statistical assumptions. Finally, the exclusion criteria were not consistent because we added two criteria, whereas the missing data were inconsistent because we did not mention them in the paper at all. After making this assessment, we included a sentence about handling missing data to the paper. This shows that our assessment protocol is not only useful to assess producibility and consistency *post hoc* but also when writing up your preregistration or paper. Our (obviously biased) assessment can be found at https://osf.io/byacg.

Table 2.

Overview of Producibility, Consistency, and Effectiveness Scores (with Standard Deviations) for the five Major and five Minor Study Parts, as well as Total Scores for the Major Study Parts.

	Producibility				Consistency				Effectiveness			
Major study parts	**0**	**1**	**2**	**Score (SD)**	**0**	**1**	**NA**	**Score (SD)**	**0**	**1**	**2**	**Score (SD)**
Measured variable (*N*=164) *	28 (17%)	61 (37%)	75 (46%)	**1.29 (0.74)**	45 (27%)	91 (55%)	28 (17%)	**0.67 (0.47)**	73 (45%)	43 (26%)	48 (29%)	**0.85 (0.85)**
Manipulated variable (*N*=218) *	49 (22%)	NA	169 (78%)	**1.55 (0.84)**	8 (4%)	146 (67%)	64 (29%)	**0.95 (0.22)**	63 (33%)	NA	155 (67%)	**1.42 (0.91)**
Dependent variable (*N*=218) *	24 (11%)	79 (36%)	115 (53%)	**1.42 (0.68)**	56 (26%)	131 (60%)	31 (14%)	**0.70 (0.46)**	87 (40%)	59 (27%)	72 (33%)	**0.93 (0.85)**
Data collection procedure (*N*=300)	38 (13%)	NA	262 (87%)	**1.75 (0.67)**	176 (59%)	84 (28%)	40 (13%)	**0.32 (0.47)**	216 (72%)	NA	84 (28%)	**0.56 (0.90)**
Statistical model (*N*=300)	23 (8%)	159 (53%)	118 (39%)	**1.32 (0.61)**	121 (40%)	135 (45%)	44 (15%)	**0.53 (0.50)**	165 (55%)	89 (30%)	46 (15%)	**0.60 (0.74)**
Inference criteria (*N*=300)	191 (64%)	NA	109 (36%)	**0.73 (0.96)**	10 (3%)	94 (31%)	196 (65%)	**0.90 (0.30)**	206 (69%)	NA	94 (31%)	**0.63 (0.93)**
Total scores				**6.65 (2.04)**				**3.53 (1.00)**				**3.96 (2.12)**
	Producibility				**Consistency**				**Effectiveness**			
Minor study parts	**0**	**1**	**2**	**Score (SD)**	**0**	**1**	**NA**	**Score (SD)**	**0**	**1**	**2**	**Score (SD)**
Measured control variable (*N*=20) **	7 (35%)	5 (25%)	8 (40%)	**1.05 (0.89)**	9 (45%)	4 (20%)	7 (35%)	**0.31 (0.48)**	16 (80%)	1 (5%)	3 (15%)	**0.35 (0.75)**
Manipulated control variable (*N*=23) **	5 (22%)	NA	18 (78%)	**1.57 (0.84)**	0 (0%)	18 (78%)	5 (22%)	**1.00 (0.00)**	5 (22%)	NA	18 (78%)	**1.57 (0.84)**
Exclusion criteria (*N*=300)	68 (23%)	NA	232 (77%)	**1.55 (0.84)**	86 (29%)	106 (35%)	108 (36%)	**0.55 (0.50)**	194 (65%)	NA	106 (35%)	**0.71 (0.96)**
Missing data (*N*=300)	159 (53%)	21 (7%)	120 (40%)	**0.87 (0.96)**	9 (3%)	37 (12%)	254 (85%)	**0.80 (0.40)**	263 (88%)	2 (1%)	35 (12%)	**0.24 (0.65)**
Statistical assumptions (*N*=300)	279 (93%)	17 (6%)	4 (1%)	**0.08 (0.32)**	2 (1%)	6 (2%)	292 (97%)	**0.75 (0.46)**	294 (98%)	5 (2%)	1 (0%)	**0.02 (0.17)**

Note Table 2. The single asterisk in Table 2 highlights that we had 82 hypotheses without a directional relationship and therefore two measured variables, and 218 hypotheses with a directional relationship and therefore a manipulated variable and a dependent variable. The double asterisk in Table 2 highlights that we only had 46 hypotheses with one or more control variables. Twenty of those were part of a non-directional hypothesis, and 26 were part of a directional hypothesis.

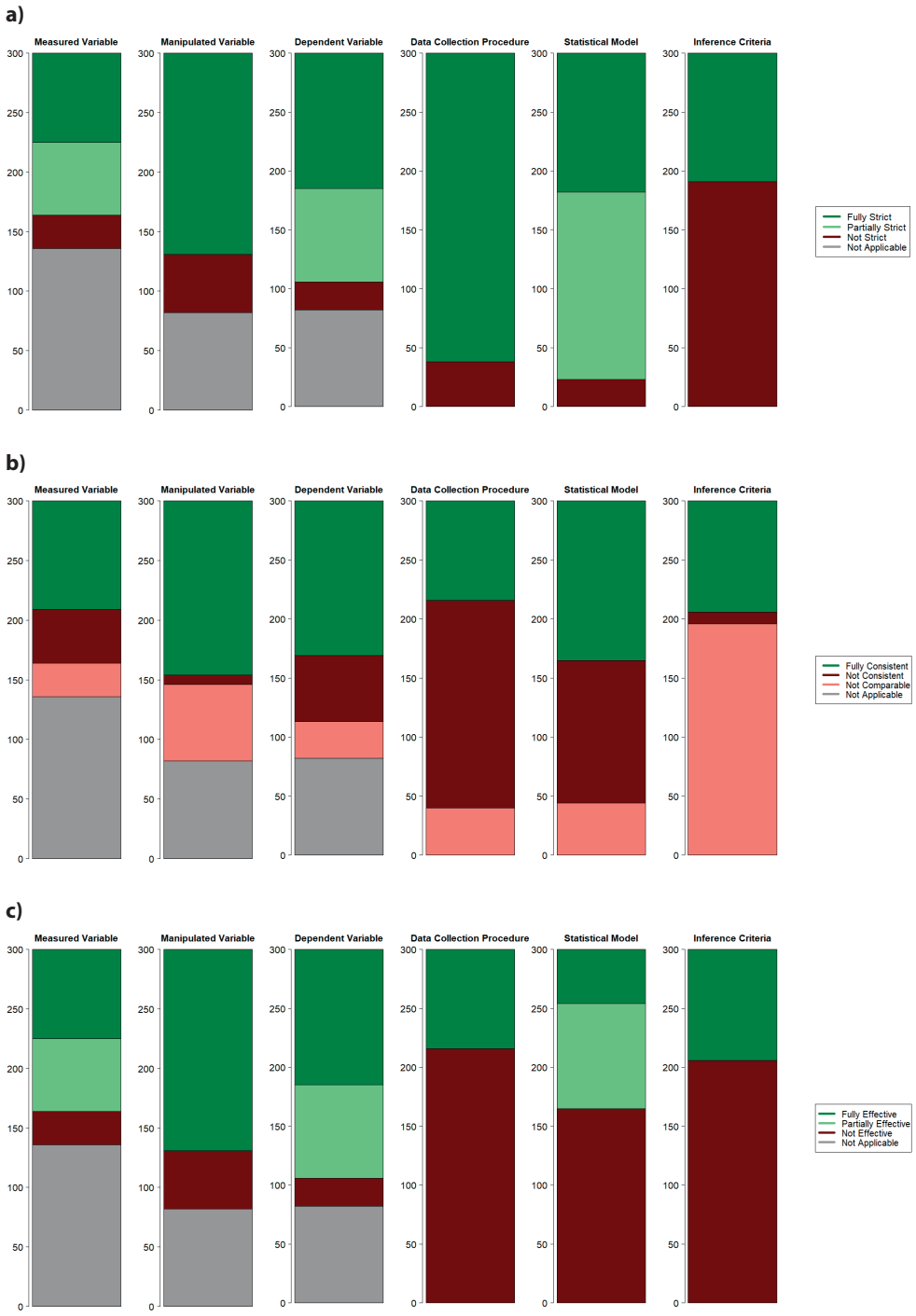


Figure 2. *Preregistration producibility scores (a), preregistration-study consistency scores (b), and effectiveness scores (c)*

Note that the consistency scores in Table 2 indicate the proportion of PSPs for which that study part was consistent out of all PSPs for which the study parts could be compared between preregistration and paper. For example, the statistical model could be compared 256 times, of which 121 (47%) were consistent. Because the inference criteria were almost never explicitly stated in the paper, we used implicit consistency instead. That is, we checked whether the authors' conclusion about the statistical result was in line with their preregistered inference criterion. For example, if the preregistration specified $\alpha = .01$ and the paper drew a conclusion in the form of "we found an effect of X on Y, $p = .007$" we would consider this as consistent and score the consistency of inference criteria with 1 point. However, if the paper specified $\alpha = .01$ and stated "we found an effect of X on Y, $p = .023$" we would consider this as inconsistent (and allocate 0 points) as a different criterion seems to be used. Note that this was a deviation from our preregistration, but that this deviation did not influence our measurement of preregistration producibility and paper reproducibility.

To allow the calculation of preregistration effectiveness for each individual PSP, all 'NA' responses for consistency were recoded to scores of 0. As can be seen in Table 2, we found mean efficiency scores below 1 (out of 2) for all study parts except for manipulated variables (1.42). Generally, the operationalizations of the variables (measured, manipulated, and dependent) were more effectively preregistered than the other study parts. A visualization of the scores for producibility, consistency, and effectiveness can be found in Figures 2-4.

We also collected data about the study elements that constitute the different study parts. In Table 3, we see that within each study part, the elements were often more or less equally producible (see the column 'Prereg producibility'). Consistency between preregistration and paper with regard to study elements (computed only for elements that were at least partially producible in the preregistration *and* reproducible in the paper) is outlined in the column 'Consistency' in Table 3. In the final column of Table 3 ('Explanations'), we provide information about the presence of authors' explanations for preregistration deviations in the final paper. Explanations of deviations were rarely provided, especially for study elements where inconsistencies were rare. We also assessed what kind of inconsistencies were most common by exploring the three study parts with the most inconsistencies: the data collection procedure, the exclusion criteria, and the statistical model. Our categorization of inconsistencies for these three study parts can be found at https://osf.io/crd3u.

Table 3

Overview of the Preregistration Producibility, Paper Reproducibility, Consistency and Authors' Explanations for Preregistration Deviations for Each Study Element.

	Prereg Producibility	Paper Reproducibility	Consistency	Explanations
Measured variable (*N*=164)				
Procedure of measurement	102 (62%)	125 (76%)	89 / 92 (97%)	0 / 3 (0%)
Potential values	87 (53%)	108 (66%)	69 / 73 (95%)	1 / 4 (25%)
Procedure to construct composite (*N*=73)	37 (51%)	45 (62%)	20 / 23 (87%)	0 / 3 (0%)
Manipulated variable (*N*=218)	169 (78%)	202 (93%)	146 / 154 (95%)	0 / 8 (0%)
Dependent variable (*N*=218)				
Procedure of measurement	184 (84%)	199 (91%)	150 / 163 (92%)	1 / 13 (8%)
Potential values	150 (69%)	177 (81%)	115 / 120 (96%)	0 / 5 (0%)
Procedure to construct composite (*N*=134)	83 (62%)	76 (57%)	54 / 57 (95%)	0 / 3 (0%)
Measured control variable (*N*=20)				
Procedure of measurement	13 (65%)	12 (60%)	8 / 8 (100%)	0 / 0
Potential values	12 (60%)	8 (40%)	5 / 7 (71%)	0 / 2 (0%)
Procedure to construct composite (*N*=8)	3 (38%)	2 (25%)	1 / 1 (100%)	0 / 0
Manipulated control variable (*N*=23)	18 (78%)	22 (96%)	18 / 23 (78%)	0 / 5 (0%)
Data collection procedure (*N*=300)				
Exact sample size	178 (59%)	176 / 178 (99%)	49 / 176 (28%)	26 / 120 (22%)
Sampling frame	84 (28%)	68 / 84 (81%)	35 / 52 (67%)	6/17 (35%)
Exclusion criteria (*N*=300)	232 (77%)	225 (75%)	106 / 192 (55%)	13 / 86 (15%)
Missing data (*N*=300)				
Definition of criteria	123 (41%)	55 (18%)	35 /42 (83%)	0 / 7 (0%)
Method of handling	138 (46%)	53 (18%)	39 / 42 (93%)	0 / 3 (0%)
Statistical model (*N*=300)				
Which model was used	256 (85%)	244 (81%)	162 / 216 (75%)	11 / 54 (20%)
Specification of variables	254 (85%)	263 (88%)	207 / 226 (92%)	5 / 22 (23%)
How the variables are used in the model	128 (43%)	110 (37%)	66 / 75 (88%)	5 / 9 (56%)
Statistical assumptions (*N*=300)				
Which assumptions are checked	20 (7%)	19 (6%)	8 / 8 (100%)	0 / 0
How assumptions are checked	4 (1%)	8 (3%)	1 / 1 (100%)	0 / 0
What is done in case of violations	19 (6%)	18 (6%)	6 / 6 (100%)	0 / 0
Inference criteria (*N*=300)	109 (36%)	37 (12%)	94 / 104 (90%)	1 / 10 (10%)

Hypothesis tests

To test whether replication studies were more effectively preregistered than original studies (Hypothesis 1) we ran three multilevel regressions (with study as the first level, and paper as the second level): one with preregistration producibility (M1a), one with preregistration-study consistency (M1b), and one with preregistration effectiveness as the dependent variable (M1c). The main independent variable *replic* was a dummy (replication vs. original study). In contrast to our hypothesis, we found no evidence that replication studies were preregistered more producibly (M = 1.31) than original studies (M = 1.35), B_1 = -0.001, $t(234,7)$ = -0.01, 99% CI = [-0.10, 0.10], p = .988, nor that preregistration-study consistency was higher for replication studies (M = 0.72) compared to original studies (M = 0.70), B_1 = -0.001, $t(49.2)$ = -0.016, 99% CI = [-0.14, 0.14], p = .988. Consequently, the effectiveness of preregistration was not higher for replication studies (M = 0.79) than for original studies (M = 0.80), B_1 = 0.03, $t(294.1)$ = 0.55, 99% CI = [-0.10, 0.15], p = .582. The regressions related to all hypotheses are presented in Table 4. We computed unstandardized coefficients for all preregistered hypothesis tests and used an alpha level of .01, as preregistered.

To compare preregistration templates in line with Hypothesis 2, we ran the same three multilevel regressions as for Hypothesis 1 twice: once with the addition of a dummy variable with value 1 if the preregistration was produced using the Open Science Framework template, and value 0 if the preregistration was produced using the AsPredicted template (M2a1, M2b1, and M2c1), and once with the addition of a dummy variable with value 1 if the preregistration was produced using the Open Science Framework template, and value 0 if the preregistration was produced using the Social Psychology template (M2a2, M2b2, and M2c2). In line with our hypothesis, we found that preregistrations based on the OSF template were more producible (M = 1.62) than preregistrations based on the AsPredicted template (M = 1.15), B_1 = 0.44, $t(170.4)$ = 8.30, 99% CI = [0.30, 0.59], p < .001, and the Social Psychology template (M = 1.31), B_1 = 0.30, $t(97.2)$ = 3.32, 99% CI = [0.07, 0.54], p = .001. Similarly, OSF preregistrations (M = 0.98) were more effective than both AsPredicted preregistrations (M = 0.69), B_1 = 0.30, $t(144.3)$ = 4.62, 99% CI = [0.16, 0.44], p < .001, and the Social Psychology preregistrations (M = 3.21), B_1 = 0.34, $t(90.6)$ = 2.79, 99% CI = [0.03, 0.66], p = .006. The higher effectiveness in OSF templates related to AsPredicted templates was likely due to differences in producibility, as there was no significant difference in preregistration-study consistency between OSF templates (M = 0.71) and AsPredicted templates (M = 0.67), B_1 = 0.04, $t(48.2)$ = 0.35, 99% CI = [-0.28, 0.36], p = .730. We could not test for a difference in preregistration-study consistency between OSF templates and Social Psychology templates because preregistration-study consistency could not be assessed for any of the Social Psychology templates as insufficient information was present to compare preregistrations and

papers. We computed unstandardized coefficients for all preregistered hypothesis tests and used an alpha level of .01, as preregistered.

Finally, to test whether preregistration effectiveness improved over time (Hypothesis 3), we again ran the same three multilevel regressions as for Hypothesis 1, with the addition of a continuous variable denoting the number of months between the registration date of the preregistration and the registration date of the first preregistration in our sample (see M3a, M3b, M3c in Table 4). As no effect of time was observed in any of the three analyses, we conclude that there was not sufficient evidence that the quality of preregistration improved over time (producibility: B_1 = 0.001, t(296.8) = 0.38, 99% CI = [-0.004, 0.006], p = .704; preregistration-study consistency: B_1 = 0.02, t(50.2) = 1.46, 99% CI = [-0.003, 0.01], p = .151; and effectiveness: B_1 = -0.001, t(266.0) = -0.61, 99% CI = [-0.01, 0.004], p = .545). We computed unstandardized coefficients for all preregistered hypothesis tests and used an alpha level of .025, as preregistered.

Exploratory analyses

The results outlined above indicate whether our sample of studies were preregistered sufficiently producible, consistent, and consequently, effective. While these results indicate the potential for *p*-hacking in a certain study, they do not speak to whether *p*-hacking actually took place. Because the research process largely takes place behind the closed doors of offices, direct evidence for *p*-hacking is almost impossible to attain. However, we can use the proxy of statistical significance to explore whether more producible, more consistent, and more effective preregistrations are associated with a lower rate of statistical significance, which would suggest less *p*-hacking in these studies. To test this, we linked each study's producibility scores, consistency scores, and effectiveness scores to whether the assessed hypothesis (see the section 'Selection of preregistered studies' for a description of how we selected hypotheses) yielded a statistically significant result. We used multilevel analyses with study as Level 1 and paper as Level 2. Data about statistical significance was derived from Van den Akker et al. (2023). The analysis for consistency did not converge because of the low number of data points (24 consistency scores with a statistically significant result, and 24 consistency scores with a non-significant result). Furthermore, we found no evidence of an association of preregistration producibility and effectiveness with statistical significance (producibility: B_1 = -0.14, t(200.4) = -1.71, 99% CI = [-0.29, 0.02], p = .088; effectiveness: B_1 = -0.12, t(227.5) = -1.72, 99% CI = [-0.26, 0.02], p = .088).

Table 4.

Results of our tests of Hypothesis 1, Hypothesis 2, and Hypothesis 3 (M3a, M3b, and M3c).

Parameters	M1a	M1b	M1c	M2a1	M2b1	M2c1	M2a2	M2b2	M2c2	M3a	M3b	M3c
Fixed effects (standard errors)												
Intercept	**1.34*** **(0.03)**	**0.71*** **(0.04)**	**0.80*** **(0.03)**	**1.16*** **(0.05)**	**0.66*** **(0.12)**	**0.67*** **(0.06)**	**1.30*** **(0.08)**	**-**	**0.65*** **(0.11)**	**1.32*** **(0.08)**	**0.50*** **(0.15)**	**0.85*** **(0.09)**
Level 1												
Replication	-0.001 (0.04)	-0.001 (0.05)	0.03 (0.05)	0.01 (0.04)	0.01 (0.06)	0.05 (0.05)	0.01 (0.04)	-	0.01 (0.07)	-0.002 (0.04)	-0.02 (0.05)	0.03 (0.05)
OSF vs. AP	-	-	-	**0.44*** **(0.05)**	0.04 (0.12)	**0.30*** **(0.07)**	-	-	-	-	-	-
OSF vs. SP	-	-	-	-	-	-	**0.30*** **(0.09)**	**-**	**0.34*** **(0.12)**	-	-	-
Months	-	-	-	-	-	-	-	-	-	0.001 (0.002)	0.004 (0.003)	-0.001 (0.002)
Random effects												
Paper-level	0.14	0.01	0.10	0.08	0.02	0.09	0.08	-	0.09	0.14	0.01	0.10

Note. Model 1a refers to the model testing the first part of Hypothesis 1 (producibility), while Model 1b and 1c test the second (consistency) and third (effectiveness) part, respectively. The same holds for the models M2 and M3, which test Hypothesis 2 and Hypothesis 3. * indicates $p < .01$.

Discussion

The number of preregistrations has greatly increased in recent years (Pennington, 2023). However, empirical evidence has been lacking as to whether preregistration achieves its goal of restricting researcher degrees of freedom. In this study, we assessed 300 preregistered psychology studies on how producible the preregistrations were and how consistent the preregistrations were with their corresponding papers. We found a mean producibility score of 1.33 out of 2 and a mean consistency score of 0.71 out of 1. Combining producibility and consistency, we found a mean score for preregistration effectiveness of 0.79 out of 2. These scores indicate that over the years 2014-2020, the practice of preregistration was not as effective as it could have been, either because preregistrations were not producible enough and/or because researchers generally deviated substantially from the preregistration. As such, the possibility for the opportunistic use of researcher degrees of freedom remained after preregistration. This finding is in line with earlier studies that assessed preregistration in economics and political science (Ofosu & Posner, 2021), in gambling (Heirene et al., 2021), and in a cross-disciplinary sample (Bakker et al., 2020).

When focusing on different study parts, we found that the operationalizations of the variables were preregistered more producibly than other study parts and that the data collection procedure, the statistical model, and the exclusion criteria were the least consistent between preregistration and paper. Moreover, we rarely encountered any concrete explanations by the authors for inconsistencies between preregistrations and papers. These results replicate previous findings that study parts that are more effectively preregistered tend to be tied to the operationalization of variables (however, see Sarafoglou, Hoogeveen, & Wagenmakers, 2023). This may be the case because the variables are the foundation of a scientific study, and researchers are more invested in properly preregistering them. More cynically, it could be argued that it is easier to *p*-hack during the statistical analysis than in the operationalization of the variables, simply because there are more researcher degrees of freedom related to the statistical analysis (Wicherts, et al., 2016). Future meta-scientific research could investigate the research process in detail to comprehensively identify the different ways a researcher could steer a study in a certain direction, and which of these ways generally biases the results most (see Stefan & Schönbrodt, 2023). Such research would shed light on which study parts to give priority when preregistering a study.

We also carried out three novel hypothesis tests. Hypothesis 1, stating an association between replication status and preregistration producibility and consistency, was not supported. Our rationale for expecting more producible preregistrations for replication studies than for original studies was that information about the to-be-replicated study

should be readily available in the paper, meaning that authors could simply include that information in their preregistration. However, this study found that study designs were often not comprehensively reported in papers with prereregistered studies, and the same issue likely holds for papers with non-preregistered studies. The vast number of reporting guidelines designed to help researchers report study details more comprehensively (see the EQUATOR Network, Simera et al., 2010) confirms this.

Additionally, we argued that preregistration-study consistency might be better for replication preregistrations because the principal goal of a replication study is to mimic the primary study. Authors of replication studies should therefore be more motivated to adhere to their preregistration than authors of original studies. However, there could be many other factors at play that influence preregistration-study consistency. It could be, for example, that the hypotheses or methodological designs of preregistered studies are simpler, which could have counteracted any motivation effect in researchers as it should be easier to adhere to a simple preregistered plan than a difficult one. Alternatively, it could be that there is a difference in motivation between researchers who conduct a replication study and researchers who conduct original studies, but that this does not hold for researchers who preregister because their motivation to adhere to the preregistration is high regardless of study type. Finally, it could simply be that our initial intuition about (researchers conducting) replication studies was wrong. In any case, we did not find sufficient evidence to establish that replication studies involve more effective preregistrations than original studies.

In line with Hypothesis 2, preregistrations based on more comprehensive templates were generally more producible and more effective than preregistrations based on less comprehensive templates. However, consistency was not significantly higher. That more comprehensive templates did not yield more consistency between preregistrations and papers may be due to a faulty assumption. We assumed that people using more comprehensive templates would be more motivated to effectively preregister their study than people using less comprehensive templates, as comprehensive templates require more work. However, it may well be that the choice of preregistration template is determined by other factors like the specific field one is in, one's knowledge of the digital Open Science space, or simply random events.

In contrast with Hypothesis 3, we did not find evidence that preregistrations became more effective over time. One reason for this could be that the early adopters of preregistration (i.e., those who authored the earliest preregistrations in our sample) were already more effective at preregistration to begin with. This would make intuitive sense because their early uptake indicates an intrinsic interest in preregistration. Our data do not allow a test of this explanation because we do not know who the early adopters

are in our dataset. It could for example be that a researcher conducted preregistrations early on outside of the scope of the Preregistration Challenge or Preregistration Badge infrastructure. Building on our results with a survey about the adoption of preregistration practices could be informative to assess the plausibility of this explanation. Alternatively, it could be that our operationalization of time did not allow a valid test of the hypothesis. Ideally, one would assess the association between time and the effectiveness of preregistration *within* authors, but the short time period yielded almost no repeated first authors, thus ruling out this approach. Future studies that use a wider time period may be able to test this hypothesis more effectively. Finally, it may well be that preregistration skills have not improved over time because learning is difficult if one is not aware of one's mistakes. While preregistration templates can function as a building block of good preregistrations, these templates often do not specify common preregistration mistakes nor detailed examples of good preregistrations. The current study established common preregistration mistakes and identified a host of high-quality preregistrations. Hopefully, these will be used by researchers to improve their preregistration skills.

In general, we did not foresee that there would be so many situations (about 15% of cases) where we could not assess the consistency between preregistration and paper for a certain study part. This occurred when either the preregistration, the paper, or both did not provide sufficient information to allow a comparison. A consequence of this lack of proper reporting is that our statistical tests about preregistration-study consistency had less statistical power than anticipated, particularly for finding small true effect sizes. We urge researchers to explicitly mention in research papers all the study parts discussed in the preregistration, even if the information seems trivial or irrelevant. If there is insufficient information to compare preregistration and paper, it is unclear whether researcher degrees of freedom were left open and readers are forced to conclude that *p*-hacking would have been possible.

Our exploratory analyses assessing the relationship between preregistration effectiveness and statistical significance did not provide sufficient evidence for the claim that more effective preregistrations better prevent *p*-hacking than less effective preregistrations. One possible explanation for the absence of an association is that we investigated not only primary hypotheses but hypotheses that were indicated by one of seven keywords (see the section 'Selection of preregistered studies'). It is plausible that primary hypotheses have a higher likelihood of being statistically significant because they were expected a priori to be supported (and that this was the reason to do the study in the first place), or because the hypothesis was selected a posteriori to become the primary hypothesis *because* it was statistically significant. In addition, the statistical power for our exploratory analyses was likely low because of the small number of studies we could

assess (N=233). Our study suggests that if an association exists, it is likely small (95% confidence interval = [-0.26, 0.02]), which raises the question of whether the added time and effort associated with an effective preregistration (Sarafoglou et al., 2023) over a less effective preregistration is worth it.

Importantly, there could also be other reasons for why more effective preregistrations would be associated with a higher likelihood of statistically significant effects. For example, researchers who diligently and conscientiously write up a producible preregistration might also conduct a priori power analyses more diligently and conscientiously, leading to higher sample sizes and higher statistical power. In that case, more effective preregistrations would also be related to statistical significance, but the contributing factor would be the researchers, not preregistration itself. Because of the implications of finding an association between preregistration and statistical significance of hypothesis tests and because of the possibility of confounding factors, we recommend conducting a confirmatory test of this hypothesis in a high-powered future study. In addition, similar confirmatory tests could be initiated to assess the validity of other benefits of preregistration (see Lakens, 2019; Sarafoglou et al., 2023; Wagenmakers & Dulith, 2016) that so far have remained largely theoretical.

Overall, our results suggest there is room for improvement in the practice of preregistration, but there are several limitations of our study that we need to consider. For example, the preregistration effectiveness scores for the data collection procedure and the statistical model may be low because our coding was quite strict. In the case of the data collection procedure, one could argue that our coding was *too* strict, specifically in the cases where an exact sample size was preregistered. As can be seen in Figure 2, many sample sizes only differed slightly between preregistration and study, sometimes by only one or two participants. As preregistered, we labeled each deviation, however small, as an inconsistency, yielding a consistency score and an effectiveness score of zero. However, slight deviations in sample size would yield only a limited potential for *p*-hacking as the addition or subtraction of one or two participants would probably not change a statistically nonsignificant ($p > .05$) to a statistically significant result ($p < .05$). Yet, such *p*-hacking is still possible. Indeed, optional stopping has been argued as one particularly potent way of getting a statistically significant result (Hartgerink, Van Aert, Nuijten, Wicherts, & Van Assen, 2016), especially in combination with other opportunistic uses of researcher degrees of freedom (Wicherts, 2017). As such, we maintain that any deviation from an explicitly stated sample size should be labeled as an inconsistency. We encourage readers to analyze our data (accessible at https://osf.io/vwgak) using their own definition of a sample size deviation to draw their own conclusions.

In the case of the statistical model, one issue is that the low scores on producibility could have arisen because we included the study element 'the way the variables were used in the model'. This element reflected factors such as mean-centering predictors or the use of robust standard errors. However, one might argue that proper preregistrations do not always require such detailed information. For example, some model specifications are so standard that mentioning them in a preregistration or paper would be seen as superfluous (e.g., the use of ordinary least squares estimation instead of weighted least squares estimation). The point here is that authors do not always need to specify detailed information about a statistical model other than the essential information captured by the other elements of the statistical model: the model itself, and the specification of the variables. However, if the authors did not specify any additional information, we did score the element 'the way the variables were used in the model' with zero points for producibility, and thus effectiveness. To correct for this, we did an exploratory analysis where we recalculated the producibility score, consistency score, and effectiveness score for the statistical model, which became 1.70 (was 1.32), 0.65 (was 0.53), and 1.00 (was 0.60), respectively. These updated scores better align with Ofosu and Posner (2021), who found that not only the variables, but also the statistical model was generally well-preregistered.

Furthermore, it could also be argued that our scoring of producibility was arbitrary. Study parts could get a score of not producible (score of 0), partially producible (score of 1), or fully producible (score of 2). Alternative scoring methods may be just as valid. As a robustness check, we therefore rescored the producibility variable in two alternative ways to see whether that affected our inferences. First, we used a binary score in which a study part received a score of 1 if at least one of the study elements was deemed to be producibly described (and 0 otherwise). Second, we used a binary score in which a study part received a score of 1 if all study elements were deemed to be producibly described (and 0 otherwise). Note that both ways correspond to most extreme scoring rules, with the difference between 'not producible' and 'partially producible' being infinitely larger than the difference between 'partially producible' and 'fully producible' (0,1,1), or infinitely smaller (0,0,1). For both these scoring methods, the results that were (not) statistically significant in the original analyses were (not) statistically significant in the new analyses, with the coefficients being in the same direction. The detailed results of the robustness analyses can be found at https://osf.io/3mxfs.

Our results also mimic those of previous studies with regard to explanations for deviations. Like Claesen et al. (2020) and Heirene et al. (2021), we found that authors rarely explain inconsistencies between preregistrations and papers. This is problematic because such omissions mean that readers cannot assess whether the deviations were reasonable and the severity of the test may be compromised (Lakens, 2019). We recommend

researchers to document the deviations from a preregistration explicitly, comprehensively, and transparently, including a rationale for why the deviations occurred and how the deviation could impact the results, perhaps employing Preregistration Planning and Deviation Documentation (Van 't Veer et al., 2019).

While we did count the number of times that the authors explained a deviation from the preregistration, we did not report on whether these deviations were reasonable because we do not presume to have the expertise required to make that judgment for each individual study. However, we do have the wordings used by the authors to explain their deviations, so interested readers could do a deep dive into our data to assess the validity of preregistration deviation explanations in psychology. In general, our data is freely available for anyone to check our coding efforts or to answer their own research questions. We believe the data we collected can be a valuable resource for meta-researchers.

Aside from comparisons with fields in the social sciences, it may also be informative to compare our results to studies in biomedicine, a field that has seen much meta-research on the topic of preregistration (often called registration in this discipline; Rice and Moher, 2019). Researchers in the United States have been mandated to register clinical trials as early as 1997 (Food and Drug Administration Modernization Act of 1997, 1997), making it possible to assess the producibility of these registrations and their consistency with the subsequent report. In general, these studies focus primarily on study outcomes and review studies show that a large proportion of clinical trial papers involves the addition, removal, or change of a primary outcome (Dwan et al., 2013: 40-62%; Jones et al., 2015: 65%; Li et al., 2018: 14% to 100%; Thibault et al., 2021: 10% to 68%). The reviews that also assess other study parts (Li et al., 2018; Thibault et al. 2021) find, like in the social sciences, that the exclusion criteria, sample size, and statistical analysis (including subgroup analyses) are the areas in which discrepancies occur most often. In review, the prevalence of discrepancies between preregistration and paper seems to be similar in the social sciences and biomedical sciences, also in terms of the types of discrepancies. A systematic comparison between the social sciences and biomedical sciences is outside the scope of this paper but would be an interesting meta-research pursuit to follow up on.

While preregistrations serve to lock in temporal relationships between planning and conducting, it should also be noted that the present study assumes that such temporal relationships were guaranteed in all the preregistrations we analyzed. However, preregistrations could be created after experiments have been carried out (Yamada, 2018). This is a problem inherent in preregistration itself, but this type of practice would be less likely to be observed in registered reports, where experimental protocols are peer-

reviewed and almost always revised before experiments are conducted (Chambers & Tzavella, 2022).

Similarly, an alternative to 'regular' preregistration could be analysis blinding, where researchers develop their analysis plan using data in which a third party removed any potentially biasing information. Sarafoglou, Hoogeveen, and Wagenmakers (2023) found that analysis blinding leads to higher consistency between preregistered and actual analysis than preregistration. For example, they found that the analysts in their study who practiced analysis blinding deviated with their exclusion criteria 2% of the time, while that was 16% for those who practiced preregistration. This practice thus seems to be a promising tool for researchers aiming to ensure the confirmatory status of their statistical analyses.

Finally, a way to improve consistency would be to have peer reviewers explicitly compare the preregistration and the actual study. Based on our experience, this comparison is often carried out haphazardly or is not carried out at all. A feasibility study on discrepancy review (TARG Meta-Research Group and Collaborators, 2022) showed that it can be effective and could feasibly be introduced as a regular practice. However, an important issue with this idea is that discrepancy review takes extra time, while reviewers already invest many unpaid hours in peer reviewing for scientific journals. If this burden increases, potential reviewers could become less tempted to accept peer review requests, leading to a potential breakdown of the system. While the feasibility study found that the extra time investment was not excessive, a full trial that looks at secondary effects of discrepancy review is desirable.

In sum, our results extend the results of other studies, making it increasingly clear that, while some researchers are good preregistrationers, much needs to be improved with regard to study preregistration. To unlock the full potential of preregistration, researchers in psychology and likely other fields should aim to write more producible preregistrations, adhere to these preregistrations more faithfully, and in case of deviations, more transparently report them. The creation of more comprehensive templates, and specific training modules to improve preregistration skills would be beneficial in this regard.

References

Bakker, M., Veldkamp, C. L., van Assen, M. A., Crompvoets, E. A., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D. T., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*(12), e3000937. https://doi.org/10.1371/journal.pbio.3000937

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

Bowman, S., DeHaven, A. C., Errington, T., Hardwicke, T. E., Mellor, D. T., Nosek, B. A., & Soderberg, C. K. (2020). OSF Prereg Template. https://doi.org/10.31222/osf.io/epgjd

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*(1), 29-42.

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), 211037. https://doi.org/10.1098/rsos.211037

Dwan, K., Gamble, C., Williamson, P. R., Kirkham, J. J., & Reporting Bias Group. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PloS one*, *8*(7), e66844.

Food and Drug Administration Modernization Act of 1997. (1997). Public Law 105-15. Retrieved from https://www.govinfo.gov/content/pkg/PLAW-105publ115/pdf/PLAW-105publ115.pdf

Hartgerink, C. H., Van Aert, R. C., Nuijten, M. B., Wicherts, J. M., & Van Assen, M. A. (2016). Distributions of p-values smaller than .05 in psychology: what is going on?. *PeerJ*, *4*, e1935.

Haven, T. L., & Van Grootel, D. L. (2019). Preregistering qualitative research. *Accountability in Research*, *26*(3), 229-244.

Heirene, R., LaPlante, D., Louderback, E. R., Keen, B., Bakker, M., Serafimovska, A., & Gainsbury, S. M. (2021). Preregistration specificity & adherence: A review of preregistered gambling studies & cross-disciplinary comparison. *PsyArXiv Preprint*. https://doi.org/10.31234/osf.io/nj4es

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, *62*(3), 221-230.

Li, G., Abbade, L. P. F., Nwosu, I., Jin, Y., Leenus, A., Maaz, M., Wang, M., Bhatt, M., Zielinski, L., Sanger, N., Bantoto, B., Luo, C., Shams, I., Shahid, H., Chang, Y., Sun, G., Mbuagbaw, L., Samaan, Z., Levine, M. A. H., Adachi, J. D., Thabane, L. (2018). A systematic review of comparisons between protocols or registrations and full reports in primary biomedical research. *BMC Medical Research Methodology, 18*(1), 1-20. https://doi.org/10.1186/s12874-017-0465-7

Malich, L., & Munafò, M. R. (2022). Introduction: Replication of Crises-Interdisciplinary Reflections on the Phenomenon of the Replication Crisis in Psychology. *Review of General Psychology*, *26*(2), 127-130.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264-269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Munafò, M. R., Chambers, C. D., Collins, A. M., Fortunato, L., & Macleod, M. R. (2020). Research culture and reproducibility. *Trends in Cognitive Sciences*, *24*(2), 91-93.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., ... & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815-818.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600-2606.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Ofosu, G. K., & Posner, D. N. (2021). Pre-analysis plans: An early stocktaking. *Perspectives on Politics*, 1-17. https://doi.org/10.1017/S1537592721000931.

Olsson-Collentine, A., van Aert, R. C. M., Bakker, M., & Wicherts, J. M. (2023). Meta-Analyzing the Multiverse: A Peek Under the Hood of Selective Reporting. Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/43yae

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. H., Govaart, G. H., Norris, E., O'Mahony, A., Parker, A. J., Todorovic, A., Pennington, C. R., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E., … Aczel, B. (2022). A community-sourced glossary of open scholarship terms. *Nature Human Behavior, 6*, 312–318. https://doi.org/10.1038/s41562-021-01269-4

Pennington, C. R. (2023). *A student's guide to open science: Using the replication crisis to reform psychology*. Open University Press.

Pfeiffer, N., & Call, M. (2022). Surpassing 100,000 Registrations on OSF: Strides in Adoption of Open and Reproducible Research [Blog Post]. Retrieved from https://www.cos.io/blog/surpassing-100000-registrations-on-osf

Preregistration Task Force. (2021). Preregistration Standards for Psychology - the Psychological Research Preregistration-Quantitative (aka PRP-QUANT) Template. ZPID (Leibniz Institute for Psychology). http://dx.doi.org/10.23668/psycharchives.4584

Rice, D. B., & Moher, D. (2019). Curtailing the use of preregistration: A misused term term. *Perspectives on Psychological Science*, *14*(6), 1105-1108.

Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E. J. (2023). Comparing analysis blinding with preregistration in the many-analysts religion project. *Advances in Methods and Practices in Psychological Science*, *6*(1), 25152459221128319.

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K. F., & Altman, D. G. (2010). Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC medicine*, *8*(1), 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.

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

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2), 220346.

TARG Meta-Research Group and Collaborators. (2022). Discrepancy review: A feasibility study of a novel peer review intervention to reduce undisclosed discrepancies between registrations and publications. *Royal Society Open Science*, *9*(7), 220142.

Thibault, R. T., Clark, R., Pedder, H., van den Akker, O. R., Westwood, S., Thompson, J., & Munafo, M. (2021). Estimating the prevalence of discrepancies between study registrations and publications: A systematic review and meta-analyses. *medRxiv*. https://doi.org/10.1101/2021.07.07.21259868

Van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2-12.

Van 't Veer, A. E., Vazire, S., Campbell, L., Feldman, G., Etz, A., & Lindsay, D. S. (2019). Preregistration Planning and Deviation Documentation (PPDD). Retrieved from https://osf.io/ywrqe

Van den Akker, O. R., van Assen, M. A. L. M., Enting, M., de Jonge, M., Ong, H., Rüffer, F. F., … Bakker, M. (2023). Selective Hypothesis Reporting in Psychology: Comparing Preregistrations and Corresponding Publications [MetaArxiv Preprint]. https://doi.org/10.31222/osf.io/nf6mq

Van den Akker, O. R., Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P., ... & Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, 5.

Van Zant, A. B., & Moore, D. A. (2015). Leaders' use of moral justifications increases policy support. *Psychological Science*, *26*(6), 934-943. https://doi.org/10.1177/0956797615572909

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

Wagenmakers, E. J., Wetzels, R., Borsboom, D., Van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on psychological science*, *7*(6), 632-638.

Wicherts, J. M. (2017). The weak spots in contemporary science (and how to fix them). *Animals*, *7*(12), 90.

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

Yamada, Y. (2018). How to crack pre-registration: Toward transparent and open science. *Frontiers in Psychology*, *9*, 1831. https://doi.org/10.3389/fpsyg.2018.01831

CHAPTER 4

Preregistration in practice: A comparison of preregistered and non-preregistered studies in psychology

Olmo R. van den Akker[1], Marcel A. L. M. van Assen[1, 2], Marjan Bakker[1], Mahmoud Elsherif[3], Tsz Keung Wong[1], Jelte M. Wicherts[1]

[1] Department of Methodology and Statistics, Tilburg University, The Netherlands
[2] Department of Sociology, Utrecht University, The Netherlands
[3] Department of Neuroscience, Psychology and Behaviour, University of Leicester, United Kingdom

This paper is published in Behavior Research Methods. See https://doi.org/10.3758/s13428-023-02277-0.

Abstract

Preregistration has gained traction as one of the most promising solutions to improve the replicability of scientific effects. In this project, we compared 193 psychology studies that earned a Preregistration Challenge prize or Preregistration Badge to 193 related studies that were not preregistered. In contrast with our theoretical expectations and prior research, we did not find that preregistered studies had a lower proportion of positive results (Hypothesis 1), smaller effect sizes (Hypothesis 2), and fewer statistical errors (Hypothesis 3) than non-preregistered studies. Supporting our Hypotheses 4 and 5, we found that preregistered studies more often contained power analyses and typically had higher sample sizes than non-preregistered studies. Finally, concerns about the publishability and impact of preregistered studies seem unwarranted as preregistered studies did not take longer to publish and scored better on several impact measures. Overall, our data indicate that preregistration has beneficial effects in the realm of statistical power and impact, but we did not find robust evidence that preregistration prevents *p*-hacking and Hypothesizing After the Results are Known (HARKing).

Keywords: *effect size, HARKing, p-hacking, preregistration, positive results, research impact*

Introduction

Researchers often hypothesize the presence of a causal effect or association between two or more variables. When a study shows evidence for such an effect or association, the result is typically branded as 'positive'. Conversely, when a study does not show such evidence, the result is typically branded as 'negative'. Although finding a positive result is not necessarily the result of better scholarship, positive results are more likely to be published (Dickersin, 1990; Ferguson & Brannick, 2012; Franco, Malhotra, & Simonovits, 2014) and are more often cited (Duyx, Urlings, Swaen, Bouter, & Zeegers, 2017) than negative results. Moreover, peer reviewers more often recommend positive results for publication than negative results because they think positive results contribute more to science (Mahoney, 1977), and researchers write up or submit positive results for publication more often than negative results because they think positive results have more publication potential (Franco, et al., 2014). Further evidence of a bias against negative results comes from studies that find that the vast majority of results in the scientific literature is and was positive (Dickersin, Chan, Chalmers, Sacks, & Smith, 1987; Sterling, 1959), particularly in psychology (Fanelli, 2010), despite the common use of underpowered designs (Bakker, Van Dijk, & Wicherts, 2012). It appears that academics perceive studies with positive results as more valuable than studies with negative results[9], possibly because the dominance of significance testing in many fields (e.g., Hubbard, 2015) leads researchers to equate positive with significance.

The premium on positive results may also shape the behavior of academics in other ways. While carrying out a study, researchers may, consciously or unconsciously, steer their study towards a positive result. Two main examples of this are HARKing (Hypothesizing After the Results are Known (Bosco, Aguinis, Field, Pierce, & Dalton, 2016; John, Loewenstein, & Prelec, 2012; Kerr, 1998; Motyl et al., 2017) and *p*-hacking (John, et al., 2012; Motyl et al., 2017). When researchers HARK, they misattribute a research result to a certain theory *after* distilling the results from the data, which is problematic because one can almost always find something of interest in a given dataset with many variables. When researchers *p*-hack, they make research decisions *contingent on* their data, often with the aim of achieving a *p*-value below .05. These so-called questionable research practices (QRPs) artificially create positive results, as the data does not always warrant the conclusion that an association between variables exists (Murphy & Aguinis, 2019; Simmons, Nelson, & Simonsohn, 2011).

9 Awareness of this issue has prompted the creation of several journals open to (e.g., PLOS One, F1000, and PeerJ), or even dedicated to publishing negative results (e.g., the Journal of Articles in Support of the Null Hypothesis, and the Journal of Pharmaceutical Negative Results).

To prevent researchers from engaging in HARKing and *p*-hacking, it has been suggested that researchers post their hypotheses, study design, and analysis plan online before collecting or looking at any data (Nosek, Ebersole, DeHaven, & Mellor, 2018; Wagenmakers, Wetzels, Borsboom, Van der Maas, & Kievit, 2012). This practice is called preregistration and would help to avoid HARKing because publicizing a study's hypotheses before data is collected makes it impossible for researchers to pretend that they theorized the study results beforehand. Similarly, preregistration would help avoid *p*-hacking because researchers have to specify most of their research decisions before data collection, restricting their freedom to make these decisions contingent on the data. Because preregistration theoretically prevents HARKing and *p*-hacking, preregistered publications should contain a lower proportion of positive results than non-preregistered publications (Hypothesis 1). This study aimed to test this hypothesis for publications in psychology.

A positive result may not be the only desirable outcome. The same may be said for a large effect size since large effect sizes indicate associations of a higher magnitude and thus more convincing evidence (Kelley & Preacher, 2012). Researchers may therefore want to *p*-hack their way to a larger effect size in a similar way as they would to a positive result (Fanelli, Costas, & Ioannidis, 2017; Ioannidis, 2008). Based on that conjecture, we also predicted that effect sizes are on average larger in non-preregistered than in preregistered studies (Hypothesis 2). This predicted effect could be driven by the premium on large effect sizes but could also be a byproduct of the premium on positive results. It could also be that non-preregistered studies have larger effect sizes because they have smaller sample sizes, as positive results require larger effect sizes to be statistically significant in smaller studies (see also Hypothesis 5 below).

Three recent studies directly compared preregistered publications to non-preregistered publications in psychology. First, Schäfer and Schwarz (2019) found a lower proportion of positive results (0.64 vs. 0.79) and lower median effect sizes (0.16 vs. 0.36) in preregistered publications (including registered reports, a type of preregistration where studies are peer reviewed *before* data collection; see Chambers & Tzavella, 2022) than in non-preregistered publications. They did not compare the proportion of positive results in published registered reports and 'regular' preregistered publications but found similar mean effect sizes in published registered reports and 'regular' preregistered publications (0.18 vs. 0.22). Second, Scheel, Schijen, and Lakens (2021) found a lower proportion of positive results in published registered reports than in non-preregistered publications (0.44 vs. 0.96). However, the authors did not compare the magnitudes of effect sizes. Finally, Toth et al. (2021) found that preregistered studies (including registered reports) included a lower proportion of positive results (0.48) than non-preregistered studies (0.66). Additionally, they investigated some other differences between preregistered studies and non-preregistered studies. In line with Bakker et al. (2020), they found that

preregistered studies more often reported a sample size rationale than non-preregistered studies (proportions of 0.72 vs. 0.29), but such rationales were not associated with larger sample sizes. A final result from Toth et al. shows that preregistered studies were more likely to discuss excluded data (0.78 vs. 0.51) and were more likely to have an a priori stopping rule (0.43 vs. 0.02).

Our project differs from these previous studies in four ways. First, we only compared 'regularly' preregistered studies to non-preregistered studies, and thus excluded registered reports. Excluding registered reports allows for a purer assessment of the effect of preregistration, as registered reports also differ from non-preregistered studies in that these reports are adjusted based on peer review.

Second, while two earlier studies did not specifically match preregistered and non-preregistered studies, we linked each preregistered study in our sample to an equivalent non-preregistered study. More specifically, Scheel et al. (2021) used a random sample of 152 psychology publications by searching for the string "test the hypothesis" in the Web of Science ESI database. Schäfer and Schwarz (2019) used a stratified random sample of 900 publications, 10 randomly selected from each of 90 journals that were themselves randomly selected from Web of Science subject categories within psychology (10 per category, but none for mathematical psychology). Only Toth et al. (2021) matched preregistered and non-preregistered studies, by using a combination of (1) non-preregistered studies in papers with an included preregistered study, and (2) non-preregistered studies in papers from the same journal issue (or the same year) as the included preregistered study. In our study, we looked at Web of Science's list of related papers (based on the number of overlapping references) for every preregistered publication and selected the first non-preregistered publication in this list with empirical data that was published in the same year as the preregistered publication. This ensured us that the preregistered and non-preregistered publications (broadly) matched on topic and publication period.

Third, rather than coding a limited set of hypotheses as in the three earlier studies, we aimed to code *all* hypotheses in a study. Scheel et al. (2021) selected only the result of the first hypothesis mentioned in a paper that was explicitly tested, Schäfer and Schwarz (2019) selected the first result related to the key research question, and Toth et al. (2021) selected the results of all hypotheses but only if they were formally stated. In our study, we took a more inclusive approach by assessing the first statistical result for *all hypotheses* in a paper, also those that are not formally stated (i.e., hypotheses that are not listed but can be found in the running text of a preregistration). We already identified hypotheses and the corresponding statistical results from preregistered studies a priori as part of another project (Van den Akker, Van Assen, et al., 2023, see also https://osf.io/z4awv).

Finally, we extend earlier studies by looking at other variables on top of effect size and the proportion of positive results and examining if they differ between preregistered and non-preregistered publications.

In a survey about QRPs, John et al. (2012) asked a sample of psychology researchers whether they ever "rounded off" a *p*-value (e.g., reported a *p*-value of .054 as less than .05). They found that a little over 20% admitted to having done so at least once, and studies screening the psychological literature indeed found that half of all papers reporting significance tests contained at least one inconsistent *p*-value (Nuijten, Hartgerink, Van Assen, Epskamp, & Wicherts, 2016). The 20% rate of admission is relatively low compared to the other QRPs in the John et al. (2012) survey. Interestingly, however, is that the authors also found that a respondent's admission to a relatively rare QRP, like incorrectly rounding off *p*-values, predicted that the respondent also engaged in other QRPs, like failing to report all study's dependent measures or deciding to collect more data after checking whether the results were significant. Incorrectly rounding off *p*-values is a QRP that cannot be prevented by preregistration but based on the finding by John et al. (2012) it may be a proxy of QRPs that *can* be prevented by preregistration like outcome switching or optional stopping (Wicherts, 2017). We therefore expected that incorrectly reported *p*-values are less prevalent in preregistered publications than in non-preregistered publications (Hypothesis 3).

The main benefit of preregistration is that it prevents HARKing and *p*-hacking but preregistration also comes with other benefits (Lakens, 2019; Sarafoglou, Kovacs, Bakos, Wagenmakers, & Aczel, 2022; Wagenmakers & Duthil, 2016). Foremost, preregistering a study requires careful deliberation about the study's hypotheses, research design, and statistical analyses. This deliberation might be spurred on by researchers' use of preregistration templates that provide guidance on what to include in a preregistration and why (e.g., Bowman et al., 2016; Haven & Van Grootel, 2019; Van den Akker et al., 2021). For example, many preregistration templates stress the importance of doing a proper power analysis to determine the study's sample size. We therefore expected that the sample sizes of preregistered studies are based on power analyses more often than the sample sizes of non-preregistered studies (Hypothesis 4).

Moreover, because studies without a power analysis often rely on sample size rationales that lead to relatively low statistical power (Bakker, Hartgerink, Wicherts, & Van der Maas, 2016), we expected that the sample sizes in preregistered studies are larger than the sample sizes in non-preregistered studies (Hypothesis 5). Indeed, Schäfer and Schwarz (2019) found that preregistered publications involved larger sample sizes than non-preregistered publications, and Maddock and Rossi (2001) showed that studies requiring a power analysis as part of a federal funding scheme had a higher power to

detect medium and small effects than other studies. On the other hand, Bakker et al. (2020) found that studies based on preregistration templates recommending power analyses did not have larger sample sizes than studies based on preregistration templates not recommending power analyses.

Some researchers have voiced worries that it is more difficult for preregistered studies than non-preregistered studies to get published. For example, researchers have expressed worries that the restrictive nature of preregistration leads to boring or messy papers without room for unexpected discoveries (Goldin-Meadow, 2016; Kornell, 2013; and see Giner-Sorolla, 2012), making it harder to get them published. However, one could also argue that preregistered studies are more likely to be published because their perceived trustworthiness may make studies with negative results more appealing. Because we do not have information about non-published preregistrations, it is difficult to investigate whether preregistered studies are harder to publish than non-preregistered studies. However, many journals do provide information about the duration of reviews. For preregistered studies, peer review should involve a comparison of the preregistration to the final manuscript, which may cause the review process to take longer. On the other hand, preregistered papers may be of higher quality or may be more clearly reported, which could result in fewer review rounds and a shorter review process. As such, we did not have a clear hypothesis about the association between preregistration and review duration, but we did examine this exploratively.

We also assessed the scientific impact of preregistered publications versus non-preregistered publications. To that end, we looked at three well-known metrics: a publication's number of citations, a publication's Altmetric Attention score, and the impact factor of the publishing journal. The number of citations and the journal impact factor have traditionally been key markers of scientific impact (Mingers & Leydesdorff, 2015). The Altmetric Attention score is relatively new and takes into account less traditional measures of impact like references in news outlets, on blogs, and on social media like Facebook and Twitter (see https://www.altmetric.com/about-our-data/the-donut-and-score for more information). Scientometric studies largely found positive relationships between traditional citation counts and both separate altmetrics (micro-blogging: r = .003, blogs: r = .12, bookmarks from online reference managers: r = .23 for CiteULike, and r = .51 for Mendeley; Bornmann, 2015) as well as the Altmetric Attention score (r = .23; Huang, Wang, & Wu, 2018). We had no a priori hypothesis about the association between preregistration and these three indicators of scientific impact.

Hypotheses

1. Preregistered studies have a lower proportion of positive results than similar non-preregistered studies

2. Preregistered studies contain smaller effect sizes than similar non-preregistered studies
3. Preregistered studies have a lower proportion of gross statistical inconsistencies than similar non-preregistered studies
4. Preregistered studies more often contain a power analysis than similar non-preregistered studies
5. Preregistered studies contain larger sample sizes than similar non-preregistered studies

Method

Sample of preregistered studies

Our sample of preregistered studies was derived from a large-scale project that investigated selective hypothesis reporting (Van den Akker, Van Assen, et al., 2023) that included published papers that earned a Preregistration Challenge prize and published papers that earned a Preregistration Badge prior to 2020. The Preregistration Challenge was a campaign organized from 2017 to 2018 by the Center for Open Science (COS) where researchers could earn $1,000 when they published a preregistered a study. Preregistration badges were also initiated by the COS; journals can decide to hand out preregistration badges to papers that include at least one preregistered study. After excluding registered report studies, studies using secondary data, and studies using non-human subjects, the earlier project included a sample of 459 preregistered studies from 259 papers.

For the current project, we only included studies for which a preregistered statistical result was retrievable in the running text, and we included only the first study of a paper to prevent dependency in the data. This led to a final sample size of 208 studies, which deviates from our preregistered sample of 210 for the following reasons. After preregistering, we noticed that one study had no retrievable result (Banks, Woznyj, Wesslen, & Ross, 2018), while another study involved changes to the preregistration after review, technically qualifying it as a registered report (Goldberg & Carmichael, 2017). While extracting the required information for the remaining 208 papers, we had to exclude an additional 12 papers because they were published in journals that were not listed in the Web of Science Core Collection (which we used to find a control paper, see below), namely: Comprehensive Results in Psychology, Psi Chi Journal of Psychological Research, BMC Psychology, and Wellcome Open Research. We also had to exclude a paper from Psychological Science because we could only find a Corrigendum on Web of

Science, rather than the actual paper. The list of the 193 remaining studies in our sample is available at https://osf.io/xzcnb.

The data collection procedure is detailed in Van den Akker, Van Assen et al. (2023) and an overview of the procedure can be found in the PRISMA flow diagram (Moher, Liberati, Tetzlaff, Altman, & PRISMA Group, 2009) in Figure 1.

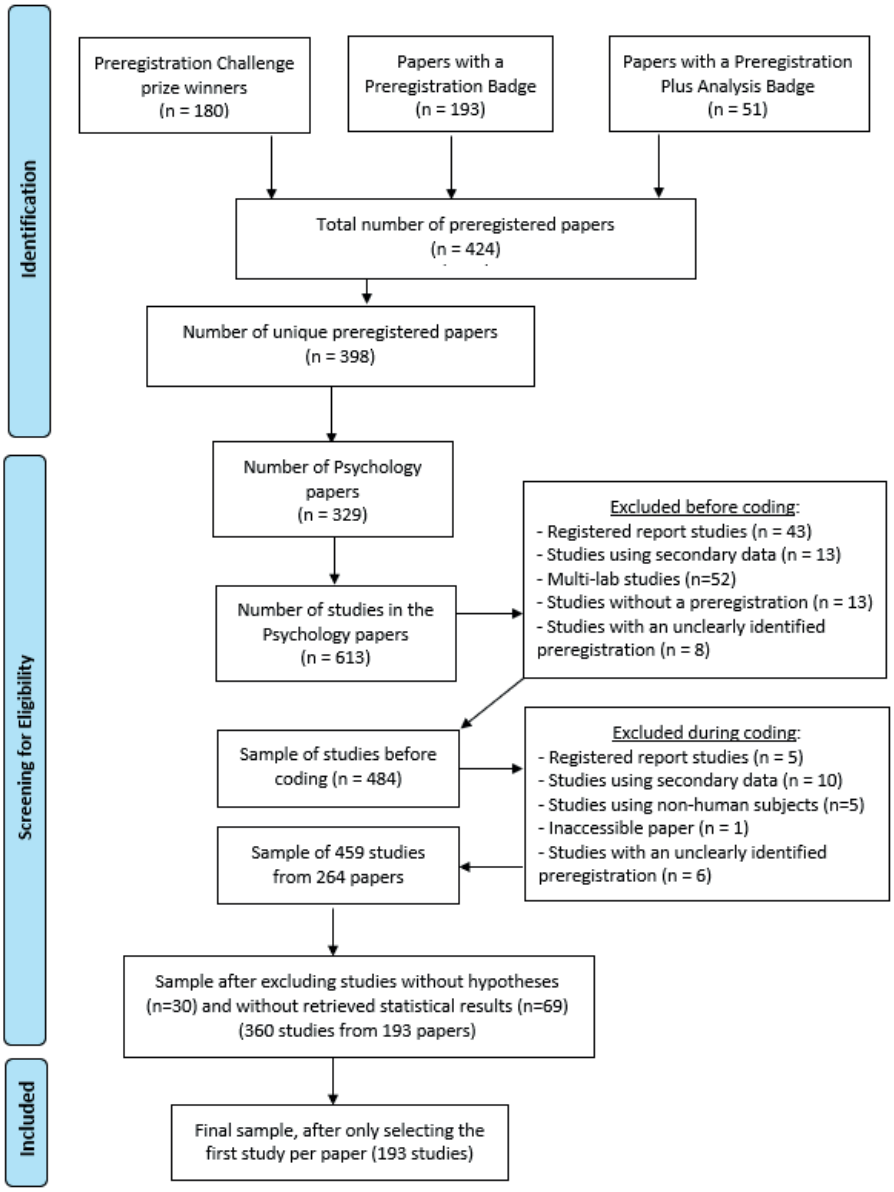


Figure 1. PRISMA flow diagram outlining the full sample selection procedure

Sample of non-preregistered studies

To create a control group for comparison with the preregistered studies in our sample, we linked each preregistered publication in our sample to a non-preregistered publication. We did so by checking Web of Science's list of related papers for every preregistered publication and selecting the first non-preregistered publication from that list that used primary quantitative data and was published in the same year as the related preregistered publication. To check whether publications were preregistered, we searched the publication for the keyword "regist". If that keyword could not be found, we assumed that the publication was not preregistered. We chose Web of Science because it covers established peer reviewed journals and the searches using its Core Collection database are reproducible (in contrast to the searches in Google Scholar, Gusenbauer & Haddaway, 2020) and because comparable studies also used Web of Science (Scheel et al., 2021; Schäfer and Schwarz, 2019), facilitating any comparisons we might want to make.

Our control group was deliberately chosen to mimic our sample of preregistered publications as closely as possible. We preferred this over a random sample of psychology papers because it could be that preregistration is more common in one subfield of psychology than in another. If that is the case, we would compare a skewed sample of preregistered publications to a representative sample of non-preregistered publications. Comparing our sample of preregistered publications to a control group of similar publications is therefore more pertinent. The list of the 193 control studies is available at https://osf.io/xzcnb.

Assessing whether a hypothesis was supported

To assess the proportion of positive results in preregistered studies, we built on the earlier project on selective hypothesis reporting (Van den Akker, Van Assen et al., 2023). In that project, hypotheses were identified in both preregistrations and their corresponding papers to see whether selective reporting took place. Hypotheses were identified by using the keywords "replicat", "hypothes", "investigat", "test", "predict", "examin", and "expect", and included if the authors predicted a relationship between two or more variables using any of these keywords (disregarding manipulation checks, and checks of statistical assumptions). We then tried to match a statistical result in the published paper to each of the preregistered hypotheses. If a match was found, we inspected the statistical output and concluded that there was a positive result when $p < .05$, or when Bayes Factors (BFs) were either smaller than 1/3 or larger than 3. If multiple results matched the preregistered hypothesis, we chose the first statistical result mentioned in that paper. If the authors specifically stated that they used a significance level smaller than .05 or a Bayes Factor criterion smaller than 1/3 or larger than 3, we used the authors' inference criteria. The end result was a fraction: the number of hypotheses with a matched positive result divided by the total number of matched results. If no *p*-value or

Bayes Factor could be retrieved, we coded this as missing data ('NA'). The protocol for the assessment of the support for preregistered hypotheses can be found at https://osf.io/fdmx4 (for preregistrations) and https://osf.io/uyrds (for publications).

We extracted the proportion of positive results for our sample of non-preregistered publications by inspecting the results sections of these publications and flagging all statistical results that were not part of a manipulation check, a check of statistical assumptions, or an exploratory test. Using all the flagged results in a paper, we calculated the proportion of positive results by assessing *p*-values and Bayes factors as we did for preregistered publications.

Effect sizes

We use the Fisher-transformed Pearson's *r* as our common effect size measure because it was found to be the most frequently reported effect size in Schäfer and Schwarz (2019) and because its interpretation is relatively straightforward. If *r* was not specified for a certain result, we calculated it based on the *t*-value or *F*-value and the accompanying degrees of freedom. In case the *F*-statistic was based on multiple contrasts or variables ($df_1 > 1$), we followed the Open Science Collaboration (2012) and computed the "correlation coefficient per degree of freedom" (r/df_1). Table 1 provides the formulas we used for these calculations. If a statistical result was not based on a *t*- or *F*-statistic (but on a *z*- or X^2-statistic for example) or a statistical result did not include sufficient information to calculate *r* we did not include the result.

Table 1

Formulas used to compute the Correlation Coefficients per Degree of Freedom

Statistic	Transformation
t	$r = \sqrt{\frac{t^2 * \frac{1}{df}}{t^2 * \frac{1}{df} + 1}}$
F	$r = \sqrt{\frac{F * \frac{df_1}{df_2}}{F * \frac{df_1}{df_2} + 1}} \sqrt{\frac{1}{df_1}}$

Note. $df = N - 1,\ df_1 = n_1 - 1,\ df_2 = n_2 - 1$

Reporting errors

We used the *statcheck* web app (Rife, Nuijten, & Epskamp, 2016) in June 2022 to count the number of "grossly" incorrectly reported *p*-values (i.e., *p*-values that do not match

their accompanying test statistic and degrees of freedom *and* for which the inconsistency changes the statistical conclusion; Nuijten, et al., 2016) and used the proportion of gross errors per study as our dependent variable. Because we used *statcheck,* we only included results in the analyses that could be extracted using that program (i.e., *t*, *F*, *r*, χ^2, and *z*-statistics). Exploratively, we also looked at "regularly" incorrectly reported *p*-values (i.e., *p*-values that do not match their accompanying test statistic and degrees of freedom but for which the inconsistency did not change the statistical conclusion).

Power analysis and sample size

We determined sample size and the presence of a power analysis as part of a project assessing the effectiveness of preregistration (Van den Akker, Bakker, et al., 2023; see https://osf.io/x7qgh for the coding protocol). Of concern here is the effective sample size (i.e., the sample size that is used to draw conclusions about the hypothesis selected using the hypothesis selection protocol, see https://osf.io/z4aw). We used the same procedure to determine the presence of a power analysis and the sample size for preregistered and non-preregistered publications.

Review duration

To compare the duration of reviews of preregistered publications and non-preregistered publications, we checked the article history of these publications to extract the submission date and the date of acceptance. The difference between the two in days was used as our measure of review duration. One potential issue with this method is that journals may not always accurately register submission dates and acceptance dates. However, we expected any inaccuracies to occur equally frequently for preregistered and non-preregistered publications. We used the same procedure to determine the submission and acceptance dates for preregistered and non-preregistered publications.

Scientific impact

We coded the number of citations, the journal impact score, and the Altmetric Attention score all in the same week (May 16th 2022 until May 20th 2022). We looked up the number of citations by searching for a manuscript on the Web of Science Core Collection database, the 2019 journal impact factor by using Web of Science's Journal Citation Reports, and the Altmetric Attention score by using the Altmetric.com bookmarklet. We used the same procedure to determine these metrics for preregistered and non-preregistered publications.

Hypothesis tests

We tested our preregistered hypotheses (see https://osf.io/mpd3u) with five bivariate regressions in which the independent variable was whether a study was preregistered or not. The dependent variables in these regressions were the proportion of positive

results in the study (Hypothesis 1), Fisher transformed effect size (Hypothesis 2), the proportion of statistical inconsistencies in the study (Hypothesis 3), the presence of a power analysis in the study (Hypothesis 4), and the log of the sample size of the study (Hypothesis 5). Hypotheses 1, 3, and 4 were tested using logistic regressions. Hypothesis 2 was tested using a multilevel linear regression with two levels: statistic (level 1), and study (level 2). Hypothesis 5 was tested using a linear regression.

Power analyses for the five hypotheses are reported in our preregistration but were based on the planned 210 rather than the actual 193 included studies. Re-running the power analyses using the same anticipated effect sizes as in our preregistration but using the actual sample size of 193 resulted in a statistical power of 1.00 for Hypothesis 1, of 0.97 for Hypothesis 2, of 0.83 for Hypothesis 3, and of 0.63 for Hypothesis 4. The updated power calculations are available at https://osf.io/m47f6.

The R-code for all hypothesis tests can be found at https://osf.io/sujfa. All data used for the analyses can be found at https://osf.io/pqnvr.

Results

We found no support for Hypothesis 1 that the proportion of positive results was lower in preregistered studies (0.69, SD = 0.38) than in non-preregistered studies (0.68, SD = 0.25), $\beta = 0.01$, 99% CI = [-0.56, 0.59], $z(366) = 0.05$, $p = .96$. For this analysis we deviated from our preregistration and excluded all null hypotheses from the sample of preregistered studies. We felt that this was warranted because we realized, in hindsight, that the calculation of the proportion of positive results for non-preregistered studies assumed that all hypotheses were directional. Excluding preregistered null-hypotheses therefore makes the above comparison between preregistered and non-preregistered studies fairer. When we did include the null-hypotheses, as we preregistered, the statistical result was as follows: $\beta = 0.01$, 99% CI = [-0.57, 0.58], $z(366) = 0.03$, $p = .98$.

We did not find support for Hypothesis 2 either. While effect sizes were numerically smaller on average for preregistered (0.29, SD = 0.24, median = 0.28) than non-preregistered studies (0.36, SD = 0.25, median = 0.30), this difference was not statistically significant, $\beta = -0.04$, 99% CI = [-0.12, 0.04], $t(1794.2) = -1.36$, $p = .175$.

Hypothesis 3 was not supported by our data either. Preregistered publications (0.001, SD = 0.01) did not have a lower proportion of gross statistical inconsistencies than non-preregistered publications (0.005, SD = 0.03), $\beta = -1.33$, 95% CI = [-7.46, 4.80], $z(216) = -0.42$, $p = .671$. When we looked at all statistical inconsistencies (including ones where

the statistical conclusion did not change), we also did not find a lower proportion of inconsistencies in preregistered publications (0.03, SD = 0.15) than in non-preregistered publications (0.09, SD = 0.19), β = -1.19, 95% CI = [-2.51, 0.13], $z(213) = -1.76$, $p = .08$.

In line with Hypothesis 4, we found that sample sizes in preregistered studies (0.55) were more often based on a power analysis than sample sizes in non-preregistered studies (0.23), $\beta = 1.38$, 99% CI = [0.81, 1.95], $z(383) = 6.17$, $p < .0001$. Accordingly, we also found support for Hypothesis 5: the sample sizes of preregistered studies (mean = 959.0, median = 216) were larger than the sample sizes of non-preregistered studies (mean = 536.6, median = 116), $\beta = 0.45$, 99% CI = [0.14, 0.76], $t(384) = 3.72$, $p = .0002$.

Preregistered exploratory analyses

We employed four bivariate regressions to explore whether preregistration influenced review duration (Exploration 1), the log of the number of citations (Exploration 2), the log of journal impact factor (Exploration 3), and the log of Altmetric Attention score (Exploration 4).

We did not find evidence that the review time of preregistered studies (257.9 days, SD = 176.6) was different from the review time of non-preregistered studies (269.4 days, SD = 213.1), $\beta = -11.54$, 95% CI = [-54.2, 31.2], $t(318) = -0.53$, $p = .597$.

Interestingly, for measures of scientific impact, the results did highlight an effect of preregistration. Preregistered publications received more citations (18.3, SD = 24.6) than non-preregistered publications (15.1, SD = 18.4), $\beta = 0.20$, 95% CI = [0.01, 0.40], $t(384) = 2.09$, $p = .038$, using $\alpha = .05$. Preregistered publications (103.9, SD = 204.0) also received a higher Altmetric attention score than non-preregistered publications (28.3, SD = 63.0) and were published in journals with a higher impact factor (4.1, SD = 1.4 vs. 3.0, SD = 1.6), Altmetric score: $\beta = 1.27$, 95% CI = [0.26, 0.44], $t(373) = 6.98$, $p < .0001$; impact factor: $\beta = 0.35$, 95% CI = [0.92, 1.63], $t(375) = 7.53$, $p < .0001$.

Conclusion and Discussion

In this project, we compared studies that earned a Preregistration Challenge prize or Preregistration Badge with similar studies that were not preregistered. Unexpectedly, we did not find that preregistered studies had a lower proportion of positive results than non-preregistered studies (Hypothesis 1) nor that they had smaller effect sizes (Hypothesis 2). Moreover, preregistered studies did not include fewer statistical inconsistencies than non-preregistered studies, as we expected (Hypothesis 3). We did find support for Hypothesis 4 and Hypothesis 5: preregistered studies more often contained a power

analysis and had larger sample sizes than non-preregistered studies. Our preregistered exploratory analyses found that there was no difference in review times for both study types, and that preregistered studies were more impactful in terms of citations, Altmetric attention scores, and journal impact factors than non-preregistered studies.

The higher statistical power and larger sample sizes in preregistrered studies compared to non-preregistered studies are important, considering earlier findings that sample sizes across psychology are often insufficient to find meaningful effects (Bakker et al., 2016; Szucs & Ioannidis, 2017). In line with our finding regarding statistical power, Maddock and Rossi (2001) found that federally funded studies (that typically included an a priori power analysis) had higher average power than studies that did not receive such funding (and typically did not include an a priori power analysis). Prior research on the link between a study's preregistration and sample size is mixed: Schäfer and Schwarz (2019) found that preregistered studies had larger sample sizes than non-preregistered studies for between-subject designs, but smaller sample sizes for within-subject designs.

Not finding an association between preregistration and positive results or effect size contrasts with earlier research (Schäfer and Schwarz, 2019; Scheel et al., 2021; Toth et al., 2021). For the proportion of positive results in preregistered studies, we would expect that our estimate would be higher than earlier estimates that were based on samples including registered report studies. This expectation stems from the idea that regular preregistrations and registered reports both prevent *p*-hacking and HARKing due to increased transparency, but registered reports additionally prevent publication bias because editors accept or reject the paper before the results of the study are known. Insofar that the samples are comparable, our estimate of 0.68 falls as expected above prior estimates by Scheel et al. (2021; 0.44, registered reports only), Schäfer and Schwarz (2019; 0.64, registered reports and regularly preregistered studies), and Toth et al. (2021; 0.48, registered reports and regularly preregistered studies). As a robustness check, it would be useful to compare our estimate to the estimates based only on the regularly preregistered studies in Schäfer and Schwarz, and Toth et al., thereby filtering out the influence of registered reports. However, neither study disclosed the particular studies they coded so this proved impossible.

Surprisingly, the proportion of positive results we found in non-preregistered studies (0.69) is lower than estimates from previous work (Fanelli, 2010: 0.92; Schäfer and Schwartz, 2021: 0.79; Scheel et al., 2021: 0.96; Sterling, Rosenbaum, & Weinkam, 1995: 0.96), with one exception (Toth et al., 2021: 0.61). The heterogeneity in the five estimates can at least partly be explained by the different methods that were used to retrieve the statistical results from the non-preregistered studies. Fanelli and Scheel et al. assessed whether authors concluded to have found positive (full or partial) or negative

(null or negative) support for the first hypothesis in the paper. One explanation for their very high estimates is that they also counted partially positive results as positive, thereby possibly including results with spin (e.g., results that the authors claimed as positive while the *p*-value was marginally significant, see Olsson-Collentine, Van Assen, & Hartgerink, 2019). Schäfer and Schwartz first identified the key research question of a study based on the title and abstract, and then extracted the first reported effect that unambiguously referred to that key research question. As Fanelli, Scheel et al., and Schäfer and Schwarz all focused on the first or key hypothesis in the paper, their high estimates may be explained by a focus on the study's pivotal hypothesis. In contrast, Toth et al. counted all hypotheses that were formally stated in the introduction section, and for which an explicit statistical conclusion could be found elsewhere in the paper. Similarly, in our study, we extracted all statistical results in the results section of non-preregistered studies except for checks of manipulations and statistical assumptions. We contend that the inclusion of other than pivotal statistical results lowered our and Toth et al.'s estimates of the proportion of positive results.

Our results-oriented approach was decided on because experiences from other projects (Van den Akker, Van Assen et al., 2023; Van den Akker, Bakker, et al., 2023) led us to expect that it would be too difficult to find the first or most important hypothesis in non-preregistered studies. While this approach seems inclusive and encompassing, it could be that we included statistical results that were not meant as hypothesis tests. However, we would argue that readers of papers tend to see all results in a results section as hypothesis tests unless they are clearly labeled as a check or as exploratory. Regardless, there is currently no well-validated method to assess the proportion of positive results of hypothesis tests in non-preregistered studies. It would help if researchers highlight in their papers what results in their results section are meant as hypothesis tests, and whether these tests were preregistered or not.

Another explanation for the similar proportion of positive results in preregistered and non-preregistered studies is that sample sizes were larger for the former (see the results for Hypothesis 5). Assuming true effect sizes are equal for preregistered and non-preregistered studies (which would be in line with the results for Hypothesis 2), we would expect higher statistical power and, thus a higher proportion of positive results for the preregistered studies. As a crude test of this explanation, we ran the analysis of Hypothesis 1 again but with sample size as a control variable. Controlling for sample size, we again did not find a difference in the proportion of positive results in preregistered and non-preregistered studies ($\beta_1 = 0.0026$, $t(364) = 0.078$, $p = .938$), demonstrating that the role of this alternative explanation is probably minor.

Finally, it is important to note that preregistered publications and non-preregistered publications can differ in yet other aspects. For example, it is likely that researchers self-select to carry out a preregistration, and researchers who preregister may be more junior, more conscientious, or more concerned with abiding by responsible research practices like preregistration. Because of these differences, causal claims about the effect of preregistration on the proportion of positive results or effect size are difficult to make. Future studies may aim to identify the characteristics of preregistering and non-preregistering researchers so that these variables could be included as control variables in studies like ours.

Taking all results together, we conclude that preregistered studies are of higher quality than non-preregistered studies in the sense that they more often contain power analyses than non-preregistered studies and typically have higher sample sizes. Moreover, concerns about the publishability of preregistered versus non-preregistered studies seem unwarranted as preregistered studies do not take longer to publish and have more impact. Our study does not provide convincing evidence that preregistration prevents *p*-hacking and HARKing of results reported in the main text of a study, as both the proportion of positive results and effect sizes are similar between preregistered and non-preregistered studies. Future research could shed more light on this. One could, for example, include preregistration as a moderator in meta-analyses on theoretically similar effects. If non-preregistered studies typically involve larger observed effects, this could be an indication of biases (publication bias and/or QRPs). Such empirical work, combined with the results from the current study, would improve our understanding of preregistration and would allow us to make more evidence-based claims about its practical value.

Open Practices Statement

The preregistration, the data, the materials and all other information relevant to this study can be found in our OSF repository at https://osf.io/pqnvr.

References

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

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543-554.

Bakker, M., Veldkamp, C. L., van den Akker, O. R., van Assen, M. A., Crompvoets, E., Ong, H. H., & Wicherts, J. M. (2020). Recommendations in pre-registrations and internal review board proposals promote formal power analyses but do not increase sample size. *PLOS One*, *15*(7), e0236079.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology*, *33*(4), 445–459. https://doi.org/10.1007/s10869-017-9528-3

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology, 69*(3), 709-750.

Bornmann, L. (2015). Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. *Scientometrics*, *103*(3), 1123-1144.

Bowman, S. D., DeHaven, A. C., Errington, T. M., Hardwicke, T. E., Mellor, D. T., Nosek, B. A., & Soderberg, C. K. (2016, January 1). OSF Prereg Template. https://doi.org/10.31222/osf.io/epgjd

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science 351*(6280), 1433-1436.

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*(1), 29-42.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *JAMA*, *263*(10), 1385-1389.

Dickersin, K., Chan, S. S., Chalmersx, T. C., Sacks, H. S., & Smith Jr, H. (1987). Publication bias and clinical trials. *Controlled Clinical Trials*, *8*(4), 343-353.

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

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLOS One*, *5*(4), e10068.

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences, 114*, 3714–3719.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*(1), 120.

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

Goldberg, M. H., & Carmichael, C. L. (2017). Language complexity, belief-consistency, and the evaluation of policies. *Comprehensive Results in Social Psychology*, *2*(2-3), 1-17.

Goldin-Meadow, S. (2016, August 31). Why preregistration makes me nervous. *APS Observer.* Retrieved from https://www.psychologicalscience.org/observer/why-preregistration-makes-me-nervous

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, *11*(2), 181-217.

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

Haven, T. L., & Van Grootel, D. L. (2019). Preregistering qualitative research. *Accountability in Research*, *26*(3), 229-244.

Huang, W., Wang, P., & Wu, Q. (2018). A correlation comparison between Altmetric Attention Scores and citations for six PLOS journals. *PLOS One*, *13*(4), e0194962.

Hubbard, R. (2015). Corrupt research: The case for reconceptualizing empirical management and social science. Sage Publications.

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

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

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137.

Kornell, N. (2013, July 29). Some concerns on regulating scientists via preregistration. Psychology Today. Retrieved from https://www.psychologytoday.com/za/blog/everybody-is-stupid-except-you/201307/some-concerns-regulating-scientists-preregistration

Lakens, D. (2019, November 18). The Value of Preregistration for Psychological Science: A Conceptual Analysis. https://doi.org/10.31234/osf.io/jbh4w

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*, 161-175.

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, *246*(1), 1-19.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group*, T. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264-269.

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., ... & Yantis, C. (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.

Murphy, K. R., & Aguinis, H. (2019). HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology, 34*(1), 1- 17.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600-2606. https://doi.org/10.1073/pnas.1708274114

Nuijten, M. B., Hartgerink, C. H., Van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205-1226.

Olsson-Collentine, A., Van Assen, M. A., & Hartgerink, C. H. (2019). The prevalence of marginally significant results in psychology over time. *Psychological Science*, *30*(4), 576-586.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657-660.

Rife, S. C., Nuijten, M. B., Epskamp, S. (2016). statcheck: Extract statistics from articles and recompute p-values [web application]. Retrieved from http://statcheck.io

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E. J., & Aczel, B. (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, *9*(7), 211997.

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

Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard Psychology literature with Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007467.

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

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. *The American Statistician, 49*(1), 108–112.

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

Toth, A. A., Banks, G. C., Mellor, D., O'Boyle, E. H., Dickson, A., Davis, D. J., ... & Borns, J. (2021). Study preregistration: An evaluation of a method for transparent reporting. *Journal of Business and Psychology*, *36*, 553-571.

Van den Akker, O. R., Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P., ... & Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology, 5*. https://doi.org/10.17605/OSF.IO/U68Y7

Van den Akker, O. R., Bakker, M., van Assen, M. A. L. M., Pennington, C. R., Verweij, L., Elsherif, M. M., … Wicherts, J. M. (2023). The effectiveness of preregistration in psychology: Assessing preregistration strictness and preregistration-study consistency. Retrieved from https://osf.io/preprints/metaarxiv/h8xjw

Van den Akker, O. R., van Assen, M. A. L. M., Enting, M., de Jonge, M., Ong, H., Rüffer, F. F., … Bakker, M. (2023). Selective Hypothesis Reporting in Psychology: Comparing Preregistrations and Corresponding Publications. https://doi.org/10.31222/osf.io/nf6mq

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

Wagenmakers, E. J., Wetzels, R., Borsboom, D., Van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632-638.

Wicherts, J. M. (2017). The weak spots in contemporary science (and how to fix them). *Animals*, *7*(12), 90.

CHAPTER 5

5

How Do Psychology Researchers Interpret the Results of Multiple Replication Studies?

Olmo R. van den Akker[1], Jelte M. Wicherts[1], Linda Dominguez Alvarez[1], Marjan Bakker[1], Marcel A. L. M. van Assen[1, 2]

[1] Department of Methodology and Statistics, Tilburg University, The Netherlands
[2] Department of Sociology, Utrecht University, The Netherlands

This paper is published in Psychonomic Bulletin & Review. See https://doi.org/10.3758/s13423-022-02235-5

Abstract

Employing two vignette studies, we examined how psychology researchers interpret the results of a set of four experiments that all test a given theory. In both studies, we found that participants' belief in the theory increased with the number of statistically significant results, and that the result of a direct replication had a stronger effect on belief in the theory than the result of a conceptual replication. In Study 2, we additionally found that participants' belief in the theory was lower when they assumed the presence of *p*-hacking, but that belief in the theory did not differ between preregistered and non-preregistered replication studies. In analyses of individual participant data from both studies, we examined the heuristics academics use to interpret the results of four experiments. Only a small proportion (Study 1: 1.6%; Study 2: 2.2%) of participants used the normative method of Bayesian inference, whereas many of the participants' responses were in line with generally dismissed and problematic vote counting approaches. Our studies demonstrate that many psychology researchers underestimate the evidence in favor of a theory if one or more results from a set of replication studies are statistically significant, highlighting the need for better statistical education.

Keywords: *multi-study paper, replication, statistical misinterpretation, heuristics, Bayesian inference, vote counting*

Introduction

Imagine the following situation: you have conducted four psychology experiments that all tested a given theory. All four experiments had a power of 50% and two out of the four experiments yielded statistically significant results. Assuming that your belief in the validity of the theory before conducting these experiments was 50%, what would your current belief in the theory be? Given that the contemporary psychology literature mainly includes statistically significant results (Fanelli, 2010, 2012; Hartgerink, Van Aert, Nuijten, Wicherts, & Van Assen, 2016; Sterling, Rosenbaum, & Weinkam, 1995), one might think the theory is valid only when all experiments yielded significant results. However, this would be mistaken. Using Bayes' rule, we can calculate that the probability of the theory being correct when two out of four results are significant is as high as 97% (see Box 1). Based on the wealth of studies that show that academics often have trouble with correctly interpreting statistical results (Aczel et al., 2018; Fischhoff, & Beyth-Marom, 1983; Gigerenzer, 2018; Hoekstra, Finch, Kiers, & Johnson, 2006; Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Kahneman, & Tversky, 1973), we suspect that this result would surprise many readers.

We carried out two between-subjects vignette studies to test academics' statistical intuitions when assessing the results of multiple (four) experiments. We decided to carry out Study 2 in 2022 because Study 1 was conducted in 2014 and a lot has changed in the meantime. For example, *p*-hacking (Friese & Frankenbach, 2020; Head, Holman, Lanfear, Kahn, & Jennions, 2015; Wicherts et al., 2016), and the statistical interpretation of replication studies (Klein et al., 2018; Maxwell, Lau, & Howard, 2015) have been discussed widely, and numerous educational materials appeared on these topics (Azevedo et al., 2019; Da Silva Frost & Ledgerwood, 2020). Increased awareness about these issues raised the question whether the results of Study 1 are still relevant to how researchers today think about the results of replication experiments. Below we will first outline the research questions that were common to both studies and then outline the research questions that were unique to each study.

To examine the relationship between belief in the theory and the number of statistically significant results, in both studies we varied the number of significant results from zero to four, out of four experiments. We expected a positive relationship but we did not have any predictions about the nature of this relationship (e.g., linear, quadratic, etc.). To uncover whether different types of replications differentially affect belief in the theory, we also presented the experiments as being either a direct or conceptual replications. We expected academics to evaluate a significant conceptual replication as providing more evidence for the validity of the theory than a significant direct replication. We based this prediction on the strong focus on novelty and generalizability in academia,

where academics might find a replication using different methods or designs more convincing (Crandall & Sherman, 2016; Giner-Sorolla, 2012; Schmidt, 2009).

In both studies, we used Bayes' rule to calculate participants' accuracy: how well their stated belief in the theory corresponded to the correct computation according to Bayes' theorem (see Box 1). Assuming that experience and knowledge positively predict the accuracy of participants' posterior beliefs, we expected a positive association between accuracy and participants' number of peer-reviewed publications and (self-reported) statistical knowledge.

In Study 1 only, we randomly allocated participants to the role of 'author' or 'reviewer' to examine if authors and reviewers differ in their assessment of the set of results. Specifically, we asked participants in these roles two questions: (1) if they would submit the set of results to a journal (author) or recommend it for publication (reviewer), and (2) whether they would run an additional replication experiment before possibly submitting (author) or whether they would demand the authors to carry out an additional replication experiment (reviewer). We did not have expectations regarding these questions.

In Study 2 only, we included a regular condition and a preregistration condition. The regular condition was equivalent to the vignette of Study 1 in that the replication studies were said to be typical for psychology. In the preregistration condition, the replication studies were said to be preregistered and aligned with their preregistrations. Thus, the two conditions would differ in the degree to which *p*-hacking could have occurred. *P*-hacking involves collecting or selecting data or analyses to render nonsignificant results significant (Head et al., 2015) and may lead to false positive results (i.e., results that are an artefact of the researcher's decisions instead of evidence in favor of an underlying theory). We therefore expected that the participants in the regular condition would have a lower belief in the theory than in the preregistration condition when confronted with significant results. In Study 2 only, we also explicitly asked whether participants considered *p*-hacking when assessing belief in the theory. For scenarios with statistically significant results, we expected participants who considered the possibility of *p*-hacking to show lower belief in the theory than participants who did not.

Finally, using individual participant data from both studies, we sought to categorize participants' assessments of the results of the four experiments into several heuristics used to weigh the evidence. We now present the methods and results of Study 1 and Study 2, and then provide more information about the Heuristic Analyses.

Box 1 – Assessing belief in the theory using Bayes' theorem

The validity of a theory (H_A) given multiple (non)significant experiments (i.e., the probability that the theory is correct given the data) depends on the power of the experiments and can be readily computed with Bayes' theorem. Formally:

$$P(H_A\ is\ true \mid data) = \frac{P(H_A\ is\ true \cap data)}{P(data)}$$

$$= \frac{P(data \mid H_A\ is\ true) * P(H_A\ is\ true)}{P(data \mid H_A\ is\ true) * P(H_A\ is\ true) + P(data \mid H_0\ is\ true) * P(H_0\ is\ true)}$$

$$= \frac{\binom{N}{k} * (1-\beta)^k * \beta^{(N-k)} * P(H_A\ is\ true)}{\binom{N}{k} * (1-\beta)^k * \beta^{(N-k)} * P(H_A\ is\ true) + \binom{N}{k} * \alpha^k * (1-\alpha)^{(N-k)} * P(H_0\ is\ true)}$$

If we assume that $P(H_A is\ true) = P(H_0\ is\ true) = 0.5$ these terms drop out. And since $\binom{N}{k}$ also drops out we obtain:

$$p(H_A|K) = \frac{(1-\beta)^k * \beta^{(N-k)}}{(1-\beta)^k * \beta^{(N-k)} + \alpha^k * (1-\alpha)^{(N-k)}} \qquad (1)$$

where (1-β) is power, α is the significance level, *N* is the total number of experiments, and *k* is the number of statistically significant results. For our example in the introduction with a power of .50, a significance level of .05, and two out of four significant results, the probability that the theory is correct is .965 (see Equation 2).

$$p(H_A|2) = \frac{(1-0.50)^2 * 0.50^{(4-2)}}{(1-0.50)^2 * 0.50^{(4-2)} + 0.05^2 * (1-0.05)^{(4-2)}} \approx 0.965 \qquad (2)$$

When only one out of four results is significant, and using the same values for α and β, the probability that the theory is correct is still .593. In case of a statistical power of .80, the posterior belief in the theory is lower than when power is 0.50 for zero, one, and two statistically significant results (see Figure 1).

Box 1 – Assessing belief in the theory using Bayes' theorem (Continued)

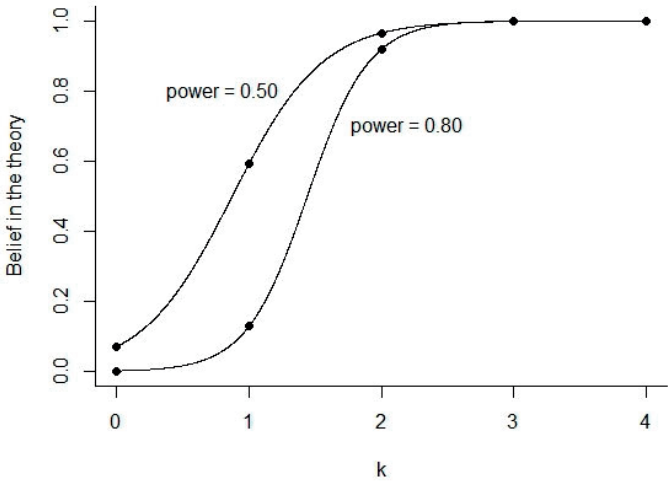


Figure 1. Belief in the theory based on Bayesian inference, as a function of statistical power (.50 and .80) and the number of statistically significant results, *k*, given prior probabilities equal to .5. The beliefs in the theory for *k*=0,1,2,3,4 are [.071, .593, .965, .998, .999] and [.002, .013, .919, .999, 1.000] for a statistical power of 0.50 and 0.80, respectively.

Method of Study 1

Sample selection

We sampled participants from social and experimental psychology who commonly conduct (as researcher) or judge (as editor) experimental research consisting of multiple studies. In both social and experimental psychology, a single study is typically not considered to be sufficient to test a theory (Murayama, Pekrun, & Fiedler, 2013), and multiple study papers are the norm (Giner-Sorolla, 2012). Using Web of Science, we selected empirical journals in social and experimental psychology published in English that had a 5-year Impact Factor higher than two in the year 2012. From social psychology, we included *Journal of Personality and Social Psychology, Journal of Experimental Social Psychology, Personality and Social Psychology Bulletin,* and *European Journal of*

Social Psychology. From experimental psychology, we included *Journal of Experimental Psychology – General, Journal of Experimental Psychology – Human Perception and Performance, Journal of Experimental Psychology – Learning, Memory, and Cognition, Quarterly Journal of Experimental Psychology,* and *Cognition & Emotion*. In total, we collected 2,449 references to articles published in 2012 and 2013. We included one additional journal for the subfield of experimental psychology to keep the number of articles between the subfields approximately equal, resulting in 1,126 articles for social psychology and 1,323 articles for experimental psychology.

To contact researchers, we retrieved contact information of the corresponding authors from Web of Science. After deleting duplicate email addresses, we ended up with 1,810 unique researchers. To contact editors, we looked up the editorial board of the selected journals and searched online for the contact details of the represented (associate) editors and reviewers, yielding contact details for 834 unique editors. Of the 2,644 potential participants of Study 1 (1,322 in each assigned role) 52 emails proved invalid, so only the remaining 2,592 researchers and editors received an invitation to participate in the survey. The invitations were sent at the end of May and the beginning of June 2014. We sent a reminder two weeks later and stopped collecting data two weeks after the reminder. After excluding the non-completers 228 academics participated in the authors' version and 277 academics in the reviewers' version, the response rates being 17.6% and 21.4%, respectively.

Procedure and materials

We used Qualtrics to conduct the survey for Study 1. Before presenting the survey, the participants in the sample were randomly assigned to the authors' version (see https://osf.io/aufn2) or to the reviewers' version (see https://osf.io/hqx4e) of the survey. The study involved eight different scenarios, each presenting the results of four experiments. All presented scenarios stated that other researchers had previously published the results of one experiment, A, and found a statistically significant effect in line with a given theory. The vignette then stated that the participant had conducted ('authors') or was asked to review ('reviewers') four experiments that replicated the findings of the original study. The first new experiment (A') was a direct replication of the earlier experiment, whereas the other three experiments (B, C, and D) were conceptual replications. Participants were presented with four out of eight possible scenarios in Table 1, where each scenario had a different number of significant results, *k*. All participants were told to imagine that their prior belief in the theory before seeing the results of the four experiments was 50% and that the number of participants, the costs of all experiments, the nominal significance level, and the statistical power in all five experiments (including the original experiment A) were typical for experimental studies in psychology.

Table 1
Summary of the Eight Different Scenarios

Scenario	A'	B	C	D	*K*
1	O	O	O	O	0
2	X	O	O	O	1
3	O	X	O	O	1
4	X	O	X	O	2
5	O	X	X	O	2
6	X	O	X	X	3
7	O	X	X	X	3
8	X	X	X	X	4

Note. X indicates significant results, whereas O indicates non-significant results. A' indicates a direct replication, whereas the remaining letters indicate conceptual replications. *K* refers to the number of statistical results in each scenario.

After providing informed consent, participants read the introduction stating that a distinction was made between direct replications and conceptual replications. We clarified that a direct replication uses the same method as the original study and tries to reproduce it as closely as possible, while a conceptual replication may use different methods or operationalizations (Schmidt, 2009). Next, participants were successively shown a table for each of the scenarios. Those tables included information about which of the experiments showed a statistically significant result and were shown at the top of every page, preventing participants from forgetting the results in the scenario (see Table 1 for all eight possible scenarios). In six scenarios (those with 1 to 3 significant results) either A' or B was significant. For instance, Scenario 2 and 3 both have one significant experimental result, but in Scenario 2 study A' (direct replication) is significant and in Scenario 3 study B (conceptual replication) is significant. Participants were randomly assigned to either scenario 2 or 3 (both with k=1), either scenario 4 or 5 (both with k=2), and either scenario 6 or 7 (both with k=3). In addition, participants were randomly assigned to either scenario 1 (with k=0) or scenario 8 (with k=4). Participants thus considered four scenarios in total.

After each scenario, participants indicated their belief in the theory on the basis of the presented evidence by means of a slider bar, with points going from low probability (0%) to high probability (100%) of the theory being correct. Participants could indicate using a text box whether they missed any information while reading the scenarios. Next, they had to indicate the statistical power, ranging between 0 and 1, they had in mind while answering the questions. The survey ended with four demographic and work-related questions; gender, year that they obtained their doctorate, number of peer-

reviewed papers published (using six categories: < 5, 5-15, 16-30, 31-50, 51-100, and > 100), and the participant's self-reported statistical knowledge on a scale from 0 (poor) to 10 (excellent)[10]. Finally, we gave participants the option to write down any comments regarding the survey or research project and we thanked them for their participation. The responses of all participants of Study 1 can be found at https://osf.io/k4us3.

To assess participants' accuracy in estimating the probability that the theory is correct, we created an accuracy variable; for every *k* in every different scenario, we calculated the root mean squared error (RMSE) comparing the participant's belief in the theory with the normative Bayesian prediction (see Equation 1, Box 1):

$$RMSE = \sqrt{\frac{1}{j}\sum_{i=1}^{j}(m_i - d_i)^2} \quad (3)$$

where *j* refers to the number of responses of a participant (typically 4), *m* refers to the responses as predicted by Bayesian inference, and *d* refers to the beliefs stated by the participants.

In the Qualtrics survey of Study 1, we also included a question whether participants would submit (as author) or accept (as reviewer) a set of studies for publication, a question whether they would require an additional experiment, and a dichotomous question about their belief in the theory. Due to space constraints, we do not present the results related to these questions in this paper but interested readers can find them at https://osf.io/vnws7.

Method of Study 2

Sample selection

For Study 2 we searched the Web of Science Core Collection for journal articles from the research areas social psychology and experimental psychology published in the years 2020 (searched on 8-2-2021) and 2021 (searched on 3-12-2021). We did not include any papers published before the 2020s because researchers' ideas about open science and statistical inference seem to be changing fast and we wanted to be able to draw conclusions about the current timeframe. Our search yielded 14,940 (for the year 2020) + 14,480 (for the year 2021) references, each with one email address. After removing duplicates, we were left with 21,120 unique email addresses. 2,632 of those were out of office, while 794 emails proved invalid. The remaining 19,694 researchers received

10 Note that participants had to provide statistical power themselves because we did not explicitly present them with the statistical power of the experiments in the vignette. In Study 2, we did provide this information.

an invitation to participate in the week of 14 February 2022. Those who did not reply received a reminder two weeks later. We stopped data collection on 25 May 2022. In total, 1,334 researchers participated in Study 2, equally distributed over the preregistration condition and the regular condition. The response rate was 6.8%.

Procedure and materials

We again used Qualtrics to develop the survey for Study 2 (see https://osf.io/xycte for the full survey). Some participants of Study 1 indicated in open comments that the survey questions were difficult to answer because the vignette lacked detailed information. Notably, twenty-two participants (4.4%) expressed confusion about the role of statistical power in our experiment. To prevent any issues, we provided more specific information in Study 2 as we stated that the significance level (= .05) and statistical power if the theory is valid (= .50) were the same for each experiment (A, A′, B, C, and D). This seemed to have helped as only fourteen participants (1.0%) expressed confusion about statistical power in Study 2. We also emphasized in the vignette of Study 2 that the prior belief of fifty-fifty pertains to the situation *after* seeing the result of the original experiment, but *before* seeing the results of the replication experiments.

The main change from Study 1 was that we randomly assigned Study 2 participants to a preregistration condition or a regular condition. In the preregistration condition participants were told that the design and analysis of all replication experiments were preregistered and that they were conducted and analyzed in line with their preregistrations. In the regular condition (corresponding to the author vignette of Study 1), participants were told to assume that the replication experiments were typical for experimental studies in psychology. We also explicitly asked about *p*-hacking: "Throughout this study, did you consider the possibility that the researcher in the scenarios made decisions through which they, consciously or subconsciously, directed their experiments toward a statistically significant result?" The responses of the participants of Study 2 can be found at https://osf.io/5dfhs.

The design, hypotheses, and statistical analyses for Study 2 were preregistered (see https://osf.io/f7vsq). Hypotheses 1 through 4 (regarding the number of statistically significant results, replication status, the number of publications, and statistical knowledge) were limited to participants in the preregistration condition to make comparison with Study 1 possible. Hypotheses 5 and 6 (regarding preregistration status, and the possibility of *p*-hacking) were related to all participants. For completeness we also ran the first four hypotheses using all participants (see https://osf.io/f7ymv for the results of these analyses).

Results of Study 1 and Study 2

All analyses were carried out using R version 4.1.2. The R-code used for the analyses can be found at https://osf.io/jq3w7 (Study 1) and https://osf.io/4cx6w (Study 2).

Table 2 shows the average belief in the theory (0-100%) for each number of statistically significant results, and the composition of significant results (direct or conceptual replication significant) for Studies 1 and 2. We used multilevel linear regression to test the hypothesized association between the number of significant results and belief in the theory. The dependent variable was a logit transformation of belief divided by 100, which makes the effective relationship between the independent variables and belief non-linear, while preserving the linear model. As expected, average belief increased with the number of statistically significant results (Study 1: β = 0.74, 95% CI = [0.69, 0.80], $p < .001$; Study 2: β = 0.74, 95% CI = [0.69, 0.79], $p < .001$).

Unexpectedly, in both studies, participants' average beliefs in the theory were higher when the *direct* replication was significant than when the *conceptual* replication was significant (Study 1: β = -0.13, 95% CI = [-0.21, -0.05], $p = .0017$; Study 2: β = -0.23, 95% CI = [-0.31, -0.15], $p < .001$). Average belief in the theory was between 1.78% points (k=1) and 2.75% points (k=3) higher for direct than for conceptual replications in Study 1, and between 1.61% points (k=1) and 6.38% points (k=3) higher in Study 2 (see Table 2).

Table 2.

Mean (Standard Deviation) of Belief in the Theory in Percentages, for Every Number of Significant Results, k, and for the Different Compositions of Statistical Results

	Participants' mean belief in Study 1	Participants' mean belief per composition in Study 1		Participants' mean belief in Study 2	Participants' mean belief per composition in Study 2	
		A' significant	B significant		A' significant	B significant
k=0	25.33 (16.83)	-	-	23.98 (18.50)	-	-
k=1	33.04 (17.14)	33.93 (17.82)	32.15 (16.42)	33.51 (19.04)	34.36 (20.35)	32.75 (18.13)
k=2	49.34 (15.27)	50.65 (15.27)	48.00 (15.19)	46.67 (17.52)	48.65 (17.49)	43.62 (18.06)
k=3	65.77 (16.18)	67.13 (16.11)	64.38 (16.16)	61.15 (18.77)	64.16 (17.16)	57.78 (19.86)
k=4	73.02 (17.65)	-	-	71.08 (18.47)	-	-

Note. k refers to the number of significant results within the scenario. 'A' significant' means that the direct replication was significant, 'B significant' that the conceptual replication was significant.

For our next hypotheses, we measured the accuracy of participants' posterior beliefs by comparing their responses to the responses predicted by Bayesian inference. In Study 1 we based this on the participants' reported power, thereby excluding participants that

did not provide a power level. In Study 2 we used a power of 0.5 as this was the power level disclosed in the vignette. We regressed participants' root mean squared error (see Equation 2) on Papers published[11] and Statistical knowledge, and found number of published papers to linearly predict higher accuracy in Study 1 (β= 0.0057, 95% CI = [0.0021, 0.0093], p = .0019), but found no such association in Study 2 (β = -0.0002, 95% CI = [-0.0004, 0.000002], p = .0537). For statistical knowledge, we found a nonsignificant association in both Study 1 (β = -0.0034, 95% CI = [-0.007, 0.00015, p = .061) and Study 2 (β = 0.008, 95% CI = [0.002, 0.014], p = .0135).

In Study 2, we also looked at the influence of preregistration status and the possible presence of *p*-hacking. Contrary to our hypothesis, we found no difference between the preregistration condition and the regular condition, β = 0.02, 95% CI = [-0.10, 0.13], p = .783 (Model 2C in Table 3). However, as expected, we found that participants had lower beliefs in the theory when they considered the researcher in the vignette to have used *p*-hacking (β = -0.31, 95% CI = [-0.43, -0.20], p < .001 (Model 2D in Table 3). As a manipulation check, we also checked whether participants more often took into account *p*-hacking in the regular condition (38.2%) than in the preregistration condition (33.9%). This was indeed the case, t(4177.6) = 2.93, p = .003.

Heuristic Analyses

Averaged results of Study 1 aligned with a heuristic where the prior belief of 50% is averaged with the percentage of statistically significant results (see Table 2). However, the results also indicated that participants varied in how *k* affected their beliefs (i.e., multilevel analyses highlighted a random slope). Therefore, we decided to analyze the data for each participant to find out which heuristics may have been used by *individual* academics. We did this for the data of both studies.

11 Even though the variable 'Papers published' was ordinal and not continuous in Study 1, a plot of the predicted values and residuals indicated no reason to suspect that the linearity assumption was violated.

Table 3.

Fixed Effects Estimates (Top) and Variance Estimates (Bottom) for the Multilevel Linear Regression of Belief in the Theory on the Number of Significant Results, the Composition of Significant Results, Preregistration Status, and the Possible Presence of p-hacking – for All Participants in Study 1 (Models 1A and 1B), Only Participants in the Preregistration Condition in Study 2 (Models 2A and 2B), and All Participants in Study 2 (Models 2C and 2D).

Parameters	Model 1A	Model 1B: 1A + Conceptual	Model 2A	Model 2B: 2A + Conceptual	Model 2C: 2A + Preregistration	Model 2D: 2A + *p*-hacking
Regression coefficients (fixed effects)						
Intercept	-1.63 (0.074) *	-1.57 (0.076) *	-1.67 (0.07) *	-1.45 (0.075) *	-1.62 (0.06) *	-1.49 (0.06) *
Level 1						
K	0.74 (0.028) *	0.74 (0.028) *	0.74 (0.26) *	0.69 (0.03) *	0.71 (0.02) *	0.72 (0.02) *
Conceptual	-	-0.13 (0.041) *	-	-0.23 (0.04) *	-	-
Preregistration	-	-	-	-	0.02 (0.06)	-
p-hacking	-	-	-	-	-	-0.31 (0.06) *
Variance components (random effects)						
Residual	0.44 (0.66)	0.43 (0.65)	0.84 (0.92)	0.48 (0.70)	0.66 (0.81)	0.58 (0.76)
Intercept	2.15 (1.47)	2.15 (1.47)	2.03 (1.43)	2.42 (1.55)	2.07 (1.44)	2.06 (1.43)
Slope	0.28 (0.53)	0.28 (0.53)	0.30 (0.55)	0.34 (0.58)	0.23 (0.47)	0.24 (0.49)
r(intercept, slope)	-0.78	-0.78	-0.74	-0.73	-0.75	-0.77

Notes. Standard errors are in parentheses. *k* refers to the number of significant results within the scenario. Conceptual is a binary variable that takes on the value of 1 if the conceptual replication was significant and 0 otherwise. Preregistration is a binary variable that takes on the value of 1 if the participant was allocated to the preregistration condition and 0 if the participant was allocated to the regular condition. *p*-hacking is a binary variable that takes on the value of 1 if the participant indicates to have taken into consideration *p*-hacking in their responses and 0 if they did not indicate this. * $p < .001$.

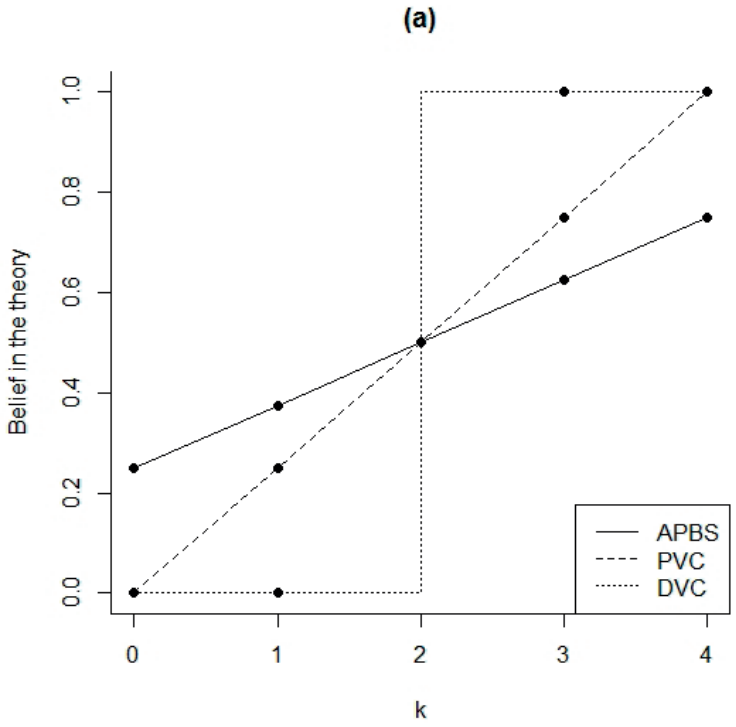


Figure 2. An overview of the three non-normative heuristics that are potentially being used by academics. "APBS" refers to averaging prior belief and significance, "PVC" refers to proportional vote counting, and "DVC" refers to deterministic vote counting.

Method Heuristic Analyses

The analyses of individual participant data were preregistered on May 19, 2018 (Study 1: https://osf.io/hjkpx) and 16 February 2022 (Study 2: https://osf.io/f7vsq). To allow accurate preregistration for Study 1, a research assistant blinded the data using a blinding protocol, R-code, and mock data can be found in the folder named 'Mock Data for Study 1' at https://osf.io/2g4wf.

In addition to Bayesian inference (Box 1), we considered three potential heuristics that academics may have used when interpreting the outcomes of multiple experiments. We label these heuristics "deterministic vote counting", "proportional vote counting", and "averaging prior belief with significance". The predictions of these three heuristics are shown in Figure 2. For simplicity's sake, none of the heuristics take into account the participants' assigned role (relevant to Study 1 only) nor the type of replication (relevant to both studies).

Deterministic vote counting and proportional vote counting are based on the possibility that academics interpret null hypothesis significance test results dichotomously (Hoekstra et al., 2006; Rosenthal & Gaito, 1963, 1964) and therefore weigh the evidence by counting the number of (non-)significant outcomes in a set of studies (Hedges & Olkin, 1980). Deterministic vote counting occurs when researchers believe the theory is true if the proportion of significant results exceeds 0.5, will believe the theory is false when that proportion is below 0.5, and have a 50/50 belief if the proportion equals 0.5. Academics employing proportional vote counting equate their belief in the theory to the proportion of significant results. Finally, when academics employ the heuristic of averaging, they simply average their prior belief in the theory before seeing the results of the four experiments (50/50 in our scenarios) with the proportion of significant results. The three heuristics will be analyzed together with the normative Bayesian inference approach outlined in Box 1.

Adhering to our preregistration, we discarded participants whose belief in the theory showed no (weakly) monotonic increase in the number of significant results (Study 1: N = 70, 13.9%; Study 2: N = 329, 24.7%), participants with three or four (out of four) missing values on the belief variable (N = 47, 9.3%; only relevant for Study 1), and/or participants with a missing value on self-assessed power (N = 1272, 24.2%; only relevant for Study 1). The remaining sample for Study 1 involved 312 participants, of which 152 were presented the author's version and 160 the reviewer's version. The remaining sample for Study 2 involved 1,005 participants, of which 493 were allocated to the preregistration condition and 512 to the regular condition.

The remaining responses were analyzed using a Bayesian model in which we calculated participants' posterior probabilities of using one of the four heuristic models (averaging prior belief and significance, proportional vote counting, deterministic vote counting, or Bayesian inference) given their responses, with prior model probabilities equal to .25 for each of the four models. In our 'weak' classification we allocate a participant to the model with the highest posterior probability (at least .25). In our 'strong' classification we allocate a participant to a model if their posterior probability for the model exceeds .75, which corresponds to a Bayes Factor of at least 3.

The posterior probability of a participant using model H_i given the data $X = \{x_1, \ldots, x_4\}$ is calculated as

$$P(H_i|X) = \frac{P(X|H_i)}{P(X|H_1)+P(X|H_2)+P(X|H_3)+P(X|H_4)} \quad (3)$$

assuming a uniform prior ($P(H_j) = .25$), and $P(X|H_j)$ denoting the likelihood of the data X (four responses) given model H_j. The likelihood of each response given a model is a

normal density with mean μ as determined by that model and standard deviation σ, truncated at 0 and 1. The standard deviation σ reflects the "random decision error" of participants. In our analysis, we used two levels of random decision error, σ = 0.10 and σ = *q*, where *q* was derived by taking each participant's lowest RMSE out of the four RMSE values (one for each model) and taking the average across all participants of those minimum values. Hence, the value of σ = *q* signifies the average misfit of participants with their best-fitting model. We chose a value of σ = 0.10 a priori based on our own statistical intuitions. More details about this procedure can be found in the preregistrations of these analyses at https://osf.io/hjkpx (Study 1) and https://osf.io/f7vsq (Study 2).

To avoid participants being classified into a heuristic while their response pattern does not fit well with any of the models, we also compared participants' response patterns against a benchmark heuristic. This benchmark heuristic is the participant's belief averaged across conditions, or simply a horizontal line corresponding to that participant's average belief. For example, if a participant stated a belief in the theory of 30%, 60%, 70%, and 100% for *k*=1,2,3,4 respectively, their average belief is 260/4 = 65%. Note that the benchmark heuristic is dependent on the data unlike any of the other heuristics. We assessed fit using the root mean squared error (Equation 2), and only allocated individuals to a model if its RMSE was lower than for the benchmark heuristic. This held for both the weak and strong classification procedures.

In Study 2 we added an explicit question about the heuristic used by the participants: "We specified four strategies that researchers may use to assess the probability of the theory being correct in the scenarios we presented. Do you consider one of them applicable to your responses throughout this study? If not, please explain what reasoning you did use to arrive at your responses."

One major alteration was made from the preregistration of the Heuristic Analysis in Study 1 because we mistakenly assumed that participants were told that the statistical power in all studies was 0.5 (like we did state in Study 2). Instead, participants were told that the power of the experiments was typical for psychology experiments. Participants thus had to imagine and report this power value themselves, which influenced the predictions based on Bayesian inference. Because of this oversight we had to calculate those predictions anew. The mean (standard deviation), and mode of self-reported statistical power in Study 1 were 0.67 (0.20), and 0.80. The post-preregistration analysis code can be found at https://osf.io/q2n7y.

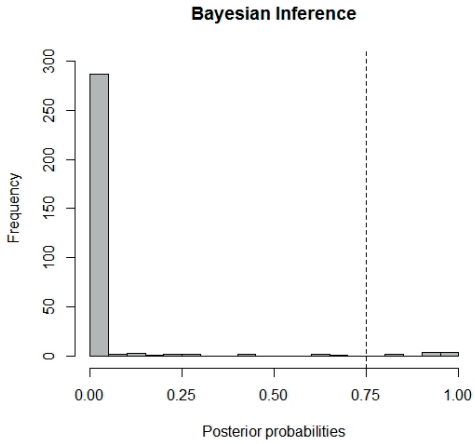
No alterations were made with regard to the preregistration of Study 2.

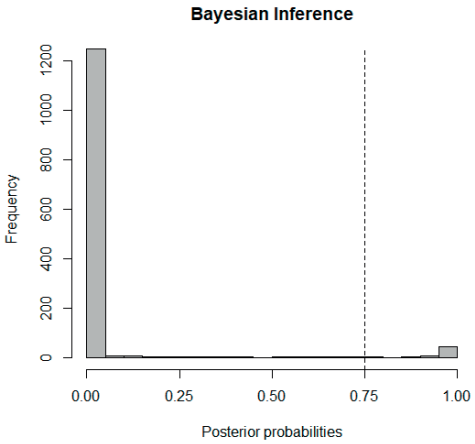
Results of Heuristic Analyses

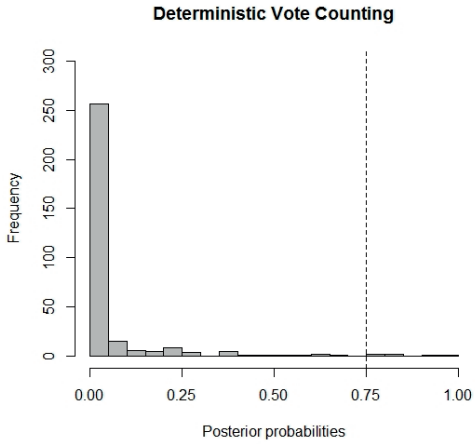
The distributions of the analyzed participants' (N = 312 in Study 1 and N = 1,005 in Study 2) posterior probabilities of using a particular model are shown in Figure 3. As evidenced by the low frequency of high posterior probabilities in the upper two panels, only a few participants appear to have used Bayesian inference. In contrast, the high frequency of high posterior probabilities in the bottom two panels of Figure 3 suggest that many participants averaged their prior belief with the number of significant results.

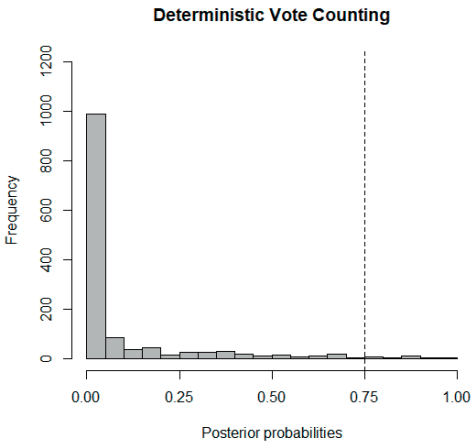
To assess the robustness of our results to alternative analytic choices, we carried out 3 (participants that faced k=0-3, participants that faced k=1-4, and the whole sample of participants) χ 2 (q, the mean of the participants' lowest RMSEs, and the a priori determined 0.1 as random decision errors) = 6 analyses for Study 1 and two analyses (q and 0.1 as random decision errors) for Study 2. For all of the analyses we implemented the weak and strong classification procedure. Because the results of all eight analyses were qualitatively similar (see an overview of all results of Study 1 at https://osf.io/wuje4 and all results of Study 2 at https://osf.io/sw7g5) we decided to only present here the weak and strong classification for the whole sample of participants with $\sigma = q = 0.118$ (Study 1) and $\sigma = q = 0.149$ (Study 2). Histograms depicting the number of participants in every category can be found in Figure 4a (strong categorization) and Figure 4b (weak categorization).

Relatively few participants used the normative "Bayesian inference" approach (Study 1 - Strong categorization: N = 6 (1.6%), Weak categorization: N = 8 (2.1%); Study 2 - Strong: N = 29 (2.2%), Weak: N = 33 (2.5%)) and "deterministic vote counting" (Study 1: Strong: N = 6 (1.6%), Weak: N = 11 (2.9%); Study 2 – Strong: N = 22 (1.6%), Weak: N = 74 (5.5%)). In contrast, a substantial number of participants used "proportional vote counting" (Study 1 - Strong: N = 49 (12.5%), Weak: N = 113 (29.6%); Study 2 – Strong: N = 43 (3.2%), Weak: N = 316 (23.7%)), and "averaging prior belief and significance" (Study 1 - Strong: N = 74 (18.9%), Weak: N = 109 (28.5%); Study 2 – Strong: N = 289 (21.7%), Weak: N = 430 (32.2%)). Using strong categorizations, we could not assign 177 participants (45.2%) and 622 (46.7%) participants in Studies 1 and 2, respectively. This happened because the RMSE for neither heuristic exceeded the RMSE of the benchmark heuristic, or because posterior probabilities of all heuristics were below 0.25. In the weak categorization we could not assign 71 participants (18.6%) in Study1, and not assign 152 (11.4%) participants in Study 2, because neither heuristic outperformed the benchmark heuristic in RMSE. Finally, in line with our preregistration we also distinguished participants with a response pattern whose belief in the theory did not show an expected (weak) monotonic increase in the number of significant results. Such an "irregular" response pattern was relatively common (Study 1: 70 participants, 18.3%; Study 2: 329 participants, 24.7%).


Bayesian Inference
Frequency
0
50
100
150
200
250
300
0.00
0.25
0.50
0.75
1.00
Posterior probabilities


Bayesian Inference
Frequency
0
200
400
600
800
1000
1200
0.00
0.25
0.50
0.75
1.00
Posterior probabilities


Deterministic Vote Counting
Frequency
0
50
100
150
200
250
300
0.00
0.25
0.50
0.75
1.00
Posterior probabilities


Deterministic Vote Counting
Frequency
0
200
400
600
800
1000
1200
0.00
0.25
0.50
0.75
1.00
Posterior probabilities

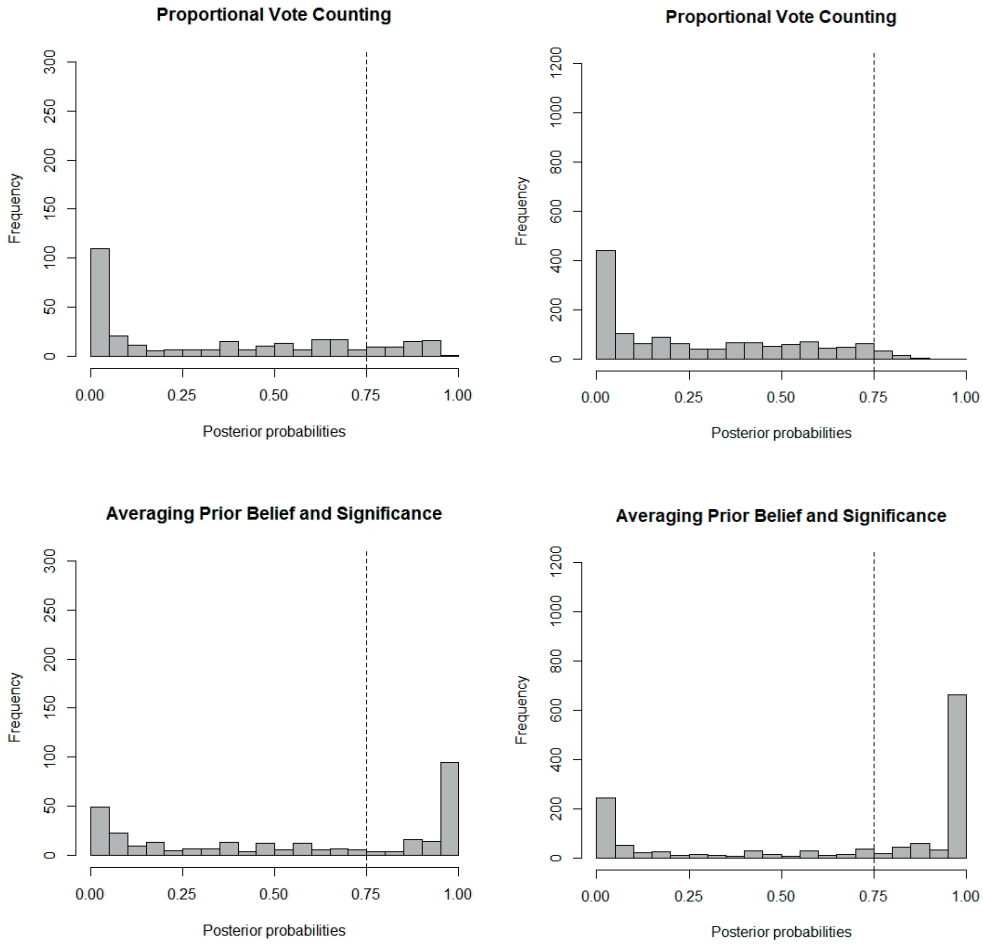


Figure 3. Frequency distributions of the participants' posterior probabilities of using a model given that they use that model or one of the other models, in the situation where the standard deviation is *q* and where all participants are included (on the left for Study 1, on the right for Study 2). The dotted line represents the threshold value for which that model is three times more likely than the other models combined.

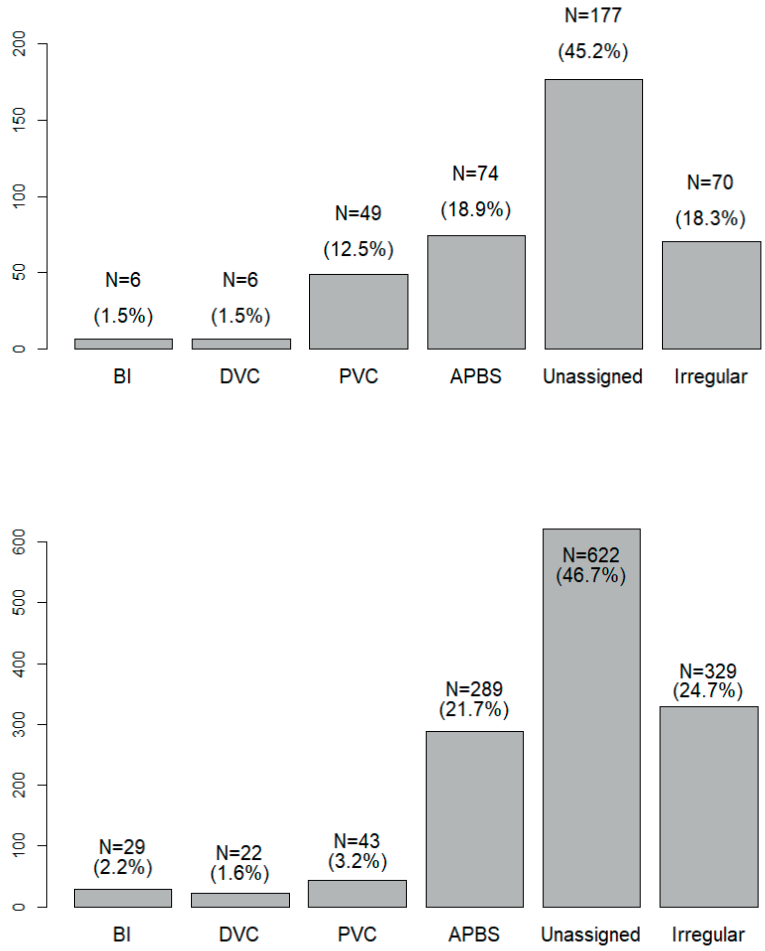


Figure 4a. Histogram presenting the strong categorization for Study 1 (top) and Study 2 (bottom). "BI" refers to Bayesian inference, "DVC" refers to deterministic vote counting, "PVC" refers to proportional vote counting, and "APBS" refers to averaging prior belief and significance.

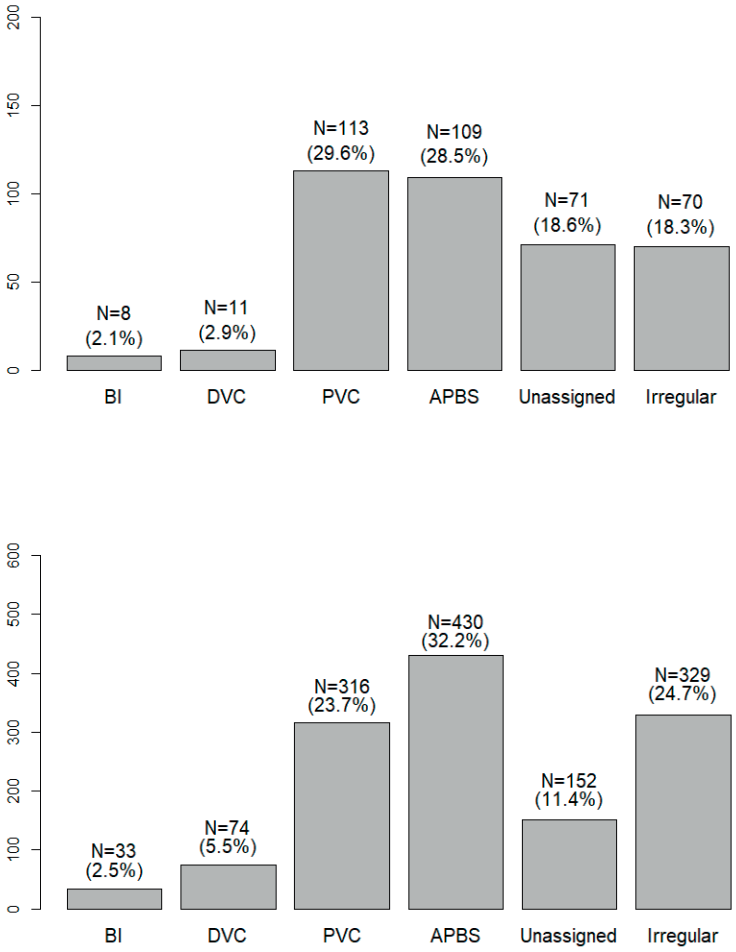


Figure 4b. Histogram presenting the weak categorization for Study 1 (top) and Study 2 (bottom). "BI" refers to Bayesian inference, "DVC" refers to deterministic vote counting, "PVC" refers to proportional vote counting, and "APBS" refers to averaging prior belief and significance.

Conclusion and Discussion

We studied how psychological researchers interpret a set of four replication experiments with varying statistical significance. Across two vignette studies we found that, on average, the number of significant results was positively related to researchers' belief in the underlying theory. Contrary to our expectations, we found that researchers valued direct replications more than conceptual replications when deciding on the validity of a theory, although this effect was small in both studies. The premium of direct replications over conceptual replications in our studies is surprising in the light of papers that question the importance of direct replications (Cesario, 2014; Schmidt,

2009) and the finding that direct replications are less often published (Makel, Plucker, & Hegarty, 2012). It is less surprising in light of the current popularity of large-scale direct replication efforts (Dang et al., 2021; Elliott et al., 2021; Klein et al., 2022). One thing to note when interpreting this result is that people's judgments tend to be influenced by initially presented values (Furnham & Boo, 2011; Tversky & Kahneman, 1974). Because the direct replication was always listed first in the table outlining the results, this "anchoring effect" could be an alternative explanation for the stronger effect of the direct replication compared to the conceptual replication.

In Study 2, we unexpectedly found that participants' belief in the theory did not differ when they assessed a set of preregistered versus a set of non-preregistered studies with statistically significant results. Perhaps our manipulation of preregistration was not strong enough, although we found that participants took *p*-hacking into account more often in the regular condition (38.2% of participants) than in the preregistration condition (33.9% of participants) indicating that our manipulation worked at least to some extent. Moreover, participants' belief in the theory in scenarios with statistically significant results was lower for those who considered *p*-hacking on behalf of the vignette researcher than for those who did not. Combining these findings, we can conclude that psychology researchers are skeptical of statistically significant results when they consider the possibility of *p*-hacking, but that they are also skeptical about the ability of preregistration to effectively prevent *p*-hacking. The latter makes sense in light of findings that preregistrations are not always sufficiently strict to prevent *p*-hacking and are also often not adhered to exactly (Bakker et al., 2020; Van den Akker, 2021).

In the Heuristic Analyses, we zoomed in on individual participant data and categorized participants' answers into three heuristics and the normative approach of Bayesian inference. Only six out of the 312 analyzed participants (1.6%) in Study 1, and 29 out of 1,334 participants (2.2%) in Study 2 used Bayesian inference, showing that few participants accurately incorporated important parameters like power (1-β) and the significance level (α) into their decisions. Instead, a large proportion of participants (27-33% using our strong categorization, 61% using our weak categorization) used (partial) vote counting approaches that underestimate the evidence in favor of a theory if two or more out of four results are statistically significant. Additionally, we were not able to categorize a substantial number of participants (45-47% using our strong categorization, 11-19% using our weak categorization), and another group of participants (18-25%) showed an irregular response pattern in which their belief in the theory did not rise with an increase in the number of statistically significant results. Taking these results together, we can conclude that many participants used invalid vote-counting or unknown approaches when interpreting situations with multiple experimental results. Future research could expand on the current study by exploring different heuristics.

A limitation of our study is the stylized nature of the vignette experiments. Indeed, many participants (Study 1: 56.8%; Study 2: 39.0%) expressed that they would prefer to have more information available in the vignette to inform their decisions. This indicates that our results may not accurately map onto real research scenarios. Although we acknowledge that practicing academics may use other available information to ground their beliefs, we were primarily interested in the effects of replication type and preregistration, and therefore designed our vignettes to vary these factors. Future research may examine what other factors affect academics' belief in a theory. One factor that may be particularly interesting is the number of experiments because including more experiments would make it easier to distinguish the vote counting rules from Bayesian inference.

Another limitation relates to our method of categorization. We preregistered an elaborate Bayesian method to categorize participants into heuristic categories (see https://osf.io/hjkpx for the preregistration related to Study and https://osf.io/f7vsq for the preregistration related to Study 2) but there are many other ways to do this. To assess the validity of our categorization method, in Study 2 we explicitly asked participants whether they used one of the four heuristics we preregistered. We measured the association between this self-categorization and our own categorizations and found a Cramer's V of 0.667. This strong association (detailed at https://osf.io/f7ymv) suggests that our method of categorization is largely in line with how the participants themselves thought of their strategies, supporting the validity of our method.

In summary, we found that psychology researchers have poor intuitions when it comes to interpreting a set of mixed experimental replication results. These poor statistical intuitions can lead to the suppression of non-significant findings (publication bias; see Ferguson & Brannick, 2012; Levine, Asada, & Carpenter, 2009), and may lead to inefficient use of resources as both authors and reviewers may require more studies to be run. Moreover, they may lead researchers to engage more frequently in *p*-hacking (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011; Wicherts, 2017). Poor statistical intuitions not only create incorrect interpretations of experimental results, but also introduce biases in the scientific literature. To avoid this, we need improved education about the interpretation of mixed results. More specifically, we would do well to discourage vote counting heuristics, which continue to appeal to many yet have been shown to be biased over 40 years ago (e.g., Hedges & Olkin, 1980). Instead, we need to focus our educational efforts on the role that Bayes' rule plays in statistical inference, possibly combined with teaching how experimental results can be synthesized using meta-analysis. Hopefully, this new focus will result in a less biased scientific literature and fairer judgments about the validity of scientific theories.

Author contributions

Linda Dominguez Alvarez designed Study 1 and collected the data under supervision of Jelte Wicherts and Marcel van Assen. Olmo van den Akker designed Study 2, while Jelte Wicherts, Marcel van Assen, and Marjan Bakker provided critical feedback. Olmo van den Akker and Marcel van Assen designed the Heuristic Analyses, while Jelte Wicherts and Marjan Bakker provided critical feedback. Olmo van den Akker analyzed the data and wrote the first draft with help from Marcel van Assen, Jelte Wicherts, and Marjan Bakker. All authors approved the final version of the manuscript for submission.

Open Practices Statement

The data, materials, and code for Study 1, Study 2, and the Heuristic Analyses are available at https://osf.io/2g4wf. The preregistrations of Study 2 and the Heuristic Analyses can also be found there.

References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., … , & Wagenmakers, E. J. (2018). Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation. *Advances in Methods and Practices in Psychological Science, 1*(3), 257-366. doi: 10.1177/2515245918773742

Azevedo, F., Parsons, S., Micheli, L., Strand, J. F., Rinke, E., Guay, S., … FORRT. (2019, December 13). Introducing a Framework for Open and Reproducible Research Training (FORRT). doi: 10.31219/osf.io/bnh7p

Bakker, M., Veldkamp, C. L., van Assen, M. A., Crompvoets, E. A., Ong, H. H., Nosek, B. A., ... & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*(12), e3000937. doi: 10.1371/journal.pbio.3000937

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

Da Silva Frost, A., & Ledgerwood, A. (2020). Calibrate your confidence in research findings: A tutorial on improving research methods and practices. *Journal of Pacific Rim Psychology*, *14*. doi: 10.1017/prp.2020.7

Dang, J., Barker, P., Baumert, A., Bentvelzen, M., Berkman, E., Buchholz, N., ... & Zinkernagel, A. (2021). A multilab replication of the ego depletion effect. *Social Psychological and Personality Science*, *12*(1), 14-24. doi: 10.1177/1948550619887702

Elliott, E. M., Morey, C. C., AuBuchon, A. M., Cowan, N., Jarrold, C., Adams, E. J., ... & Voracek, M. (2021). Multilab direct replication of Flavell, Beach, and Chinsky (1966): Spontaneous verbal rehearsal in a memory task as a function of age. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211018187. doi: 10.1177/25152459211018187

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS one*, *5*(4), e10068. doi: 10.1371/journal.pone.0010068

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

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*(3), 239. doi: 10.1037/0033-295X.90.3.239

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

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*(1), 35-42. doi: 10.1016/j.socec.2010.10.008

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science, 1*(2), 198-218. doi: 10.1177/2515245918771329

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*(6), 562-571. doi: 10.1177/1745691612457576

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

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

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, *88*(2), 359-369. doi: 10.1037/0033-2909.88.2.359

Hoekstra, R., Finch, S., Kiers, H. A., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review*, *13*(6), 1033-1037. doi: 10.3758/BF03213921

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157- 1164. doi: 10.3758/s13423-013-0572-3

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

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237. doi: 10.1037/h0034747

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Sowden, W. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443- 490. doi: 10.1177/2515245918810225

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

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. doi: 10.1037/a0039400.

Murayama, K., Pekrun, R., & Fiedler, K. (2013). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review*, 1088868313496330. doi: 10.1177/1088868313496330

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *The Journal of Psychology*, *55*(1), 33-38. doi: 10.1080/00223980.1963.9916596

Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in interpretation of levels of significance. *Psychological Reports, 15*(2), 570. doi: 10.2466/pr0.1964.15.2.570

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

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

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. *The American Statistician*, *49*(1), 108-112. doi: 10.1080/00031305.1995.10476125

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124-1131. doi: 10.1126/science.185.4157.1124

Van den Akker, O. R. (2021). Preregistration in the Social Sciences: Empirical Evidence of its Effectiveness. Presentation at Metascience 2021. https://www.youtube.com/watch?v=jitIUlmDZS8

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 1832. doi: 10.3389/fpsyg.2016.01832

Wicherts, J. M. (2017). The weak spots in contemporary science (And how to fix them). *Animals*, *7*(12), 90. doi: 10.3390/ani7120090

CHAPTER 6

6

Preregistration of secondary data analysis: A template and tutorial

Olmo R. van den Akker[1], Sara J. Weston[2], Lorne Campbell[3], William J. Chopik[4], Rodica Ioana Damian[5], Pamela E. Davis-Kean[6], Andrew N. Hall[7], Jessica E. Kosie[8], Elliott Kruse[9], Jerome Olsen[10,11], Stuart J. Ritchie[12], K.D. Valentine[13], Anna E. van 't Veer[14], Marjan Bakker[1]

[1]Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands
[2]Department of Psychology, University of Oregon, Eugene, OR, USA
[3]Department of Psychology, University of Western Ontario, London, ON, CAN
[4]Department of Psychology, Michigan State University, East Lansing, MI, USA
[5]Department of Psychology, University of Houston, Houston, TX, USA
[6]Department of Psychology, University of Michigan, Ann Arbor, MI, USA
[7]Department of Psychology, Northwestern University, Evanston, IL, USA
[8]Department of Psychology, Princeton University, Princeton, NJ, USA
[9]EGADE Business School, Tec de Monterrey, Ciudad de México, México
[10]Department of Applied Psychology: Work, Education and Economy, University of Vienna, Vienna, Austria
[11]Max Planck Institute for Research on Collective Goods, Bonn, Germany
[12]Social, Genetic and Developmental Psychiatry Centre, King's College, London, UK
[13]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA
[14]Methodology and Statistics Unit, Institute of Psychology, Leiden University, Leiden, The Netherlands

This paper is published in Meta-Psychology. See https://doi.org/10.15626/MP.2020.2625

Abstract

Preregistration has been lauded as one of the solutions to the so-called 'crisis of confidence' in the social sciences and has therefore gained popularity in recent years. However, the current guidelines for preregistration have been developed primarily for studies where new data will be collected. Yet, preregistering secondary data analyses---where new analyses are proposed for existing data---is just as important, given that researchers' hypotheses and analyses may be biased by their prior knowledge of the data. The need for proper guidance in this area is especially desirable now that data is increasingly shared publicly. In this tutorial, we present a template specifically designed for the preregistration of secondary data analyses and provide comments and a worked example that may help with using the template effectively. Through this illustration, we show that completing such a template is feasible, helps limit researcher degrees of freedom, and may make researchers more deliberate in their data selection and analysis efforts.

Introduction

Preregistration has been lauded as one of the key solutions to the replication crisis in the social sciences, mainly because it has the potential to prevent *p*-hacking by restricting researcher degrees of freedom, but also because it improves transparency and study planning, and can reduce publication bias. However, despite its growing popularity, preregistration is still in its infancy and preregistration practices are far from optimal[12] (Claesen, Gomes, Tuerlinckx, & Vanpaemel, 2019; Veldkamp et al., 2018). Moreover, the current guidelines for preregistration are primarily relevant for studies in which new data will be collected. In this paper, we suggest that preregistration is also attainable when testing new hypotheses with pre-existing data and provide a tutorial on how to effectively preregister such secondary data analyses.

Secondary data analysis involves the analysis of existing data to investigate research questions, often in addition to the main ones for which the data were originally gathered (Grady, Cummings, & Hulley, 2013). Analyzing these datasets comes with its own challenges (Cheng & Phillips, 2014; Smith et al., 2011). For instance, common secondary datasets often include many different variables from many different respondents, sometimes measured at different points in time (e.g., the World Values Survey, Inglehart et al., 2014; the Wisconsin Longitudinal Study, Herd, Carr, & Roan, 2014). This provides ample opportunity for researchers to *p*-hack and increases the likelihood of obtaining spurious statistically significant results (Weston, Ritchie, Rohrer, & Przybylski, 2019).

In addition, because secondary data are often extensive and difficult to collect initially, researchers frequently analyze the same dataset multiple times to answer different research questions. Researchers are therefore not likely to come to a dataset with completely fresh eyes, and may have insight regarding associations between at least some of the variables in the dataset. Such prior knowledge may steer the researchers toward a hypothesis that they already know is in line with the data. This practice is called HARKing (Hypothesizing After Results Are Known; Kerr, 1998) and can lead to false positive results (Rubin, 2017). If HARKing goes undisclosed, it is not possible for third parties to evaluate whether the statistical tests for the hypotheses are well founded, as statistical hypothesis tests (e.g., null hypothesis significance tests, NHST) are only valid when the hypotheses are drawn up a priori (Wagenmakers, Wetzels, Borsboom, Van der Maas, & Kievit, 2012; but see Devezer et al., 2020).

12 The benefits of preregistration can only be reaped fully if preregistration documents are sufficiently detailed and there are no undisclosed discrepancies between the preregistration and the actual study. If this is not the case, preregistration runs the risk of being an empty signal of scientific rigor (Pham & Oh, 2021).

Because secondary data analyses are particularly sensitive to data-driven researcher decisions, preregistering them is especially important. Other options exist to increase error control and illustrate sensitivity to flexibility in data analysis, however. For example, a multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016) or specification-curve analysis (Simonsohn, Simmons, & Nelson, 2015) would be useful if researchers are unsure about which specific analysis is most suitable to test their hypothesis. In these approaches, all *plausible* analytic specifications are implemented to get an overall picture of the evidence without the need to choose a specific (and potentially biased) statistical analysis. This makes it impossible for researchers to cherry-pick variables or analyses based on their prior knowledge. However, it would still be possible to cherry-pick the range of analyses, and it is difficult to weight the results from the different analyses in an unbiased manner. It would thus be appropriate to complement these methods with a preregistration, especially when the aim is to limit the potential for *p*-hacking and HARKing, for both primary and secondary data analysis.

To facilitate the preregistration of secondary data analyses, a session was organized at the Society for the Improvement of Psychological Science (SIPS, see https://improving-psych.org) conference in 2018 with the aim of creating an expert-generated preregistration template specifically tailored to secondary data analysis. Providing guidance on how to preregister is vital as preregistration is hard and requires practice and effort to be effective (Nosek et al., 2019). Participants in the session were experts on or had experience with secondary data analysis, preregistration, or both, thereby providing a good mix of expertise for the task at hand. The session began with analyzing the standard OSF Preregistration template (Bowman et al., 2016) and through successive rounds of discussion and testing, participants decided whether items could be edited, omitted, or added to make the template suitable for secondary data analysis. The resulting first draft of the template was further improved in the months following the conference through a digital back and forth involving the preregistration of an actual secondary data analysis. These efforts---the generation of the template and the preregistration of an example analysis---culminated in the preregistration template presented here.

Specific templates like this can greatly facilitate preregistration as it gives the author guidance about what to include in the preregistration so that all researcher degrees of freedom are covered (Veldkamp et al., 2018). As such, the template would also be well-suited as a framework for a registered report submission that focuses on secondary data. Some of the questions in the preregistration template for secondary data analysis are similar to the questions in more 'traditional' templates; others aim to solve the challenges unique to the preregistration of secondary data analysis, such as the increased need for transparency about the process leading up to the preregistration.

The template presented here is not the only preregistration template for secondary data analysis. Mertens and Krypotos (2019) simultaneously developed a template consisting of 10 questions based on the AsPredicted template (see https://aspredicted.org). Our template differs from that template in two ways. First, it involves 25 questions and therefore captures a wider array of researcher degrees of freedom. For example, our template includes specific questions about defining and handling outliers, and the specification of robustness checks, both of which give leeway for data-driven decisions in secondary data analyses (Weston et al., 2019). Moreover, a more comprehensive template gives researchers the option to use as many or as few of the questions as they want, in order to tailor their preregistration to specific study needs. Second, our template comes with elaborate comments and a worked example that we hope makes the preregistration of secondary data analysis more concrete. We think both these contributions are helpful to researchers looking to preregister their secondary data analysis.

Using the template to preregister a secondary data analysis: Template questions, example answers, and guiding comments

Part 1: Study information

Q1: Provide the working title of your study.
A1: Do religious people follow the golden rule? Assessing the link between religiosity and prosocial behavior using data from the Wisconsin Longitudinal Study.
Comment(s): We specifically mention the data set we are using so that readers know we are preregistering a secondary data analysis. Clarifying this from the outset is helpful because readers may look at such preregistrations differently than they look at preregistrations of primary data analyses.

Q2: Name the authors of this preregistration.
A2:
Josiah Carberry (JC) – ORCID iD: https://orcid.org/0000-0002-1825-0097
Pomona Sprout (PS) – Personal webpage: https://en.wikipedia.org/wiki/Hogwarts_staff#Pomona_Sprout
Comment(s): When listing the authors, add an ORCID iD or a link to a personal webpage so that you and your co-authors can be easily identified. This is particularly important when preregistering secondary data analyses because you may have prior knowledge about the data that may influence the contents of the preregistration. If a reader has access to a personal profile that lists prior research, they can judge whether any prior knowledge of the data is plausible and whether it potentially biased the data analysis.

That is, whether it introduced systematic error in the testing because researchers selected or encouraged one outcome or answer over others (Merriam-Webster, n.d.).

Q3: List each research question included in this study.
A3:
RQ1 = Are more religious people more prosocial than less religious people?
RQ2 = Does the relationship between religiosity and prosociality differ for people with different religious affiliations?
Comment(s): Research questions are often used as a stepping stone for the development of specific and testable hypotheses and can therefore be phrased on a more conceptual level than hypotheses. Note that it is perfectly fine to skip the research questions and only preregister your hypotheses.

Q4: Please provide the hypotheses of your secondary data analysis. Make sure they are specific and testable, and make it clear what your statistical framework is (e.g., Bayesian inference, NHST). In case your hypothesis is directional, do not forget to state the direction. Please also provide a rationale for each hypothesis.
A4: "Do to others as you would have them do to you" (Luke 6:31). This golden rule is taught by all major religions, in one way or another, to promote prosociality (Parliament of the World's Religions, 1993). *Religious prosociality* is the idea that religions facilitate behavior that is beneficial for others at a personal cost (Norenzayan & Shariff, 2008). The encouragement of prosocial behavior by religious teachings appears to be fruitful: a considerable amount of research shows that religion is positively related to prosocial behavior (e.g., Friedrichs, 1960; Koenig, McGue, Krueger, & Bouchard, 2007; Morgan, 1983). For instance, religious people have been found to give more money to, and volunteer more frequently for, charitable causes than their non-religious counterparts (e.g., Grønbjerg & Never, 2004; Lazerwitz, 1962; Pharoah & Tanner, 1997). Also, the more important people viewed their religion, the more likely they were to do volunteer work (Youniss, McLellan, & Yates, 1999). Based on the above we expect that religiosity is associated with prosocial behavior in our sample as well.
To assess this prediction, we will test the following hypotheses using a null hypothesis significance testing framework:
H0(1) = In men and women who graduated from Wisconsin high schools in 1957, there is no association between religiosity and prosociality
H1(1) = In men and women who graduated from Wisconsin high schools in 1957, there is a positive association between religiosity and prosociality
Comment(s): Just like in primary data analysis, a good hypothesis is specific (i.e., it includes a specific population), quantifiable, and testable. A one-sided hypothesis is suitable if theory, previous literature, or (scientific) reasoning indicates that your effect of interest is likely to be in a certain direction (e.g., A < B). Note that we provided de-

tailed information about the theory and previous literature in our answer. This is crucial for secondary data analysis because it allows the reader to assess the thought process behind the hypotheses. Readers can then judge for themselves whether they think the hypotheses logically follow from the theory and previous literature or that they may have been tainted by the authors' prior knowledge of the data. Ideally, your preregistration already contains the framework for the introduction of the final paper. Moreover, writing up the introduction now instead of post hoc forces you to think clearly about the way you arrived at the hypotheses and may uncover flaws in your reasoning that can then be corrected before data collection begins.

Part 2: Data description

Q5: Name and describe the dataset(s), and if applicable, the subset(s) of the data you plan to use. Useful information to include here is the type of data (e.g., cross-sectional or longitudinal), the general content of the questions, and some details about the respondents. In the case of longitudinal data, information about the survey's waves is useful as well.
A5: To answer our research questions we will use a dataset from the Wisconsin Longitudinal Study (WLS; Herd, Carr, & Roan, 2014). The WLS provides long-term data on a random sample of all the men and women who graduated from Wisconsin high schools in 1957. The WLS involves twelve waves of data. Six waves were collected from the original participants or their parents (1957, 1964, 1975, 1992, 2004, and 2011), four were collected from a selected sibling (1977, 1994, 2005, and 2011), one from the spouse of the original participant (2004), and one from the spouse of the selected sibling (2006). The questions vary across waves and are related to domains as diverse as socio-economic background, physical and mental health, and psychological makeup. We will use the subset consisting of the 1957 graduates who completed the follow-up 2003-2005 wave of the WLS dataset because it includes specific modules on religiosity and volunteering.
Comment(s): Like the WLS data we use in our example, many large-scale datasets are outlined in detail in an accompanying paper. It is important to cite papers like this, but also to mention the most relevant information in the preregistration so that readers do not have to search for the information themselves. Sometimes information about the dataset is not readily available. In those cases, be especially candid with the information you have about the dataset because the data you provide may be the only information about the data available to readers of the preregistration.

Q6: Specify the extent to which the dataset is open or publicly available. Make note of any barriers to accessing the data, even if it is publicly available.
A6: The dataset we will use is publicly available, but you need to formally agree to acknowledge the funding source for the Wisconsin Longitudinal Study, to cite the data release in any manuscripts, working papers, or published articles using these data, and

to inform WLS about any published papers for use in the WLS bibliography and for reporting purposes. To do this you need to submit some information about yourself on the website (https://www.ssc.wisc.edu/wlsresearch/data/downloads). You will then receive an email with a download link.
Comment(s): It is important to check whether the data is open or publicly available also to other researchers. For example, it could be that you have access via the organization providing the data (explain this in your answer to Q7), but that does not necessarily mean that it is publicly available to others. An example of publicly available data that is difficult to access would be data for which you need to register a profile on a website, or for which the owners of the data need to accept your request before you can have access.

Q7: How can the data be accessed? Provide a persistent identifier or link if the data are available online, or give a description of how you obtained the dataset.
A7: The data can be accessed by going to the following link and searching for the variables that are specified in Q12 of this preregistration: https://www.ssc.wisc.edu/wlsresearch/documentation/browse/?label=&variable=&wave_108=on&searchButton=Search
Comment(s): When available, report the dataset's persistent identifier (e.g., a DOI) so that the data can always be retrieved from the Internet. In our example, we could only provide a link, but we added instructions for the reader to retrieve the data. In general, try to bring the reader as close to the relevant data as possible, so instead of giving the link to the overarching website, give the link to the part of the website where the data can easily be located.

Q8: Specify the date of download and/or access for each author.
A8:
PS: Downloaded 12 February 2019; Accessed 12 February 2019.
JC: Downloaded 3 January 2019 (estimated); Accessed 12 February 2019.
We will use the data accessed by JC on 12 February 2019 for our statistical analyses.
Comment(s): State here for each author when the dataset was initially downloaded (e.g., for previous analyses or merely to obtain the data) and when either metadata or the actual data (specify which) was first accessed (e.g., to identify variables of interest or to help fill out this form). Also, specify the author whose downloaded data you will use for the statistical analyses. This information is crucial in light of the reproducibility of your study because it is possible that the data has been edited since you last downloaded or accessed it. If you cannot retrieve when you downloaded or accessed the data, estimate those dates. In case you collected the data yourself to answer another research question, please state the date you first looked at the data. Finally, because not everybody will use the same date format it is important to state the date you downloaded or accessed

the data unambiguously. For example, avoid dates like 12/02/2019 and instead use 12 February 2019 or December 2nd, 2019.

Q9: If the data collection procedure is well documented, provide a link to that information. If the data collection procedure is not well documented, describe, to the best of your ability, how data were collected.
A9: The WLS data was and is being collected by the University of Wisconsin Survey Center for use by the research community. The origins of the WLS can be traced back to a state-sponsored questionnaire administered during the spring of 1957 at all Wisconsin high school to students in their final year. Therefore, the dataset constitutes a specific sample not necessarily representative of the United States as a whole. Most panel members were born in 1939, and the sample is broadly representative of white, non-Hispanic American men and women who completed at least a high school education. A flowchart for the data collection can be found here: https://www.ssc.wisc.edu/wlsresearch/about/flowchart/cor459d7.pdf
Comment(s): While describing the data collection procedure, pay specific attention to the representativeness of the sample, and possible biases stemming from the data collection. For example, describe the population that was sampled from, whether the aim was to acquire a representative / regional / convenience sample, whether the data collectors were aware of this aim, the data collectors' recruitment efforts, the procedure for running participants, whether randomization was used, and whether participants were compensated for their time. All of this information can be used to judge whether the sample is representative of a wider population or whether the data is biased in some way, which crucially determines the conclusions that can be drawn from the results. In addition, thinking about the representativeness of a dataset is a crucial part of the planning stage of the research. For example, you might come to the conclusion that the dataset at hand is not suitable after all and opt for a different dataset, thereby preventing research waste. Finally, it is good practice to describe what entity originally collected the data (e.g., your own lab, another lab, a multi-lab collaboration, a (national) survey collection organization, a private organization) because different data sources may have different purposes for collecting the data, which may also result in biased data.

Q10: Some studies offer codebooks to describe their data. If such a codebook is publicly available, link to it here or upload the document. If not, provide other available documentation. Also provide guidance on what parts of the codebook or other documentation are most relevant.
A10: The codebook for the dataset we use can be found here: https://www.ssc.wisc.edu/wlsresearch/documentation/waves/?wave=grad2k. We will mainly use questions from the mail survey about religion and spirituality, and the phone survey on volunteering, but will also use some questions from other modules (see the answer to Q12).

Comment(s): Any documentation is welcome here, as readers will use this documentation to make sense of the dataset. If applicable, provide the codebook for the entire dataset but guide the reader to the relevant parts of the codebook so they do not have to search for the relevant parts extensively. Alternatively, you can create your own data dictionaries/codebooks (Arslan, 2019; Buchanan et al., 2019). If, for some reason codebook information cannot be shared publicly, provide an explanation.

Part 3: Variables

Q11: If you are going to use any manipulated variables, identify them here. Describe the variables and the levels or treatment arms of each variable (note that this is not applicable for observational studies and meta-analyses). If you are collapsing groups across variables this should be explicitly stated, including the relevant formula. If your further analysis is contingent on a manipulation check, describe your decisions rules here.
A11: Not applicable.
Comment(s): Manipulated variables in secondary datasets usually originate from another study investigating another research question. You may, therefore, need to adapt the manipulated variable to answer your own research question. For example, it may be necessary to relabel or even omit one of the treatment arms. Please provide a careful log of all these adaptations so that readers will have a clear grasp of the variable you will be using and how it differs from the variable in the original dataset. Any resources mentioned in the answer to Q10 may be useful here as well.

Q12: If you are going to use measured variables, identify them here. Describe both outcome measures as well as predictors and covariates and label them accordingly. If you are using a scale or an index, state the construct the scale/index represents, which items the scale/index will consist of, how these items will be aggregated, and whether this aggregation is based on a recommendation from the study codebook or validation research. When the aggregation of the items is based on exploratory factor analysis (EFA) or confirmatory factor analysis (CFA), also specify the relevant details (EFA: rotation, how the number of factors will be determined, how best fit will be selected, CFA: how loadings will be specified, how fit will be assessed, which residuals variance terms will be correlated). If you are using any categorical variables, state how you will code them in the statistical analyses.
A12:
Religiosity (IV): Religiosity is measured using a newly created scale with a subset of items from the Religion and Spirituality module of the 2004 mail survey (described here: https://www.ssc.wisc.edu/wlsresearch/documentation/waves/?wave=grad2k&module=gmail_religion). The scale includes general questions about how religious/spiritual the individual is and how important religion/spirituality

is to them. Importantly, the questions are not specific to a particular denomination and are on the same response scale. The specific variables are as follows:

1. il001rer: How religious are you?
2. il002rer: How spiritual are you?
3. il003rer: How important is religion in your life?
4. il004rer: How important is spirituality in your life?
5. il005rer: How important was it, or would it have been if you had children, to send your children for religious or spiritual instruction?
6. il006rer: How closely do you identify with being a member of a religious group?
7. il007rer: How important is it for you to be with other people who are the same religion as you?
8. il008rer: How important do you think it is for people of your religion to marry other people who are the same religion?
9. il009rer: How strongly do you believe that one should stick to a particular faith?
10. il010rer: How important was religion in your home when you were growing up?
11. il011rer: When you have important decisions to make in your life, how much do you rely on your religious or spiritual beliefs?
12. il012rer: How much would your spiritual or religious beliefs influence your medical decisions if you were to become gravely ill?

The levels of all of these variables are indicated by a Likert scale with the following options: (1) Not at all; (2) Not very; (3) Somewhat; (4) Very; (5) Extremely, as well as 'System Missing' (the participant did not provide an answer) and 'Refused' (the participant refused to answer the question). Variables il006rer, il008rer, and il012rer additionally include the option 'Don't know' (the participant stated that they did not know how to answer the question). We will use the average score (after omitting non-numeric and 'Don't know' responses) on the twelve variables as a measure of religiosity. This average score is constructed by ourselves and was not already part of the dataset.

Prosociality (DV): In line with previous research (Konrath, Fuhrel-Forbis, Lou, & Brown, 2012), we will use three measures of prosociality that measure three aspects of engagement in other-oriented activities (see Brookfield, Parry, & Bolton, 2018 for the link between prosociality and volunteering). The prosociality variables come from the Volunteering module of the 2004 phone survey. The codebook of that module can be found here: https://www.ssc.wisc.edu/wlsresearch/documentation/waves/?wave=grad2k&module=gvol). The three measures of prosociality we will use are:

1. gv103re: Did the graduate do volunteer work in the last 12 months?
    a. This dichotomous variable assesses whether or not the participant has engaged in any volunteering activities in the last 12 months. The levels of this variable are yes/no. Yes will be coded as '1', no will be coded as '0'.

2. gv109re: Number of graduate's other volunteer activities in the past 12 months.
    a. This variable is a summary index providing a quantitative measure of the participant's volunteering activities. Scores on this variable range from 1 to 5 and reflect the number of the previous five questions to which the participant answered YES. The previous five questions assess whether or not the participant volunteered at any of the following organization types: (1) religious organizations; (2) school or educational organization; (3) political group or labor union; (4) senior citizen group or related organization; (5) other national or local organizations. For each of these questions the answer 'yes' is coded as 1 and the answer 'no' is coded as 0.
3. gv111re: How many hours did the graduate volunteer during a typical month in the last 12 months?
    a. This is a numerical variable that provides information on how many hours per month, on average, the participant volunteered.

The three variables will be treated as separate measures in the dataset and do not require manual aggregation.

Number of Siblings (Covariate):
We will include the participant's number of siblings as a control variable because many religious families are large (Pew Research Center, 2015) and it can be argued that cooperation and trust arise more naturally in larger families because of the larger number of social interactions in those families. To measure participants' number of siblings we used the variable gk067ss: The total number of siblings ever born from the 2004 phone survey Siblings module (see https://www.ssc.wisc.edu/wlsresearch/documentation/waves/?wave=grad2k&module=gsib). This is a numerical variable with the possibility for the participant to state "I don't know". At the interview participants were instructed to include "siblings born alive but no longer living, as well as those alive now and to include step-brothers and step-sisters and children adopted by their parents."

Agreeableness (Covariate):
We will include the summary score for agreeableness (ih009rec, see https://www.ssc.wisc.edu/wlsresearch/documentation/waves/?wave=grad2k&module=gmail_values) in the analysis as a control variable because a previous study (on the same dataset, see the answer to Q18) we were involved in showed a positive association between agreeableness and prosociality. Because previous research also indicates a positive association between agreeableness and religiosity (Saroglou, 2002) we need to include agreeableness as a control variable to disentangle the influence of religiosity on prosociality and the influence of agreeableness on prosociality. The variable ih009rec is a sum score of the variables ih003rer-ih008rer (To what extent do you agree that you see

yourself as someone who is talkative / is reserved [reverse coded] / is full of energy / tends to be quiet [reverse coded] / who is sometimes shy or inhibited [reverse coded] / who generates a lot of enthusiasm). All of these were scored from 1 to 6 (1 = "agree strongly", 2 = "agree moderately", 3 = "agree slightly", 4 = "disagree slightly", 5 = "disagree moderately", 6 = "disagree strongly"), while participants could also refuse to answer the question. If a participant refused to answer one of the questions, that participant's score was not included in the sum score variable ih009rec.
Comment(s): If you are using measured variables, describe them in such a way that readers know exactly what variables will be used in the statistical analyses. Because secondary datasets often involve many measured variables, there is ample room to select variables after doing an analysis. It is therefore essential to be exhaustive here. Variables you do not mention here should not pop up in your analysis later unless you have a good reason for it. As you can see, we clearly label the function of each variable, the specific items related to that variable, and the item's response options. It could be that you choose to combine items into an index or scale that have not been combined like that in previous studies. Carefully detail this process and indicate that you constructed the index or scale yourself to avoid confusion. Finally, note that we include covariates to be able to make statements about the causal effect of religion on prosociality. This is common practice in the social sciences, but causal inference is complex and there may be better solutions in other situations, and even in this situation. Please see Rohrer (2018) for more information about causation in observational data.

Q13: Which units of analysis (respondents, cases, etc.) will be included or excluded in your study? Taking these inclusion/exclusion criteria into account, indicate the (expected) sample size of the data you'll be using for your statistical analyses to the best of your knowledge. In the next few questions, you will be asked to refine this sample size estimation based on your judgments about missing data and outliers.
A13: Initially, the WLS consisted of 10,317 participants. As we are not interested in a specific group of Wisconsin people, we will not exclude any participants from our analyses. However, only 7,265 participants filled out the questions on prosociality and the number of siblings in the phone survey and only 6,845 filled out the religiosity items in the mail survey (Herd et al., 2014). This corresponds to a response rate of 73% and 69% respectively. Because we do not know whether the participants that did the mail survey also did the phone survey, our minimum expected sample size is 10,317 * 0.73 * 0.69 = 5,297.
Comment(s): Provide information on the total sample size of the dataset, the sample size(s) of the wave(s) you are going to use (if applicable), as well as the number of participants that provided data on each of the questions and/or scales to be used in the data analyses. In our sample we do not exclude any participants, but if you have a research question about a certain group you may need to exclude participants based

on one or more characteristics. Be very specific when describing these characteristics so that readers with no knowledge of the data are able to redo your moves easily.

For our WLS dataset, it is impossible to know the exact sample size without inspecting the data. If that is the case, provide an estimate of the sample size. If you provide an estimate, try to be conservative and pick the lowest sample size of the possible options. If it is impossible to provide an estimate, it is also possible to mask the data. For example, it is possible to add random noise to all values of the dependent variable. In that case, it is impossible to pick up any real effects and you are essentially blind to the data. Similarly, it is possible to blind yourself to real effects in the data by having someone relabel the treatment levels so you cannot link them to the treatment levels anymore. These and other methods of data blinding are clearly described by Dutilh, Sarafoglou, and Wagenmakers (2019).

Q14: What do you know about missing data in the dataset (i.e., overall missingness rate, information about differential dropout)? How will you deal with incomplete or missing data? Based on this information, provide a new expected sample size.
A14: The WLS provides a documented set of missing codes. In Table 1 below you can find missingness information for every variable we will include in the statistical analyses. 'System missing' refers to the number of participants that did not or could not complete the questionnaire. 'Partial interview' refers to the number of participants that did not get that particular question because they were only partially interviewed. The rest of the codes are self-explanatory.
Importantly, some respondents refused to answer the religiosity questions. These respondents apparently felt strongly about these questions, which could indicate that they are either very religious or very anti-religious. If that is the case, the respondent's propensity to respond is directly associated with their level of religiosity and that the data is missing not at random (MNAR). Because it is not possible to test the stringent assumptions of the modern techniques for handling MNAR data we will resort to simple listwise deletion. It must be noted that this may bias our data as we may lose respondents who are very religious or anti-religious. However, we believe this bias to be relatively harmless given that our sample still includes many respondents that provided extreme responses to the items about the importance of the different facets of religion (see https://www.ssc.wisc.edu/wlsresearch/documentation/waves/?wave=grad2k&module=gmail_religion). Moreover, because our initial sample size is very large, statistical power is not substantially compromised by omitting these respondents. That being said, we will extensively discuss any potential biases resulting from missing data in the limitations section of our paper.

Table 1

An overview of the missing values for all variables we will use in our analyses

Variable	System missing	Don't know	Inappropriate	Refused	Not ascertained	Partial interview	Could not code	**Remaining**	**Remaining (%)**
il001rer	3,471	0	0	190	0	0	0	**6,656**	**64**
il002rer	3,471	0	0	212	0	0	0	**6,634**	**64**
il003rer	3,471	0	0	191	0	0	0	**6,655**	**65**
il004rer	3,471	0	0	241	0	0	0	**6,605**	**64**
il005rer	3,471	0	0	201	0	0	0	**6,645**	**64**
il006rer	3,471	1	0	201	0	0	0	**6,644**	**64**
il007rer	3,471	0	0	192	0	0	0	**6,654**	**65**
il008rer	3,471	1	0	199	0	0	0	**6,646**	**64**
il009rer	3,471	0	0	219	0	0	0	**6,627**	**64**
il010rer	3,471	0	0	190	0	0	0	**6,656**	**65**
il011rer	3,471	0	0	190	0	0	0	**6,656**	**65**
il012rer	3,471	1	0	198	0	0	0	**6,647**	**64**
gv103re	3,052	0	3,955	1	0	182	0	**3,127**	**30**
gv109re	3,052	0	4,590	0	0	182	0	**2,493**	**24**
gv111re	3,052	50	4,716	0	0	182	0	**2,317**	**23**
gk067ss	3,052	21	0	0	0	0	0	**7,244**	**70**

Employing listwise deletion leads to an expected minimum number of 10,317 * 0.30 * 0.70 * 0.64 = 1,387 participants for the binary logistic regression, and an expected minimum number of 10,317 * 0.24 * 0.70 * 0.64 = 1,109 (gv109re) and 10,317 * 0.23 * 0.70 * 0.64 = 1,063 (gv111re) for the linear regressions.

Comment(s): Provide descriptive information, if available, on the amount of missing data for each variable you will use in the statistical analyses and discuss potential issues with the pattern of missing data for your planned analyses. Also provide a plan for how the analyses will take into account the presence of missing data. Where appropriate, provide specific details how this plan will be implemented. This can be done by specifying a step-by-step protocol for how you will impute any missing data. You could first explain how you will assess whether the data are missing at random (MAR) missing completely at random (MCAR) or missing not at random (MNAR), and then state that you will use technique X in case of MAR data, technique Y in case of MCAR data, and technique Z in case of MNAR data. For an overview of the types of missing data, and the different techniques to handle missing data, see Lang & Little (2018). Note that the missing data technique we used in our example, listwise deletion, is usually not the best

way to handle missing data. We decided to use it in this example because it gave us the opportunity to illustrate how researchers can describe potential biases arising from their analysis methods in a preregistration.

If you cannot specify the exact number of missing data because the dataset does not provide that information, provide an estimate. If you provide an estimate, try to be conservative and pick the lowest sample size of the possible options. If it is impossible to provide an estimate, you could also mask the data (see Dutilh, Sarafoglou, & Wagenmakers, 2019). It is good practice to state all missingness information with relation to the total sample size of the dataset.

Q15: If you plan to remove outliers, how will you define what a statistical outlier is in your data? Please also provide a new expected sample size. Note that this will be the definitive expected sample size for your study and you will use this number to do any power analyses.

A15: The dataset probably does not involve any invalid data since the dataset has been previously 'cleaned' by the WLS data controllers and any clearly unreasonably low or high values have been removed from the dataset. However, to be sure we will create a box and whisker plot for all continuous variables (the dependent variables gv109re and gv111re, the covariate gk067ss, and the scale for religiosity) and remove any data point that appears to be more than 1.5 times the IQR away from the 25th and 75th percentile. Based on normally distributed data, we expect that 2.1% of the data points will be removed this way, leaving 1,358 out of 1,387 participants for the binary regression with gv103re as the outcome variable and 1,086 out of 1,109 participants, and 1,041 out of 1,063 participants for the linear regressions with gv109re and gv111re as the outcome variables, respectively.

Comment(s): Estimate the number of outliers you expect for each variable and calculate the expected sample size of your analysis. The expected sample size is required to do a power analysis for the planned statistical tests (Q21) but also prevents you from discarding a significant portion of the data during or after the statistical analysis. If it is impossible to provide such an estimate, you can mask the data and make a more informed estimation based on these masked data (see Dutilh, Sarafoglou, & Wagenmakers, 2019). If you expect to remove many outliers or if you are unsure about your outlier handling strategy, it is good practice to preregister analyses including and excluding outliers. To see how decisions about outliers can influence the results of a study, see Bakker and Wicherts (2014) and Lonsdorf et al. (2019). For more information about outliers in the context of preregistration, see Leys, Delacre, Mora, Lakens, and Ley (2019).

Q16: Are there sampling weights available with this dataset? If so, are you using them or are you using your own sampling weights?

A16: The WLS dataset does not include sampling weights and we will not use our own sampling weights as we do not seek to make any claims that are generalizable to the national population.

Comment(s): Because secondary data samples may not be entirely representative of the population you are interested in, it can be useful to incorporate sampling weights into your analysis. You should state here whether (and why) you will use sampling weights, and provide specifics on exactly how you will use them. To implement sampling weights into your analyses, we recommend using the "survey" package in R (Lumley, 2004).

Part 4: Knowledge of data

Q17: List the publications, working papers (in preparation, unpublished, preprints), and conference presentations (talks, posters) you have worked on that are based on the dataset you will use. For each work, list the variables you analyzed, but limit yourself to variables that are relevant to the proposed analysis. If the dataset is longitudinal, also state which wave of the dataset you analyzed.

Importantly, some of your team members may have used this dataset, and others may not have. It is therefore important to specify the previous works for every co-author separately. Also mention relevant work on this dataset by researchers you are affiliated with as their knowledge of the data may have been spilled over to you. When the provider of the data also has an overview of all the work that has been done using the dataset, link to that overview.

A17: Both authors (PS and JC) have previously used the Graduates 2003-2005 wave to assess the link between Big Five personality traits and prosociality. The variables we used to measure the Big Five personality traits were ih001rei (extraversion), ih009rei (agreeableness), ih017rei (conscientiousness), ih025rei (neuroticism), and ih032rei (openness). The variables we used to measure prosociality were ih013rer ("To what extent do you agree that you see yourself as someone who is generally trusting?"), ih015rer ("To what extent do you agree that you see yourself as someone who is considerate to almost everyone?"), and ih016rer ("To what extent do you agree that you see yourself as someone who likes to cooperate with others?). We presented the results at the ARP conference in St. Louis in 2013 and we are currently finalizing a manuscript based on these results.

Additionally, a senior graduate student in JC's lab used the Graduates 2011 wave for exploratory analyses on depression. She linked depression to alcohol use and general

health indicators. She did not look at variables related to religiosity or prosociality. Her results have not yet been submitted anywhere.

An overview of all publications based on the WLS data can be found here: https://www.ssc.wisc.edu/wlsresearch/publications/pubs.php?topic=ALL.

Comment(s): It is important to specify the different ways you have previously used the data because this information helps you to establish any knowledge of the data you may already have. This prior knowledge will need to be provided in Q18. If available, include persistent identifiers (e.g. a DOI) to any relevant papers and presentations.

Understandably, there is a subjectivity involved in determining what constitutes "relevant" work or "relevant" variables for the proposed analysis. We advise researchers to use their professional judgment and when in doubt always mention the work or variable so readers can assess their relevance themselves. In the worked example, the exploratory analysis by the student in JC's lab is probably not relevant, but because of the close affiliation of the student to JC, it is good to include it anyway.

Listing previous works based on the data also helps to prevent a common practice identified by the American Psychological Association (2019) as unethical: the so-called "least publishable unit" practice (also known as "salami-slicing"), in which researchers publish multiple papers on closely related variables from the same dataset. Given that secondary datasets often involve many closely related variables, this is a particularly pernicious issue here.

Q18: What prior knowledge do you have about the dataset that may be relevant for the proposed analysis? Your prior knowledge could stem from working with the data first-hand, from reading previously published research, or from codebooks. Also provide any relevant knowledge of subsets of the data you will *not* be using. Provide prior knowledge for every author separately.

A18: In a previous study (mentioned in Q17) we used three prosociality variables (ih013rer, ih015rer, and ih016rer) that may be related to the prosociality variables we use in this study. We found that ih013rer, ih015rer, and ih016rer are positively associated with agreeableness (ih009rec). Because previous research (on other datasets) shows a positive association between agreeableness and religiosity (Saroglou, 2002) agreeableness may act as a confounding variable. To account for this we will include agreeableness in our analysis as a control variable. We did not find any associations between prosociality and the other Big Five variables.

Comment(s):

It is important to denote your prior knowledge diligently because it provides information about possible biases in your statistical analysis decisions. For example, you may have learned at an academic conference or in a footnote of another paper that the correlation between two variables is high in this dataset. If you do a test of this hypothesis, you already know the test result, making the interpretation of the test invalid (Wagenmakers, et al., 2012). In cases like this, where you have *direct* knowledge about a hypothesized association, you should disregard doing a confirmatory analysis altogether or do one based on a different dataset.

Any *indirect* knowledge about the hypothesized association does not preclude a confirmatory analysis but should be transparently reported in this section. In our example, we mentioned that we know about the positive association between agreeableness and prosociality, which may say something about the direction of our hypothesized association given the association between agreeableness and religiosity. Moreover, this prior knowledge urged us to add agreeableness as a control variable. Thus, aside from improving your preregistration, evaluating your prior knowledge of the data can also improve the analyses themselves.

All information like this that may influence the hypothesized association is relevant here. For example, restriction of range (Meade, 2010), measurement reliability (Silver, 2008), and the number of response options (Gradstein, 1986) have been shown to influence the association between two variables. You may have provided univariate information regarding these aspects in previous questions. In this section, you can write about how they may affect your hypothesized association.

Do note that it is unlikely that you are able to account for all the effects of prior knowledge on your analytical decisions. For example, you may have prior knowledge that you are not consciously aware of. The best way to capture this unconscious prior knowledge is to revisit previous work, think deeply about any information that might be relevant for the current project, and present it here to the best of your ability. This exercise helps you reflect on potential biases you may have and makes it possible for readers of the preregistration to assess whether the prior knowledge you mentioned is plausible given the list of prior work you provided in Q17.

Of course, it is still possible that researchers purposefully neglect to mention prior knowledge or provide false information in a preregistration. Even though we believe that deliberate deceit like this is rare, at the end of our template we require researchers to formally "promise" to have truthfully filled out the template and that no other prereg-

istration exists on the same hypotheses and data. A violation of this formal statement can be seen as misconduct, and we believe researchers are unlikely to cross that line.

Part 5: Analyses

Q19: For each hypothesis, describe the statistical model you will use to test the hypothesis. Include the type of model (e.g., ANOVA, multiple regression, SEM) and the specification of the model. Specify any interactions and post-hoc analyses and remember that any test not included here must be labeled as an exploratory test in the final paper.

A19: Our first hypothesis will be tested using three analyses since we use three variables to measure prosociality. For each, we will run a directional null hypothesis significance test to see whether a positive effect exists of religiosity on prosociality. For the first outcome (gv103re: Did the graduate do volunteer work in the last 12 months?) we will run a logistic regression with religiosity, the number of siblings, and agreeableness as predictors.

For the second and third outcomes (gv109re: Number of graduate's other volunteer activities in the past 12 months; gv111re: How many hours did the graduate volunteer during a typical month in the last 12 months?) we will run two separate linear regressions with religiosity, the number of siblings, and agreeableness as predictors.

The code we will use for all these analyses can be found at https://osf.io/e3htr.

Comment(s): Think carefully about the variety of statistical methods that are available for testing each of your hypotheses. One of the classic "Questionable Research Practices" is trying multiple methods and only publishing the ones that "work" (i.e., that support your hypothesis). Almost every method has several options that may be more or less suited to the question you are asking. Therefore, it is crucial to specify *a priori* which one you are going to use and how.

If you can, include the code you will use to run your statistical analyses, as this forces you to think about your analyses in detail and makes it easy for readers to see exactly what you plan to do. Ideally, when you have loaded the data in a software program you only have to press one button to run your analyses. If including the code is impossible, describe the analyses such that you could give a positive answer to the question: "Would a colleague who is not involved in this project be able to recreate this statistical analysis?"

Q20: If applicable, specify a predicted effect size or a minimum effect size of interest for all the effects tested in your statistical analyses.

A20: For the logistic regression with 'Did the graduate do volunteer work in the last 12 months?' as the outcome variable, our minimum effect size of interest is an odds of 1.05. This means that a one-unit increase on the religiosity scale would be associated with a 1.05 factor change in odds of having done volunteering work in the last 12 months versus not having done so.

For the linear regressions with 'The number of graduate's volunteer activities in the last 12 months", and "How many hours did the graduate volunteer during a typical month in the last 12 months?' as the outcome variables, the minimum regression coefficients of interest of the religiosity variables are 0.05 and 0.5, respectively. This means that a one-unit increase in the religiosity scale would be associated with 0.05 extra volunteering activities in the last 12 months and with 0.5 more hours of volunteering work in the last 12 months. All of these smallest effect sizes of interest are based on our own intuition.

To make comparisons possible between the effects in our study and similar effects in other studies the unstandardized linear regression coefficients will be transformed into standardized regression coefficients using the following formula: $\beta_i = B_i(s_i/s_y)$, where β_i is the unstandardized regression coefficient of independent variable i, and s_i and s_y are the standard deviations of the independent and dependent variable respectively.

Comment(s): A predicted effect size is ideally based on a representative preliminary study or meta-analytical result. If those are not available, it is also possible to use your own intuition. For advice on setting a minimum effect size of interest, see Lakens, Scheel, & Isager (2018) and Funder and Ozer (2019).

Q21: Present the statistical power available to detect the predicted effect size(s) or the smallest effect size(s) of interest, OR present the accuracy that will be obtained for estimation. Use the sample size after updating for missing data and outliers, and justify the assumptions and parameters used (e.g., give an explanation of why anything smaller than the smallest effect size of interest would be theoretically or practically unimportant).

A21: The sample size after updating for missing data and outliers is 1,358 for the logistic regression with gv103re as the outcome variable, and 1,086 and 1,041 for the linear regressions with gv109re and gv111re as the outcome variables, respectively. For all three analyses this corresponds to a statistical power of approximately 1.00 when assuming our minimum effect sizes of interest. For the linear regressions we additionally assumed the variance explained by the predictor to be 0.2 and the residual variance to be 1.0 (see figure below for the full power analysis of the regression with the lowest sample size). For the logistic regression we assumed an intercept of -1.56 corresponding

to a situation where half of the participants have done volunteer work in the last year (see the R-code for the full power analysis at https://osf.io/f96rn).

Comment(s): Advice on conducting a power analysis using G*Power can be found in Faul, Erdfelder, Buchner, and Lang (2009). Advice on conducting a power analysis using R can be found here: https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html. Note that power analyses for secondary data analyses are unlike power analyses for primary data analyses because we already have a good idea about what our sample size is based on our answers to Q13, Q14, and Q15. Therefore, we are primarily interested in finding out what effect sizes we are able to find for a given power level or what our power is given our minimum effect size of interest. In our example, we chose the second option. When presenting your power analysis be sure to state the version of G*Power, R, or any other tool you calculated power with, including any packages or add-ons, and also report or copy all the input and results of the power analysis.

Q22: What criteria will you use to make inferences? Describe the information you will use (e.g. specify the *p*-values, effect sizes, confidence intervals, Bayes factors, specific model fit indices), as well as cut-off criteria, where appropriate. Will you be using one- or two-tailed tests for each of your analyses? If you are comparing multiple conditions or testing multiple hypotheses, will you account for this, and if so, how?

A22: We will make inferences about the association between religiosity and prosociality based on the *p*-values and the size of the regression coefficients of the religiosity variable in the three main regressions. We will conclude that a regression analysis supports our hypothesis if both the *p*-value is smaller than .01 *and* the regression coefficient is larger than our minimum effect size of interest. We chose an alpha of .01 to account for the fact that we do a test for each of the three regressions (0.05/3, rounded down). If the conditions above hold for all three regressions, we will conclude that our hypothesis is fully supported, if they hold for one or two of the regressions we will conclude that our hypothesis is partially supported, and if they hold for none of the regressions we will conclude that our hypothesis is not supported.

Comment(s): It is crucial to specify your inference criteria before running a statistical analysis because researchers have a tendency to move the goalposts when making inferences. For example, almost 40% of *p*-values between 0.05 and 0.10 are reported as "marginally significant", even though these values are not significant when compared to the traditional alpha level of 0.05, and the evidential value of these *p*-values is low (Olsson-Collentine, Van Assen, & Hartgerink, 2019). Similarly, several studies have found that the majority of studies reporting *p*-values do not use any correction for multiple comparisons (Cristea & Ioannidis, 2018; Wason, Stecher, & Mander, 2014), perhaps be-

cause this lowers the chance of finding a statistically significant result. For an overview of multiple-comparison correction methods relevant to secondary data analysis, see Thompson, Wright, Bissett, and Poldrack (2019).

Q23: What will you do should your data violate assumptions, your model not converge, or some other analytic problem arises?

A23: When the distribution of the number of volunteering hours (gv111re) is significantly non-normal according to the Kolmogorov-Smirnov test (Massey, 1951), and/or (b) the linearity assumption is violated (i.e., the points are asymmetrically distributed around the diagonal line when plotting observed versus the predicted values), we will log-transform the variable.

Comment(s): It is, of course, impossible to predict every single way that things might go awry during the analysis. One of the variables may have a strange and unexpected distribution, one of the models may not converge because of a quirk of the correlational structure, and you may even encounter error messages that you have never seen before. You can use your prior knowledge of the dataset to set up a decision tree specifying possible problems that might arise and how you will address them in the analyses. Thinking through such a decision tree will make you less overwhelmed when something does end up going differently than expected.

However, note that decision trees come with their own problems and can quickly become very complex. Alternatively, you might choose to select analysis methods that make assumptions that are as conservative as possible; preregister robustness analyses which test the robustness of your findings to analysis strategies that make different assumptions; and/or pre-specify a single primary analysis strategy, but note that you will also report an exploratory investigation of the validity of distributional assumptions (Williams & Albers, 2019). Of course, there are pros and cons to all methods of dealing with violations, and you should choose a technique that is most appropriate for your study.

Q24: Provide a series of decisions about evaluating the strength, reliability, or robustness of your focal hypothesis test. This may include within-study replication attempts, additional covariates, cross-validation efforts (out-of-sample replication, split/hold-out sample), applying weights, selectively applying constraints in an SEM context (e.g., comparing model fit statistics), overfitting adjustment techniques used (e.g., regularization approaches such as ridge regression), or some other simulation/sampling/bootstrapping method.

A24: To assess the sensitivity of our results to our selection criterion for outliers, we will run an additional analysis without removing any outliers.

Comment(s): There are many methods you can use to test the limits of your hypothesis. The options mentioned in the question are not supposed to be exhaustive or prescriptive. We included these examples to encourage researchers to think about these methods, all of which serve the same purpose as preregistration: improving the robustness and replicability of the results.

Q25: If you plan to explore your dataset to look for unexpected differences or relationships, describe those tests here, or add them to the final paper under a heading that clearly differentiates this exploratory part of your study from the confirmatory part.

A25: As an exploratory analysis, we will test the relationship between scores on the religiosity scale and prosociality after adjusting for a variety of social, educational, and cognitive covariates that are available in the dataset. We have no specific hypotheses about which covariates will attenuate the religiosity-prosociality relation most substantially, but we will use this exploratory analysis to generate hypotheses to test in other, independent datasets.

Comment(s): Whereas it is not presently the norm to preregister exploratory analyses, it is often good to be clear about which variables will be explored (if any), for example, to differentiate these from the variables for which you have specific predictions or to plan ahead about how to compute these variables.

Part 6: Statement of integrity

The authors of this preregistration state that they filled out this preregistration to the best of their knowledge and that no other preregistration exists pertaining to the same hypotheses and dataset.

Summary

In this tutorial we presented a preregistration template for the analysis of secondary data and have provided guidance for its effective use. We are aware that the number of questions (25) in the template may be overwhelming but it is important to note that not every question is relevant for every preregistration. Our aim was to be inclusive and cover all bases in light of the diversity of secondary data analyses. Even though none of the questions are mandatory, we do believe that an elaborate preregistration is preferable over a concise preregistration simply because it restricts more researcher degrees of freedom. We therefore recommend that authors answer as many questions in as much detail as possible. And, if questions are not applicable, it would be good practice to also specify why this is the case so that readers can assess your reasoning.

Effectively preregistering a study *is* challenging and can take a lot of time but, like Nosek et al. (2019) and many others, we believe it can improve the interpretability, verifiability and rigor of your studies and is therefore more than worth it if you want both yourself and others to have more confidence in your research findings.

The current template is merely one building block toward a more effective preregistration infrastructure and, given the ongoing developments in this area, will be a work in progress for the foreseeable future. Any feedback is therefore greatly appreciated. Please send any feedback to the corresponding author, Olmo van den Akker (ovdakker@gmail.com).

Acknowledgments

The authors would like to thank the participants of the Society for the Improvement of Psychological Science (SIPS) conference in 2018 that helped to create the first draft of the preregistration template but were unable to help out with the subsequent extensions (Brian Brown, Oliver Clark, Charles Ebersole, and Courtney Soderberg).

This research is based on data from the Wisconsin Longitudinal Study, funded by the National Institute on Aging (R01 AG009775; R01 AG033285).

References

American Psychological Association. (2019). Publication Practices & Responsible Authorship. Retrieved from https://www.apa.org/research/responsible/publication

Arslan, R. C. (2019). How to Automatically Document Data With the codebook Package to Facilitate Data Reuse. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245919838783

Bakker, M., & Wicherts, J. M. (2014). 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. https://doi.org/10.1037/met0000014

Bowman, S. D., DeHaven, A. C., Errington, T. M., Hardwicke, T. E., Mellor, D. T., Nosek, B. A., & Soderberg, C. K. (2016, January 1). OSF Prereg Template. https://doi.org/10.31222/osf.io/epgjd

Brookfield, K., Parry, J., & Bolton, V. (2018). Getting the measure of prosocial behaviors: A comparison of participation and volunteering data in the national child development study and the linked social participation and identity study. *Nonprofit and Voluntary Sector Quarterly*, *47*(5), 1081-1101. https://doi.org/10.1177/0899764018786470

Buchanan, E. M., Crain, S. E., Cunningham, A. L., Johnson, H. R., Stash, H. E., Papadatou-Pastou, M., … , & Aczel, B. (2019, May 20). Getting Started Creating Data Dictionaries: How to Create a Shareable Dataset. https://doi.org/10.31219/osf.io/vd4y3

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ..., & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637-644. https://doi.org/10.1038/s41562-018-0399-z

Cheng, H. G., & Phillips, M. R. (2014). Secondary analysis of existing data: opportunities and implementation. *Shanghai Archives of Psychiatry*, *26*(6), 371-375. https://doi.org/10.11919/j.issn.1002-0829.214171

Claesen, A., Gomes, S. L. B. T., Tuerlinckx, F., & Vanpaemel, W. (2019, May 9). Preregistration: Comparing Dream to Reality. https://doi.org/10.31234/osf.io/d8wex

Cristea, I. A., & Ioannidis, J. P. (2018). P values in display items are ubiquitous and almost invariably significant: A survey of top science journals. *PloS one*, *13*(5), e0197440. https://doi.org/10.1371/journal.pone.0197440

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020). The case for formal methodology in scientific reform. BioRxiv, 2020.04.26.048306. https://doi.org/10.1101/2020.04.26.048306

Dutilh, G., Sarafoglou, A., & Wagenmakers, E. J. (2019). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, 1-28. https://doi.org/10.1007/s11229-019-02456-7

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

Friedrichs, R. W. (1960). Alter versus ego: An exploratory assessment of altruism. *American Sociological Review*, 496-508. http://doi.org/10.2307/2092934

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156-168. https://doi.org/10.1177/2515245919847202

Gradstein, M. (1986). Maximal correlation between normal and dichotomous variables. *Journal of Educational Statistics*, *11*(4), 259-261.

Grady, D. G., Cummings, S. R., & Hulley, S. B. (2013). Research using existing data. Designing clinical research, 192-204. Retrieved from https://pdfs.semanticscholar.org/343e/04f26f768c9530f58e1847aff6a4e072d0be.pdf

Grønbjerg, K. A., & Never, B. (2004). The role of religious networks and other factors in types of volunteer work. *Nonprofit Management and Leadership*, *14*(3), 263-289. https://doi.org/10.1002/nml.34

Herd, P., Carr, D., & Roan, C. (2014). Cohort profile: Wisconsin longitudinal study (WLS). *International Journal of Epidemiology*, *43*(1), 34-41. https://doi.org/10.1093/ije/dys194

Inglehart, R., C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, & B. Puranen et al. (eds.). (2014). World Values Survey: Round Six - Country-Pooled Datafile Version: www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. JD Systems Institute.

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

Koenig, L. B., McGue, M., Krueger, R. F., & Bouchard, T. J. Jr. (2007). Religiousness, antisocial behavior, and altruism: Genetic and environmental mediation. *Journal of Personality*, *75*(2), 265-290. https://doi.org/10.1111/j.1467-6494.2007.00439.x

Konrath, S., Fuhrel-Forbis, A., Lou, A., & Brown, S. (2012). Motives for volunteering are associated with mortality risk in older adults. *Health Psychology*, *31*(1), 87-96.
http://doi.org/10.1037/a0025226

Lang K. M. & Little T. D. (2018). Principled Missing Data Treatments. *Prevention Science*, *19*(3), 284-294. https://doi.org/10.1007/s11121-016-0644-5

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259-269. https://doi.org/10.1177/2515245918770963

Lazerwitz, B. (1962). Membership in voluntary associations and frequency of church attendance. *Journal for the Scientific Study of Religion*, *2*(1), 74-84. http://doi.org/10.2307/1384095

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre- registration. *International Review of Social Psychology*, *32*(1). http://doi.org/10.5334/irsp.289

Lonsdorf, T. B., Klingelhöfer-Jens, M., Andreatta, M., Beckers, T., Chalkia, A., Gerlicher, A., …, & Merz, C. J. (2019). How to not get lost in the garden of forking paths: Lessons learned from human fear conditioning research regarding exclusion criteria. https://doi.org/10.31234/osf.io/6m72g

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(1), 1-19. http://doi.org/10.18637/jss.v009.i08

Massey F. J. Jr. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association, 46*(253), 68-78.

Meade, A. W. (2010). Restriction of Range. In N. J. Sand (Ed.), *Encyclopedia of Research Design*. SAGE Publishing. Retrieved from https://sk.sagepub.com/reference/researchdesign/n388.xml

Merriam-Webster (n.d.). Bias. In *Merriam-Webster.com dictionary*. Retrieved January 26, 2021, from https://www.merriam-webster.com/dictionary/bias.

Mertens, G., & Krypotos, A.M. (2019). Preregistration of Analyses of Preexisting Data. *Psychologica Belgica*, *59*(1), 338–352. http://doi.org/10.5334/pb.493

Morgan, S. P. (1983). A research note on religion and morality: Are religious people nice people? *Social Forces*, 61(3), 683-692. http://doi.org/10.2307/2578129

Norenzayan, A., & Shariff, A. F. (2008). The origin and evolution of religious prosociality. *Science*, *322*(5898), 58-62. http://doi.org/10.1126/science.1158757

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistation is hard, and worthwhile. *Trends in Cognitive Sciences, 23*(10), 815-818. https://doi.org/10.1016/j.tics.2019.07.009

Olsson-Collentine, A., Van Assen, M. A., & Hartgerink, C. H. (2019). The prevalence of marginally significant results in psychology over time. *Psychological Science*, *30*(4), 576-586. https://doi.org/10.1177/0956797619830326

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. https://doi.org/10.1177/1745691612465253

Parliament of the World's Religions. (1993). Toward a global ethic: An initial declaration. Retrieved from https://www.weltethos.org/1-pdf/10-stiftung/declaration/declaration_english.pdf

Pew Research Center. (2015). America's changing religious landscape. *Pew Research Center*. Retrieved from https://www.pewforum.org/2015/05/12/americas-changing-religious-landscape

Pham, M. T., & Oh, T. T. (2021). Preregistration is neither sufficient nor necessary for good science. *Journal of Consumer Psychology, 31*(1), 163-176. https://doi.org/10.1002/jcpy.1209

Pharoah, C., & Tanner, S. (1997). Trends in charitable giving. *Fiscal Studies*, *18*(4), 427-443. https://doi.org/10.1111/j.1475-5890.1997.tb00272.x

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science, 1*(1), 27–42. https://doi.org/10.1177/2515245917745629

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

Saroglou, V. (2002). Religion and the five factors of personality: A meta-analytic review. *Personality and Individual Differences*, *32*(1), 15-25. https://doi.org/10.1016/S0191-8869(00)00233-6

Silver, N. C. (2008). Attenuation. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods*. SAGE Publishing. Retrieved from http://methods.sagepub.com/reference/encyclopedia-of-survey-research-methods/n24.xml

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General, 144*(6), 1146-1152. https://doi.org/10.1037/xge0000104

Smith, A. K., Ayanian, J. Z., Covinsky, K. E., Landon, B. E., McCarthy, E. P., Wee, C. C., & Steinman, M. A. (2011). Conducting high-value secondary dataset analysis: An introductory guide and resources. *Journal of General Internal Medicine*, *26*(8), 920-929. https://doi.org/10.1007/s11606-010-1621-5

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712. https://doi.org/10.1177/1745691616658637

Thompson, W. H., Wright, J., Bissett, P. G., & Poldrack, R. A. (2019). Dataset Decay: the problem of sequential analyses on open datasets. bioRxiv, 801696. https://doi.org/10.1101/801696

Veldkamp, C. L. S., Bakker, M., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., ..., & Wicherts, J. M. (2018). Ensuring the quality and specificity of preregistrations. https://doi.org/10.31234/osf.io/cdgyh

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(6), 632–638. https://doi.org/10.1177/1745691612463078

Wason, J. M., Stecher, L., & Mander, A. P. (2014). Correcting for multiple-testing in multi-arm trials: Is it necessary and is it done? *Trials, 15*(1), 364.
https://doi.org/10.1186/1745-6215-15-364

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of pre-Existing Datasets. *Advanced Methods and Practices in Psychological Science.* https://doi.org/10.1177/2515245919848684

Williams, M. N., & Albers, C. (2019). Dealing with distributional assumptions in preregistered research. *Meta-Psychology*, 3. https://doi.org/10.15626/MP.2018.1592

Youniss, J., McLellan, J. A., & Yates, M. (1999). Religion, community service, and identity in American youth. *Journal of Adolescence*, *22*(2), 243-253. https://doi.org/10.1006/jado.1999.0214

CHAPTER 7

Increasing the transparency of systematic reviews: Presenting a generalized registration form

Olmo R. van den Akker[1] *, Gjalt-Jorn Ygram Peters[2] *, Caitlin J. Bakker[3], Rickard Carlsson[4], Nicholas A. Coles[5], Katherine S. Corker[6], Gilad Feldman[7], David Moreau[8], Thomas Nordström[4], Jade S. Pickering[9], Amy Riegelman[10], Marta K. Topor[11], Nieky van Veggel[12], Siu Kit Yeung[7], Mark Call[13], David T. Mellor[13], & Nicole Pfeiffer[13]

[1] Department of Methodology & Statistics, Tilburg University, The Netherlands
[2] Department of Theory, Methodology & Statistics, Open University, The Netherlands
[3] Discovery Technologies Unit, University of Regina, Canada
[4] Department of Psychology, Linnaeus University, Sweden
[5] Center for the Study of Language and Information, Stanford University, United States
[6] Department of Psychology, Grand Valley State University, United States
[7] Department of Psychology, University of Hong Kong, Hong Kong
[8] School of Psychology and Center for Brain Research, University of Auckland, New Zealand
[9] Department of Psychology, University of York, United Kingdom
[10] Social Sciences Library, University of Minnesota, United States
[11] School of Psychology, University of Surrey, United Kingdom
[12] School of Animal and Human Sciences, Writtle University College, United Kingdom
[13] Center for Open Science, United States

* Olmo R. van den Akker and Gjalt-Jorn Ygram Peters contributed equally to this project

This paper is published in Systematic Reviews. See https://doi.org/10.1186/s13643-023-02281-7

Abstract

This paper presents a generalized registration form for systematic reviews that can be used when currently available forms are not adequate. The form is designed to be applicable across disciplines (i.e., psychology, economics, law, physics, or any other field) and across review types (i.e., scoping review, review of qualitative studies, meta-analysis, or any other type of review). That means that the reviewed records may include research reports as well as archive documents, case law, books, poems, etc. Items were selected and formulated to optimize broad applicability instead of specificity, forgoing some benefits afforded by a tighter focus. This PRISMA 2020 compliant form is a fallback for more specialized forms and can be used if no specialized form or registration platform is available. When accessing this form on the Open Science Framework website, users will therefore first be guided to specialized forms when they exist. In addition to this use case, the form can also serve as a starting point for creating registration forms that cater to specific fields or review types.

Background

Systematic reviews are systematic in the sense that they involve a systematic process to transparently, reproducibly, and often exhaustively identify and synthesize the literature on a given research topic. Even though objectivity is desirable for systematic reviews, the process is not immune to bias. Systematic reviewers are well aware of this, and many initiatives have been undertaken to identify and prevent biases. In 2011, a registry of systematic reviews (PROSPERO) was created to help researchers prospectively register health-related systematic review protocols (Booth et al., 2020). This registry was an important step in making the systematic review process more transparent as it facilitated documentation of the process and justifications for deviations from the protocol. The registry also allowed third parties to check the extent to which completed systematic reviews (as presented in journal articles) are carried out in line with the protocol, making it easier to identify decisions that may have introduced bias (e.g., a change in the criteria for study inclusion or the omission of an analysis without a valid rationale). Finally, PROSPERO allows researchers to check whether similar endeavors are underway prior to engaging in a systematic review, facilitating collaboration and synergy. In all, PROSPERO makes the systematic review process more transparent, and makes it feasible to identify and address biases so that they are less likely to influence the results of systematic reviews.

When registering a systematic review in the PROSPERO registry, researchers are presented with a registration form that they can use to specify their protocol (see Appendix 1 for the PROSPERO registration form). However, this form is optimized for health-related systematic reviews (either in humans or in animals). This serves PROSPERO's goal well but is necessarily exclusive to other systematic reviews. Specifically, PROSPERO directly excludes all systematic reviews without health outcomes, systematic reviews that are non-interventional, scoping reviews, evidence maps, and qualitative systematic reviews. PROSPERO's focus on health-related reviews also manifests itself through the items included in the registration form. For example, the form prompts specification of the "disease, condition or healthcare domain being studied", and the form assumes that some kind of intervention will take place, including mandatory fields where the intervention(s)/exposure(s) and the comparator(s)/control are specified, even though much research does not involve interventions.

To enable researchers to register systematic reviews for which PROSPERO is not suitable, we developed a generalized form for registering systematic reviews that is designed to be applicable across disciplines (i.e., psychology, economics, law, physics, or any other field) and across review types (i.e., scoping review, review of qualitative studies, meta-analysis, or any other type of review). This means that the reviewed records may include

research reports as well as archive documents, case law, books, poems, etc**.** Therefore, our selection of items and formulation of each item were optimized for broad applicability instead of specificity. Such generic formulation means some benefits afforded by a tighter focus (e.g., on a given method) may have been forgone. This form, therefore, is well suited as a fallback for more specialized forms and can be used if no specialized form or registration platform is available. When accessing this form on the Open Science Framework website (https://osf.io) users will therefore first be guided to specialized forms when they exist. If such a specialized form does not exist, we encourage users to reflect on whether this generalized form suits their needs, or whether it would be better to adapt the form into a form that better caters to the user's specific field or review type. As such, this generalized form can also function as a starting point for creating new registration forms.

To select items for this form, we assessed the items of several reporting guidelines and guides, most notably the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement. PRISMA was published to help researchers create reproducible reports of their systematic reviews, and with that alleviate biases in the reporting phase of systematic reviews (Liberati et al., 2009; Moher, Liberati, Tetzlaff, & Altman, 2009; for an updated version see Page et al., 2021). Inspired by PRISMA, additional reporting guidelines have been developed for specific disciplines (e.g., ROSES is tailored to systematic reviews in environmental research, and MOOSE is tailored to systematic reviews in epidemiology) and specific types of reviews (e.g., PRISMA-IPD is tailored to systematic reviews of individual participant data, and PRISMA-DTA is tailored to systematic reviews of diagnostic test accuracy). Where the Generalized Systematic Review Registration Form is optimized for accurate and comprehensive a priori documentation of systematic review procedures, reporting guidelines were optimized for application after completion of a systematic review. Because of these different end goals, reporting guidelines like PRISMA lack detail with respect to decisions that are important regarding the planning of a systematic review. In contrast, this form includes several decisions that are important to transparently document before data collection for the systematic review begins. At the same time, some PRISMA items can only be filled out once a systematic review is finished.

Nonetheless, there is also considerable overlap: these reporting guidelines do partly capture the same information as registration forms. Therefore, for each item in this form, we specified the corresponding PRISMA item (PRISMA items P1-P22 and P25-27 were applicable; P16-P23 cover reporting of results and P24 refers to registration forms like this). Researchers planning to use a specific reporting standard to report the results of their review, should enter the information required by that reporting standard in the corresponding (overarching) fields of this form.

The Generalized Systematic Review Registration Form is the result of a collaborative effort of several groups of researchers that independently identified the need for a systematic review registration form that is not restricted to a specific context. These groups initially started to build such forms based on their own research needs but when they learnt about each other's initiatives through Twitter and academic conferences they decided to combine resources and create this form. These existing resources were the PRISMA statement outlined above, a preregistration template specifically designed for non-interventional research (Topor et al., 2022), a registration form drafted at the conference of the Society for the Improvement of Psychological Science in 2018, and a registration form drafted for systematic reviews in animal research. We included all items that overlapped between two or more resources in the new form. For the remaining items, we decided collectively through Zoom-meetings and e-mail discussions whether and how to include them. In the final stage, the form was presented to experts in the scientific community to solicit feedback to improve generalizability and usability even more (Center for Open Science, 2023). Based on this feedback, the template was polished into its current state.

Instructions

To align with general use and open science best practices, when you fill out the form on the Open Science Framework, all items are mandatory. Being as comprehensive as possible makes your registration more useful for readers, funders, yourself, and others, so check carefully whether you did not accidently omit an item. If an item asks about a procedure you do not plan to use or is not applicable, indicate that in the corresponding field (including, ideally, the underlying reason).

You should be transparent about any deviations from the preregistration and provide the rationale for these deviations in your final review. If you already foresee some deviations when filling out the form (e.g., you anticipate that you will not have enough studies in a moderator group), provide a contingency plan for these deviations in the relevant parts of the registration. In addition, we recommend publishing updated registrations, allowing you to document and justify your decisions along the way in the same uniform format.

The aim of this registration form is to be optimally inclusive (i.e., to be usable for registration of any systematic review, regardless of scientific discipline or review type). This inclusivity is also signified by the fact that this form has been used for published papers involving scoping reviews, systematic reviews, narrative reviews, and meta-analyses from areas as diverse as psychology, political science, and biomedicine [Coen, Vezzoli, & Zogmaister, 2022; Chaxiong, Dimian, & Wolff, 2022; Hughes, Irwin, & Nestor, 2023; Evans et al., 2023; Yeung, Yay, & Feldman, 2022). Moreover, since it was made public on the

Open Science Framework (10 April 2023), the form has already been used 68 times as per 2 May 2023, which amounts to more than 20 completed registrations per week. Readers can see how the template is being used in research on the OSF Registry (https://bit.ly/osf-sytematic-reviews). Given that the form is relatively new and awareness of the form is expected to grow, we expect this average to increase more in later months and years.

Because this aim precludes 1:1 correspondence with the existing reporting guidelines, we want to emphasize that this form is also intended as a basis to develop more specialized forms that do correspond closely to more specific reporting guidelines. Such specialized forms can include, for example, additional fields, added comments, and worked examples. This form is included in the *preregr* R package (Van Eijk, Jiao, & Peters, 2023), and the underlying *preregr* form specification (https://osf.io/by27q) can be used to develop adapted versions of this form (e.g., Gültzow, Neter, & Zimmermann, 2023). Note that *preregr* can also be used to produce an R Markdown template containing this form presented in this paper, including the item labels and descriptions using the command "preregr::form_to_rmd_template('genSysRev_v1', file = 'C:/path/to/file.Rmd');".

The registration form is presented below. The items of the form are denoted by GSRRF, while the corresponding items from the PRISMA checklist are denoted by PRISMA.

Metadata

This metadata applies only to the registration you are creating, and will not be applied to your OSF project.

[GSRRF-1] **Title:**

Prisma: 1.

[GSRRF-2] **Contributors:**

[GSRRF-3] **Subjects:**

[GSRRF-4] **Tasks and roles:**

Describe the expected tasks and roles of each author/contributor, for example using the Contributor Roles Taxonomy (CRediT).

Review methods

In this section, you register the general type, background and goals of your review.

[GSRRF-5] **Type of review:**

This can be, for example, a meta-analysis, evidence map, or a qualitative review. Also indicate whether you used any guidelines, tools or checklists to prepare your protocol, and if so, which ones. For more information, see: Tricco, A. C., Tetzlaff, J., Moher, D. (2011). PRISMA: 1.

[GSRRF-6] **Review stages:**

Indicate the stages in which you will conduct this review. Common stages are, in this order, the sections of this form: Search, Screening, Extraction, Synthesis. Sometimes other stages are distinguished, such as Preparation, Critical Appraisal, and Reporting. Additionally, it can be beneficial to include pilot stages for screening and extraction, while mentioning any updates to the preregistration. The stages could then look like: Preparation, Search, Pilot Screening (100 hits), Prereg Update, Screening, Pilot Extraction (10 sources), Prereg update, Extraction, Synthesis.

[GSRRF-7] **Current review stage:**

Indicate in which stage from the list you specified in the "Review stages" item you are at this moment (i.e., when you freeze this registration). Note that in many contexts, only registrations in earlier stages count as preregistrations. For example, you can use a table to indicate whether you started and/or finished with a certain stage as is customary for PROSPERO registrations. In addition, if this is not the first preregistration (but a second or third update, e.g., after pilot screening or pilot extraction), you can make that explicit here.

[GSRRF-8] **Start date:**

Indicate the planned start date, or if you already started, the actual start date.

[GSRRF-9] **End date:**

Indicate the planned end date, or if you already completed the review, the actual end date. You can use resources such as PredicTER.org to estimate how long a review will take to complete.

[GSRRF-10] **Background:**

Introduce the topic of your review, its aims, and/or provide a short summary of known literature and what your review adds to this literature. You can describe why the review is needed, as well as which reviews already exist on this or related topics. PRISMA: 2 and 3.

[GSRRF-11] **Primary research question(s):**

List the specific questions this review is meant to answer (i.e., the questions that ultimately informed the decisions made when designing the search strategy, and screening, extraction, and synthesis plans). You may find it helpful to refer to frameworks such as PICOS where appropriate to pinpoint your research questions. Note that all analyses pertaining to primary research questions should normally be reported in the final report. PRISMA: 4.

[GSRRF-12] **Secondary research question(s):**

List additional research questions that you will examine, but that took less central roles in informing the review's design. Note that all analyses pertaining to secondary research questions should normally be reported in the final report. PRISMA: 4.

[GSRRF-13] **Expectations / hypotheses:**

Describe any hypotheses (common for quantitative approaches) and/or expectations you have. These can pertain to your research questions, the types of sources you will find, social and political contexts, and contextual information that you know may color your interpretations and decisions (common for qualitative approaches). PRISMA: 3

[GSRRF-14] **Dependent variable(s) / outcome(s) / main variables:**

List the dependent / outcome / main variables you are interested in. If this review concerns one or more associations, list the outcome variable(s) or dependent variables. If this review does not concern one or more associations (e.g., in reviews of single variables such as prevalences, or descriptive reviews), list the main variables of interest here. PRISMA: 10a.

[GSRRF-15] **Independent variable(s) / intervention(s) / treatment(s):**

If this review's research question(s) concerns one or more associations or effects, list the variable(s) that theoretically cause them or are assumed to otherwise explain the dependent variable(s) / outcome(s). If this is a manipulation, treatment, or intervention, make sure to describe it in full: that means also describing all groups, including any control group(s) or comparator(s). PRISMA: 10b.

[GSRRF-16] **Additional variable(s) / covariate(s):**

Here, list any additional variables you are interested in that were not included in the two lists above, such as covariates, moderators, or mediators. PRISMA: 10b.

[GSRRF-17] **Software:**

List the software you will use for the review, for instance to store and screen search results, extract data, keep track of decisions, and to synthesize the results. Include version numbers and the operating systems, if applicable. PRISMA: 13d.

[GSRRF-18] **Funding:**

List the funding sources for everybody that is involved in this review at this stage. If the work is unfunded, please state this as such. PRISMA: 25.

[GSRRF-19] **Conflicts of interest:**

List any potential conflicts of interest (e.g., if there is a potential outcome of this review that can in any way have negative or positive effects for anybody involved in this review in terms of funding, prestige, or opportunities). If there are no conflicts of interest, please state this as such. PRISMA: 26.

[GSRRF-20] **Overlapping authorships:**

Declare whether you expect that anyone involved in this review is a co-author of one of the studies that will likely be included in the review (based on your search strategy) and, if so, how you will address potential bias (i.e., that reviewer is not involved in screening, data extraction, quality assessment, or synthesis of that study). If you are confident that this does not represent a conflict of interest, explain why you think so. PRISMA: 26.

Search strategy

In this section, you register your search strategy: the procedures you designed to obtain all (potentially) relevant sources to review (e.g., articles, books, preprints, reports, case law, policy papers, archived documents).

[GSRRF-21] **Databases:**

List the databases you will search (e.g., ArXiv, PubMed, Scopus, Web of Science, PsycINFO, AGRIS, BioOne, PubChem). Note that these are different from interfaces (see below and here). PRISMA: 6.

[GSRRF-22] **Interfaces:**

For each database, list the interface you used to search that database (e.g., Ovid or EBSCO). Some databases are provided by the same organisation, in which case the interface can have the same name (e.g., PubMed, ArXiv). For more information about the distinction, see here. PRISMA: 6.

[GSRRF-23] **Grey literature:**

List your strategies for locating grey literature (i.e., sources not indexed in the databases you search) such as pre-prints (e.g., disciplinary repositories such as ArXiv or PsyArXiv or university repositories using for example, Dspace), dissertations and theses, conference proceedings and abstracts, government/industry reports etc. PRISMA: 6.

[GSRRF-24] **Inclusion and exclusion criteria:**

List the specific inclusion and exclusion criteria that you used to inform your search strategy. Also list the framework(s) you used to establish your exclusion and inclusion criteria and use them to develop your search query, if any. Examples of the latter are PICO (Population, Intervention, Comparison, Outcome) and SPIDER (Sample, Phenomenon of Interest, Design, Evaluation, Research type), but many more exist (see here for an overview based on the medical and health sciences). PRISMA: 5 and 13a.

[GSRRF-25] **Query strings:**

For each database/interface combination, list the query you will input (note that the available fields and operators can differ by database and by interface). The query string can be based on, for example, your inclusion criteria, the entities you want to extract

(see "extraction") and design requirements (e.g., qualitative studies, RCTs, or prevalence studies). PRISMA: 7.

[GSRRF-26] **Search validation procedure:**

Explain whether you plan to employ a search validation procedure, and if so, describe the procedure. Templates are available here. PRISMA: 7.

[GSRRF-27] **Other search strategies:**

List any additional search strategies you aim to employ, such as using the ascendancy approach (look through other sources cited in your included sources), the descendancy approach (look through the sources that cite your included sources using systems such as Crossref), or using other systems such as CoCites. PRISMA: 7.

[GSRRF-28] **Procedures to contact authors:**

Describe your procedures for contacting authors. Will you contact authors? When? How will you follow-up on your first contact? Do you plan to share meta-data about those communications, and if so, how do you ask authors' permission for that? Note that templates are available at https://osf.io/q8stz/. PRISMA: 7.

[GSRRF-29] **Results of contacting authors:**

Describe whether you plan to report the outcomes of contacting the authors (e.g., how many authors responded, how many authors sent data), and if so, how. PRISMA: 16a.

[GSRRF-30] **Search expiration and repetition:**

Depending on how quickly the literature in an area expands, searches can have limited expiration dates; and for living reviews, repetition is planned regardless of ideas about expiration. Will you repeat your search (for example, in the case of a living review), and if so, how many months or years after your first search? PRISMA: 7.

[GSRRF-31] **Search strategy justification:**

Search strategies are often compromises, balancing pragmatic considerations with scientific rigour. Here, describe the justifications for your decisions about the databases, interfaces, grey literature strategies, query strings, author contact procedures, and search expiration date. PRISMA: 7.

[GSRRF-32] **Miscellaneous search strategy details:**

Here, you can describe any details that are not captured in the other fields in this section. PRISMA: 7.

Screening

In this section, you register your screening procedure: the procedure you designed to eliminate all irrelevant sources from the results of the search strategy (and retain the relevant sources).

[GSRRF-33] **Screening stages:**

Describe the stages you will use for screening. For example, if you expect many hits, you may want to first screen based on titles only, in a second round also include abstracts and keywords, and in a third round screen based on full texts. Also indicate for each round whether the screening is done by a computer (e.g., AI), a human, or a computer supervised by a human. Don't forget to describe the deduplication procedure, if you implement it. PRISMA: 8.

[GSRRF-34] **Screened fields / blinding:**

Describe which bibliographic fields (e.g., title, abstract, authors) are visible during the screening, and which fields are blinded. For example, journal names, authors, and publication years can be hidden from screeners in an effort to minimize bias. PRISMA: 8.

[GSRRF-35] **Used exclusion criteria:**

List the specific exclusion criteria that you apply in your screening to eliminate sources from the set of sources identified in your search. Note that inclusion criteria are typically used to inform the search strategy; during screening, as soon as an exclusion criterion is met, an entry is excluded, and so, inclusion criteria are reformulated into exclusion criteria where applicable. PRISMA: 8.

[GSRRF-36] **Screener instructions:**

List or upload the instructions provided to the screener(s). PRISMA: 8.

[GSRRF-37] **Screening reliability:**

For each screening round, list the number of screeners and the procedure used to ensure independent screening.This can also mean that you declare that you only use one screener, use multiple screeners that work together, or that you will not implement procedures to ensure that the screening is conducted independently. Also explain the test you will use, if any, to assess screener agreement. PRISMA: 8.

[GSRRF-38] **Screening reconciliation procedure:**

If you use more than one screener, describe the procedure to deal with divergent screener decisions for each screener round (e.g., through discussion or input from an additional screener). PRISMA: 8.

[GSRRF-39] **Sampling and sample size:**

Describe whether you plan to use all sources included through the screening procedure, or whether you plan to sample from these sources (note that in most cases, all studies identified at this stage are kept). In case of the latter, describe the procedure you plan to use, the sample size analyses you conducted or will conduct, and the resulting required sample size if that is already available. If you plan to refrain from drawing conclusions, or draw more nuanced conclusions, describe that here as well. Finally, describe what you will do if a minimum required sample size or power is not reached (for your main analysis and any supplementary analyses). PRISMA: 8.

[GSRRF-40] **Screening procedure justification:**

Screening procedures are often compromises, balancing pragmatic considerations with scientific rigour. Here, describe the justifications for your decisions about the screening rounds, blinding, in/exclusion criteria, assurance, and reconciliation procedures. PRISMA: 8.

[GSRRF-41] **Data management and sharing:**

Describe whether and how you plan to share the sources you obtained from the searches in the databases (see Search Strategy) and the decisions each screener made in each screening round. List both the file format (e.g., BibTeX, RIS, CSV, XLSX), the repository, and any potential embargos or conditions for access. PRISMA: 27.

[GSRRF-42] **Miscellaneous screening details:**

Here, you can describe any details that are not captured in the other fields in this section. PRISMA: 8.

Extraction

In this section, you register your plans for data extraction: the procedures you designed to extract the data you are interested in from the included sources. Examples of such data are text fragments, effect sizes, study design characteristics, year of publication, characteristics of measurement instruments, final verdicts and associated penalties in a legal system, company turnovers, sample sizes, or prevalences.

[GSRRF-43] **Entities to extract:**

List all entities that will be extracted from each included source. Entities can be, for example, 1) variables such as values of independent and dependent variables, and potential moderators (e.g., means, standard deviations); 2) estimations of associations between variables or effect sizes (e.g., Pearson's r or Cohen's d); 3) qualitative data fragments (e.g., interview material or synthesized themes); 4) descriptions of the used methods such as the included studies' designs, sample sizes, sample characteristics, time between data collection sessions, and blinding procedures; 5) metadata such as authors, institutions, and year of publication; 6) and (other) risk of bias indicators. PRISMA: 10a, 10b, and 12.

[GSRRF-44] **Extraction stages:**

Describe the stages you will use for extraction. Examples of stages are: a training stage, a reliability verification stage, and a final extraction stage; or first extracting primary data and in a second stage risk of bias information; or two extractors working sequentially or in parallel. Also indicate for each stage whether the extraction is done by a computer (e.g., AI), a human, or a computer supervised by a human. PRISMA: 9.

[GSRRF-45] **Extractor instructions:**

List or upload the instructions provided to the extractors (i.e., those performing the data extraction). PRISMA: 9.

[GSRRF-46] **Extractor blinding:**

If blinding is used, describe the procedure used to blind extractors from the research questions, hypotheses, and/or specific roles of each entity to extract in this review. For example, extractors can be research assistants who are not informed of the study's background or research questions, but who are trained to extract entities using the coding instructions you developed for each entity; or entity extraction can be crowdsourced to citizen scientists. PRISMA: 9.

[GSRRF-47] **Extraction reliability:**

For each extraction round, list the number of extractors and the procedure used to ensure independent extraction (this can also mean that you declare that you use one extractor, or will not implement procedures to ensure that the extractions are conducted independently). Also explain the test you will use, if any, to assess extractor agreement. PRISMA: 9.

[GSRRF-48] **Extraction reconciliation procedure:**

For each extraction round, describe the procedure to deal with divergent extraction decisions (if applicable, i.e., if you use more than one extractor). PRISMA: 9.

[GSRRF-49] **Extraction procedure justification:**

Extraction procedures are often compromises, balancing pragmatic considerations with scientific rigour. Here, describe the justifications for your decisions about the justification of each entity that will be extracted, the extraction rounds, reliability assurance, and reconciliation procedures. PRISMA: 9.

[GSRRF-50] **Data management and sharing:**

Describe whether and how you will share the files with the extracted entities (as specified in the corresponding field above; i.e., everything extracted from every source, including metadata, method characteristics, variables, associations, etc). List both the file format (e.g., CSV, XLSX, Rdata), the repository, and any potential embargos or conditions for access. Describe efforts made to share FAIR, 5-star open data, if any such efforts will be made. PRISMA: 27.

[GSRRF-51] **Miscellaneous extraction details:**

Here, you can describe any details that are not captured in the other fields in this section. PRISMA: 9.

Synthesis and Quality Assessment

In this section, you register the procedure for the review's synthesis: the procedure you designed to use the data that was extracted from each source to answer your research question(s). This often includes transforming the raw extracted data, verifying validity, applying predefined inference criteria, interpreting results, and presenting results. Additionally, you register procedures you designed to assess bias in individual sources and the synthesis itself.

[GSRRF-52] **Planned data transformations:**

Describe your plans for transforming the raw extracted data. This may include converting effect sizes to other metrics (e.g., convert all metrics to Pearson correlation coefficients); recoding or (re)categorizing extracted qualitative data fragments (e.g., coding extracted music genres within an existing taxonomy); and aggregating extracted data prior to the main synthesis procedures (e.g., compute the mean of a variable over all samples in one source). Applying these transformations to the raw extracted entities from the Extraction section should yield data that corresponds to the variables of interest listed in the Review Methods section. PRISMA: 13b.

[GSRRF-53] **Missing data:**

Describe how you will deal with missing data (i.e., cases where it is not possible to extract one or more entities from the source material, and your efforts to obtain the missing information, for example by contacting the authors, are not fruitful). PRISMA: 10b.

[GSRRF-54] **Data validation:**

Describe your process of ensuring that the data are correct and useful (e.g., identifying outliers, identifying retractions, or triangulating with other sources). Also describe your criteria for assessing data validity and how you will deal with data violating those criteria. PRISMA: 10b.

[GSRRF-55] **Quality assessment:**

Describe the analyses you plan to do to assess and weigh the quality of the included sources with respect to your research question(s). Examples of tools to use for quality evaluation are Cochrane's Risk of Bias 2 tool, GRADE, and GRADE-CERQual. PRISMA: 11.

[GSRRF-56] **Synthesis plan:**

Describe the specific procedure you will apply to arrive at an answer to the research question(s). For example, in meta-analyses this is the full analysis plan, including any planned subgroup analyses and moderator analyses, the (multilevel) model specification, and preferably the analysis code. For a qualitative review, it is the procedure you plan to use to collate your results into a coherent picture. If you distinguish synthesis tiers (e.g., primary and secondary analysis, or confirmatory and exploratory analyses), list them and indicate which procedures you plan to use for each. Also specify what you will do if parts of the plan can't be properly executed. PRISMA: 13c, 13d, and 13e.

[GSRRF-57] **Criteria for conclusions / inference criteria:**

If you plan to draw your conclusions based on pre-specified criteria (e.g., a minimal effect size of interest, a significance level, or a saturation point), list these here. PRISMA: 20b.

[GSRRF-58] **Synthesist blinding:**

Describe the procedure, if any, used to blind synthesists (i.e., the persons synthesizing the extracted data to arrive at answers to your research question(s)) from the research questions, hypotheses, and/or specific roles of each extracted entity/variable in this review. For example, for meta-analyses, an analyst external to the main research team can be engaged to perform the analyses without knowing the study's hypotheses. For qualitative reviews, for the synthesis, other researchers can be involved who are unaware of and are not informed about the research process and expectations. PRISMA: 13d.

[GSRRF-59] **Synthesis reliability:**

List the number of synthesists and the procedure used to ensure independent synthesis (this can also mean that you declare that you use one synthesist, or will not implement procedures to ensure that the syntheses are conducted independently). PRISMA: 13d.

[GSRRF-60] **Synthesis reconciliation procedure:**

Describe the procedure to deal with divergent synthesis decisions (if relevant). PRISMA: 13d.

[GSRRF-61] **Publication bias analyses:**

Describe the analyses you plan to do to assess publication bias (if any). For an overview of commonly used publication bias correction methods, see Table 1 in Van Aert, Wicherts, & Van Assen (2019). PRISMA: 14.

[GSRRF-62] **Sensitivity analyses / robustness checks:**

Describe the sensitivity analyses or robustness checks you plan to conduct (if any). PRISMA: 13f and 15.

[GSRRF-63] **Synthesis procedure justification:**

Extraction procedures are sometimes compromises, balancing pragmatic considerations with scientific rigour. Here, describe the justifications for your decisions about your planned transformations (e.g., if based on assumptions, how do you know those are feasible), your data integrity and missing data checks and corrections, your synthesis plan, the criteria you chose to drive your conclusions/inferences (if any), and your procedures for blinding, and reliability assurance/reconciliation if you use multiple synthesists. PRISMA: 13d.

[GSRRF-64] **Synthesis data management and sharing:**

Describe whether and how you will share the files with the analysis scripts, notes, and outputs. List both the file format (e.g., R scripts, Rmarkdown files, plain text files, Open Document files), the repository, and any potential embargos or conditions for access. See here for a generic example of an analysis script. PRISMA: 27.

[GSRRF-65] **Miscellaneous synthesis details:**

Here, you can describe any details that are not captured in the other fields in this section. PRISMA: 13d.

References

Booth, A., Mitchell, A. S., Mott, A., James, S., Cockayne, S., Gascoyne, S., & McDaid, C. (2020). An assessment of the extent to which the contents of PROSPERO records meet the systematic review protocol reporting items in PRISMA-P. *F1000*, 9, 773. https://doi.org/10.12688/f1000research.25181.2

Center for Open Science (2022). Systematic Review Registration Template Community Call. https://www.youtube.com/watch?v=1bAReid9Ffw&ab_channel=CenterforOpenScience. Accessed May 3, 2023.

Chaxiong, P., Dimian, A. F., & Wolff, J. J. (2022). Restricted and repetitive behavior in children with autism during the first three years of life: A systematic review. *Frontiers in Psychology*, 13, 986876. https://doi.org/10.3389/fpsyg.2022.986876

Coen, S., Vezzoli, M., & Zogmaister, C. (2022). Theoretical and methodological approaches to activism during the COVID-19 pandemic—between continuity and change. *Frontiers in Political Science*, 89, 844591. https://doi.org/10.3389/fpos.2022.844591

Evans, T. R., Burns, C., Essex, R., Finnerty, G., Hatton, E., Clements, A. J., ... et al. (2023). A systematic scoping review on the evidence behind debriefing practices for the wellbeing/emotional outcomes of healthcare workers. *Frontiers in Psychiatry*, 14, 1078797. https://doi.org/10.3389/fpsyt.2023.1078797

Gültzow, T., Neter, E., & Zimmermann, H. (2023). Making Research Look Like the World Looks: Introducing the 'Inclusivity & Diversity Add-On for Preregistration Forms' Developed During an EHPS2022 Pre-Conference Workshop. https://doi.org/10.31219/osf.io/r2e7a

Hughes, L. M., Irwin, M. G., & Nestor, C. C. (2023). Alternatives to remifentanil for the analgesic component of total intravenous anesthesia: a narrative review. *Anaesthesia*, 78(5), 620–625. https://doi.org/10.1111/anae.15952

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., ... et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Annals of Internal Medicine*, 151(4), W65–94. https://doi.org/10.1016/j.jclinepi.2009.06.006

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, 151(4), 264–269. https://doi.org/10.1136/bmj.b2535

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International Journal of Surgery*, 88, 105906. https://doi.org/10.1136/bmj.n71

Topor, M. K., Pickering, J. S., Mendes, A. B., Bishop, D., Büttner, F., Elsherif, M. M., ... et al. (2022). An integrative framework for planning and conducting Non-Intervention, Reproducible, and Open Systematic Reviews (NIRO-SR). *Meta-Psychology*. https://doi.org/10.31222/osf.io/8gu5z

Van Eijk, N. L., Jiao, H., & Peters, G.-J. Y. (2023). Making Preregistration Accessible: An R Package to Make Machine-Readable Preregistrations and Create New Preregistration Forms. *PsyArXiv*. https://doi.org/j3c3

Yeung, S. K., Yay, T., & Feldman, G. (2022). Action and inaction in moral judgments and decisions: Meta-analysis of omission bias omission-commission asymmetries. *Personality and Social Psychology Bulletin*, 48(10), 1499–1515. https://doi.org/10.1177/01461672211042315

CHAPTER 8

Summary and discussion

In this concluding chapter, I review the empirical evidence presented in Chapters 2 to 6 on whether preregistration achieves its two main goals: increasing transparency and reducing bias (Hardwicke & Wagenmakers, 2023). I also discuss the key features of the preregistration templates presented in Chapters 7 and 8. Afterwards, I discuss any implications of the findings in this dissertation for the future of preregistration in psychology and provide concrete recommendations that may help improve preregistration uptake and effectiveness, and with that, the quality and replicability of research in psychology and beyond.

Summary

In Chapter 2, we assessed the extent of selective hypothesis reporting in psychological research by comparing the hypotheses found in a set of 459 preregistrations to the hypotheses found in the corresponding papers. We found that more than half of the preregistered studies contained omitted hypotheses (N_s = 224; 52%) or added hypotheses (N_s = 227; 57%), and about one-fifth of studies contained hypotheses with a direction change (N_s = 79; 18%). We only found few studies with hypotheses that were demoted from primary to secondary importance (N_s = 2; 1%) and no studies with hypotheses that were promoted from secondary to primary importance. However, this small number may have to do with the fact that categorizing hypotheses as primary or secondary is not as common in psychology compared to fields like medicine. In all, 60% of studies included at least one hypothesis in one or more of these categories, indicating a substantial bias in presenting and selecting hypotheses by preregistering researchers and possibly reviewers/editors. Contrary to our expectations, we did not find sufficient evidence that added hypotheses and changed hypotheses were more likely to be statistically significant than non-selectively reported hypotheses. For the other types of selective hypothesis reporting, we might have lacked sufficient statistical power to detect this relationship. Finally, we found that replication studies were less likely to include selectively reported hypotheses than original studies. Thus, selective hypothesis reporting is problematically common in psychological research.

In Chapter 3, we assessed the effectiveness of preregistration in restricting potentially biasing researcher degrees of freedom. We used an extensive protocol to assess the producibility of preregistrations (i.e., the extent to which the study can be properly conducted based on the information in the preregistration) and the consistency between preregistration and publications of 300 preregistered psychology studies. We found that preregistrations often lack methodological details and that deviations from preregistered plans were rarely disclosed. For example, only 22% - 35% of deviations for the data collection procedure and about 15%-20% of deviations for the exclusion criteria and

statistical model were disclosed. Combining the producibility and consistency results highlights that biases due to researcher degrees of freedom remain possible in many preregistered studies. More comprehensive registration templates typically yielded more producible and, hence, better preregistrations. We did not find that the effectiveness of preregistrations differed over time or between original and replication studies. Furthermore, we found that operationalizations of variables were generally more effectively preregistered than other study parts. Inconsistencies between preregistrations and published studies were mainly encountered for data collection procedures, statistical models, and exclusion criteria.

The results in Chapters 2 and 3 highlight that researchers often do not preregister their studies optimally. Notably, comparing preregistrations and the corresponding papers was challenging because researchers often switched variable names, used inconsistent notations, or used ambiguous language. Accentuating these issues, preregistrations and papers were often written in different formats; preregistrations were often structured based on preregistration templates, whereas papers were often structured based on journal requirements. Even though preregistration currently does not fulfill its full potential, it could still help reduce the number of false positives in the psychological literature to some extent.

In Chapter 4, we compared 193 psychology studies that earned a Preregistration Challenge prize or Preregistration Badge to 193 similar studies that were not preregistered. In contrast with our theoretical expectations and prior research (Schäfer & Schwarz, 2019; Toth et al., 2021), we did not find that preregistered studies had a lower proportion of positive results (Hypothesis 1), smaller effect sizes (Hypothesis 2), and fewer statistical errors (Hypothesis 3) than non-preregistered studies. Supporting our Hypotheses 4 and 5, we found that preregistered studies more often contained power analyses and typically had higher sample sizes than non-preregistered studies. Both these study characteristics are associated with higher research quality. Finally, concerns about the publishability and impact of preregistered studies seem unwarranted as preregistered studies did not take longer to publish and scored better on several impact measures. Overall, our data indicate that preregistration has beneficial effects in the realm of statistical power and impact, but we did not find robust evidence in this study that preregistration prevents *p*-hacking and Hypothesizing After the Results are Known (HARKing).

In Chapter 5, we conducted two vignette studies to examine how psychology researchers interpret the results of a set of four replications that are either preregistered or not. Only a small proportion (Study 1: 1.6%; Study 2: 2.2%) of participants used the normative method of Bayesian inference, whereas many of the participants' responses were in line with generally dismissed and problematic vote counting approaches. These two

studies demonstrated that many psychology researchers underestimate the evidence in favor of a theory if one or more results from a set of replication studies are statistically significant, highlighting the need for better statistical education. In both studies, we found that participants' belief in the theory increased with the number of statistically significant results and that the result of a direct replication had a stronger effect on belief in the theory than the result of a conceptual replication. In Study 2, we additionally found that participants' belief in the theory was lower when they assumed the presence of *p*-hacking, but that belief in the theory did not differ between preregistered and non-preregistered replication studies. In analyses of individual participant data from both studies, we examined the heuristics academics use to interpret the results of four experiments.

In Chapters 6 and 7 we took a more practical approach and developed two preregistration templates. These templates are timely given the result in Chapter 3 that preregistration templates help writing more producible preregistrations. In Chapter 6, we presented a preregistration template specifically aimed at secondary data analyses, in which new analyses are carried out using existing data. Such a template is important given that researchers' hypotheses and analyses may be biased by their prior knowledge of the data. The need for proper guidance in this area is especially clear now that data are increasingly shared publicly. In this tutorial, we presented a template for the preregistration of secondary data analyses and provided comments and a worked example that may help with using the template effectively. Through this illustration, we showed that completing such a template is feasible, helps limit researcher degrees of freedom, and may make researchers more deliberate in their data selection and analysis efforts.

In Chapter 7, we presented a generalized registration form for systematic reviews that can be used when currently available forms are not adequate. The form is designed to be applicable across disciplines (i.e., psychology, economics, law, physics, or any other field) and across review types (i.e., scoping review, review of qualitative studies, meta-analysis, or any other type of review). So, the reviewed records may include research reports, but also archive documents, case law, books, poems, etc. Items were selected and formulated to optimize broad applicability instead of specificity, forgoing some benefits afforded by a tighter focus. This PRISMA 2020 compliant form is a fallback for more specialized forms and can be used if no specialized form or registration platform is available. When accessing this form on the Open Science Framework website, users will therefore first be guided to specialized forms when they exist. In addition to this use case, the form can serve as a starting point for creating registration forms that cater to specific fields or review types.

Discussion

According to Hardwicke and Wagenmakers (2023), preregistration has two main objectives: (1) increasing transparency and (2) reducing bias. Bias is defined as a systematic deviation of results (and interpretation of those results) from the truth and is typically reflected by false positive results and overinflated effect sizes. The studies presented in Chapters 2 to 4 of this dissertation clearly demonstrate that preregistration fulfills the first objective. Preregistration made it possible for us to assess study plans and compare them to the realized studies, which is impossible for the vast number of studies in the psychological literature that have not been preregistered and for which the study plans are obscured or vague to begin with.

Aside from being useful for meta-scientific projects like ours, the increased transparency also allows readers to more accurately assess the validity of the claims made in papers or, more formally, the severity of the empirical tests in papers (Lakens, 2019). Specifically, transparency allows readers to assess the extent to which researcher degrees of freedom were left open in a study. The more researcher degrees of freedom were left open, the higher the likelihood of *p*-hacking and HARKing, the less falsifiable the study hypotheses, and the less severe the hypothesis tests. However, preregistration effectiveness does not *necessarily* imply severe tests. It could be that a researcher restricted all researcher degrees of freedom but chose an inappropriate statistical test that would have supported a hypothesis even if it were untrue. Consequently, preregistration is not sufficient for severity; it merely provides information about the degree of severity. Or, in terms of the two objectives of Hardwicke and Wagenmakers (2023): increased transparency does not necessarily reduce bias.

It also became clear in this dissertation that it is hard to assess whether preregistration reduces bias in psychology. Comparing preregistrations with papers was a challenge across the board, from hypotheses to statistical models and from variables to inference criteria. This is arguably most problematic in the case of hypotheses, which can be viewed as the building blocks of science. If such elemental parts of a research study are already rooted in ambiguous language, like we found in Chapter 2, it makes it difficult or even impossible to assess the other parts of a study. This is exemplified by the fact that from the 459 studies with preregistered hypotheses that we considered in Chapter 2, only 300 studies involved hypotheses that were sufficiently consistent between preregistration and paper for us to check in Chapter 3. We believe that this discrepancy is largely due to the ambiguous language used when researchers specify their hypotheses, impeding our ability to match hypotheses that were intended to be identical. Moreover, ambiguous hypotheses in preregistrations leave open researcher degrees of freedom because researchers can argue in favor of more research choices (like the choice of

control variables or the statistical model) that 'fit' the ambiguous hypothesis than in the case of producible and falsifiable hypotheses. The language used in preregistrations and subsequent papers should be unambiguous and consistent to counter biases fully.

One way to facilitate preregistration-paper comparisons could be to encourage more journal editors to take up the registered reports format. In registered reports, where peer review takes place before data collection, preregistrations typically already involve an introduction and a methods section, just like in research papers. This alignment between preregistrations and papers should facilitate comparisons between the two. Moreover, in the registered reports format, second stage reviewers are asked to pay particular attention to any changes between preregistration and paper. The explicit alignment between preregistration and paper in registered reports might be the reason that the proportion of positive results in registered reports is lower than in non-preregistered papers, an indication of less bias (Scheel, Schijen, & Lakens, 2021). Another reason may be that acceptance decisions for registered reports are made before data collection and are thus not contingent on the results. At the moment, the registered report format is offered by more than 300 journals, and more than 600 registered reports have been published (Chambers and Tzavella, 2022). While this is promising, it pales in comparison to the more than 130,000 'regular' preregistrations posted on the Open Science Framework, and the more than 350,000 registrations posted on ClinicalTrials.gov (see https://osf.io/registries/discover). The lower popularity of registered reports could be because traditional journals struggle to incorporate this new format into their workflow, or because researchers fail to see the added value of registered reports. Because the registered report format shows so much potential in reducing bias in the scientific literature, I believe it is important to address these issues.

Recommendation 1: *More journals should allow researchers to submit registered reports.*

Recommendation 2: *Funders should encourage or even mandate that any confirmatory, hypothesis-testing studies they fund are submitted as registered reports.*

Recommendation 3: *Researchers should be educated about the potential of registered reports to increase transparency and reduce bias in confirmatory, hypothesis-testing studies.*

As registered reports become more popular, it must be stressed that there are some growing pains that need to be resolved. For example, Hardwicke et al. (2018) found that only half of the journals that accepted stage 1 protocols made them publicly available and that the registration and reporting of registered reports often lacked standardization. Moreover, researchers were often faced with substantial time delays because the review process took two stages instead of one. This is particularly problematic in case

of publicly funded research that often involves a fixed project duration. In their paper reviewing the current state of registered reports, Chambers and Tzavella (2022) suggest solutions for these issues like rapid review (i.e., a network in which reviewers agree to evaluate submissions within a short time frame), scheduled review (i.e., editors acquire reviewers and schedule reviews at the same that authors are preparing the manuscript), and observer-evaluator review (i.e., researchers upload study protocols, code, and data to a virtual space while reviewers review these on a rolling basis), but these should first be empirically validated before they can be employed to relieve the growing pains.

Recommendation 4: *Meta-researchers should initiate studies to empirically identify the causes and test potential solutions for the lack of transparency, the lack of standardization, and the time delays associated with registered reports.*

An option to improve preregistration-paper comparisons outside of the registered report format is to have journal editors recommend or even require that 'regular' reviewers specifically assess preregistration-paper consistency. This would create an incentive for researchers to use consistent language in both the preregistration and the paper making them easier to compare. The downside of this could be that the review time increases, putting the peer review system under more strain than it already is. This is a risk based on our experiences comparing preregistrations and papers in this study, which could sometimes take hours per preregistration-paper pair. On the other hand, a recent study piloting a procedure for discrepancy reviews showed that preregistration-paper comparisons can be effectively implemented in the peer review process without much extra costs (TARG Meta-Research Group and Collaborators, 2022). A way to decrease the review burden would be to specifically link the study parts presented described in papers to their corresponding study parts in preregistrations. This would drastically decrease review time because reviewers would not need to search for the specific location that a study part is described. Ideally, researchers would click on (or hover over) a certain part of a paper to directly see what the preregistration said about that part of the study. Having a separate section in a paper or in supplementary materials dedicated to outlining all deviations from the preregistered plan could also prove useful.

Recommendation 5: *Meta-researchers should conduct studies assessing how preregistration-paper comparisons can be implemented into the peer review system as efficiently as possible.*

Recommendation 6: *Journals should invest in the technical specifications necessary to improve the efficiency of preregistration-paper comparisons.*

While preregistration-paper comparisons can be challenging, they are sometimes not even possible at all. This happens in instances where researchers do not include the necessary information in the preregistration and/or the paper. For example, Chapter 3 showed that measured variables, dependent variables, data collection procedures, and statistical models could not be compared in about 15% of the studies. For manipulated variables, this was even worse with almost 30%. While this incompatibility is largely because *preregistrations* did not provide sufficient information, the *papers* also regularly failed to provide the required information to allow preregistration-paper comparisons. Incomplete reporting in published papers has already been widely acknowledged in the scientific community (Chalmers, 1990; Simera et al., 2010), but Chapters 2 and 3 highlight that we must also focus on incomplete reporting in preregistrations. That is, preregistrations are currently not producible enough.

Recommendation 7: *Meta-researchers should empirically assess preregistration templates on the extent to which they allow for producible preregistrations, possibly using newly developed guidelines.*

Preregistration templates with specific goals (e.g., those tailored to a certain type of study) like we provide in Chapters 6 and 7 could help with this issue as such templates provide researchers with prompts about what study parts and study elements to report in a preregistration. Specific education about the practice of preregistration would be another step forward. Currently, many students are already accustomed to the concept of preregistration by having to specify thesis proposals. While this is a good first step, these thesis proposals could be embedded in concrete preregistration training programs. Such tailor-made preregistration training would give aspiring researchers a good starting point to use preregistration during their first (and later) steps in academia. When this proves successful, similar training modules could be developed for more experienced researchers.

Recommendation 8: *(Under)graduate methods curricula in psychology should include a specific course around thesis proposals that outlines best preregistration practices and resources and discusses the theoretical and practical benefits and challenges of preregistration.*

In Chapter 4, we tried to assess whether preregistration prevents *p*-hacking and HARKing. We used the proportion of positive results and the size of effects in papers as proxies for these questionable research practices. We did not find a difference in the number of positive results and the size of effects, which would indicate that although some researcher degrees of freedom are restricted there are still many left over for researchers to engage in *p*-hacking and/or HARKing. However, there are alternative explanations

for the unexpected result in Chapter 4. For example, in contrast to previous studies that found positive results to be less common in preregistered vs. non-preregistered papers (Schäfer & Schwarz, 2019; Toth et al., 2021), we included virtually *all* statistical results in our control sample of non-preregistered studies. If we had focused on key statistical results, like in the previous studies, the proportion of positive results may have been higher, and we may have found the expected relationship between preregistration and positive results. This approach would probably be more prudent because it could be that we included statistical results that were not meant as hypothesis tests.

More generally, preregistered publications and non-preregistered publications can differ in other aspects than preregistration status. For example, researchers probably self-select into preregistration. Researchers who preregister may be more conscientious or more concerned with abiding by responsible research practices like preregistration than researchers who do not. On the other hand, it could be that researchers only decide to preregister hypotheses that they are unsure of as the preregistration might give them more certainty that their study will be published regardless of the result. In short, we need to be careful about making causal claims about the effect of preregistration on the proportion of positive results or effect size. Looking at related meta-research may be prudent here. For example, evidence from simulation studies (Stefan & Schönbrodt, 2003)) and studies comparing meta-analyses to preregistered multilab replications (Kvarven, Strømland, & Johannesson, 2020) do imply preregistration can be effective. That being said, this dissertation does not provide convincing evidence that preregistration currently achieves the goal of reducing bias. Future studies may aim to identify the characteristics of preregistering and non-preregistering researchers so that these variables could be included as control variables in studies like ours.

Recommendation 9: *Meta-researchers should study further whether preregistration is effective in reducing bias, taking into consideration potential confounding variables.*

That this dissertation could not establish solid evidence for preregistration as a tool to reduce bias raises the question whether preregistration is worthwhile, a question also reflected on by other researchers (Nosek et al., 2019; Szollosi et al, 2020). However, I believe it is. The research in this dissertation as well as other studies (Claesen, Gomes, Tuerlinckx, & Vanpaemel, 2019; Schäfer & Schwarz, 2019; Toth et al., 2021) shows that preregistration allows one to identify the selective reporting of hypotheses and results, and the use of vague or ambiguous language. In more formal terms, preregistration increases the transparency of the research process and allows one to assess the severity of the tests in research papers. This makes preregistration valuable in and of itself, even without considering any secondary benefits of preregistration, like that it improves the methodological quality of research (Sarafoglou, Kovacs, Bakos, Wagenmakers, & Aczel,

2022), and allows meta-research on variations in study design and statistical analysis. At the same time, this dissertation makes clear that there is much to be improved regarding preregistration practices and preregistration infrastructure. Yet, this should not be surprising given that preregistration has only had a salient place in the field of psychology for about a decade. Evidence indicates that students feel that doing preregistrations improves their preregistration skills (Pownall et al., 2023; Sarafoglou et al) so in due time it is likely that the scientific community can better extract the full potential of preregistration.

Recommendation 10: *Meta-researchers should conduct studies assessing whether and how preregistration skills improve over time and with experience*.

Aside from assessing whether the main goals of preregistration are achieved, Chapter 4 also answered some questions about other issues surrounding preregistration. For example, it alleviated some concerns that were levied about preregistration by Kornell (2013) and Goldin-Meadow (2016). They argued that preregistration does not allow for exploration and thus serendipities. This claim is simply untrue as preregistration merely requires one to distinguish between confirmatory and exploratory analyses, not forego exploratory analyses altogether. Additionally, they argued that preregistration would lead to unstructured or uninteresting papers because the restrictive nature of preregistration sometimes gets in the way of a 'clean' narrative. While this can be true, from a scientific perspective a less readable paper with valid results is more valuable than a readable paper with questionable results (Giner-Sorolla, 2012). Moreover, we found that preregistered studies did not take longer to be published and even had higher scores on several impact measures than non-preregistered studies. Concerns about the publishability of preregistered studies thus seem unwarranted but it would be good to study this more thoroughly. Other concerns, for example that preregistration will serve as an empty signal of good research could have merit and thus should also be carefully studied.

Recommendation 11: *Meta-researchers should study whether statistically non-significant studies that are preregistered are more publishable than statistically non-significant studies that are not preregistered.*

Recommendation 12: *Meta-researchers should conduct studies to assess the validity of concerns levied against the practice of preregistration.*

In Chapter 5, preregistration played a less major role than in Chapters 2-4. We found that many researchers use simplistic vote-counting heuristics with low statistical power (Hedges & Olkin, 1980) instead of the normative method of Bayesian inference when assessing sets of replication studies. This indicates that the statistical intuition of many

researchers is suboptimal (see also Aczel et al., 2018; Gigerenzer, 2018; Hoekstra, Finch, Kiers, & Johnson, 2006; Hoekstra, Morey, Rouder, & Wagenmakers, 2014) and that the width and depth of statistical education should be improved. Preregistration could help with this because it prompts researchers to do power analyses (see Chapter 4), which would help researchers discern and better understand the relationship between statistical significance, effect size, and sample size. Chapter 5 also showed that researchers' belief did not differ for preregistered and non-preregistered studies but that their belief in a theory typically was lower when they assumed that *p*-hacking took place. Even though the study involved a vignette instead of practical research scenarios, we concluded that psychology researchers are skeptical of statistically significant results when they consider the possibility of *p*-hacking, but that they are also skeptical about the ability of preregistration to effectively prevent *p*-hacking. The latter makes sense in light of findings in Chapters 2 and 3 that preregistrations are not always sufficiently producible to prevent *p*-hacking and are also often not adhered to exactly.

Recommendation 13: *Creators of preregistration templates should include items prompting researchers to do a power analysis if the planned study is confirmatory.*

Recommendation 14: *Meta-researchers should further study how the results of preregistered studies are evaluated compared to the results of non-preregistered studies.*

The empirical evidence outlined above can be characterized as a mixed bag when looking at the effectiveness of preregistration. But whether it is effective or not, it cannot be denied that preregistration has rapidly gained popularity in the scientific community. This is exemplified by the usage of the preregistration templates we described in Chapters 6 and 7. As of 7 July 2023, the template for secondary data analysis (Chapter 6) has been used 1,117 times to create a preregistration on the Open Science Framework, and the registration form for systematic reviews (Chapter 7) has been used 374 times. This shows that researchers are eager to preregister their studies and there is a lot of momentum. While these are impressive numbers, it is important to keep in mind the empirical evidence presented in this dissertation. Preregistration only functions to reduce bias if preregistrations are producible and if preregistrations align with the eventual papers. It is therefore of vital importance that those who preregister do it well. The widespread use of preregistration templates, and the increased uptake of the registered reports format should help, because it makes the process more structured. Specific education about preregistration would help to inform (aspiring) researchers more clearly about the goals, potential, challenges, and intricacies of preregistration. Finally, more meta-research like that in this dissertation could provide a more solid evidence-base about the practical benefits and downsides of preregistration, not only in psychology but also in other disciplines.

References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., … , & Wagenmakers, E. J. (2018). Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation. *Advances in Methods and Practices in Psychological Science, 1*(3), 257- 366.

Chalmers, I. (1990). Underreporting research is scientific misconduct. *JAMA, 263*, 1405-1408.

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*(1), 29-42.

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), 211037.

Goldin-Meadow, S. (2016). Why preregistration makes me nervous. *APS Observer*, *29*.

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science, 1*(2), 198-218

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.

Hardwicke, T. E., & Wagenmakers, E. J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour*, *7*(1), 15-26.

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin, 88*(2), 359–369.

Hoekstra, R., Finch, S., Kiers, H. A., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of *p* values. *Psychonomic Bulletin & Review*, *13*(6), 1033-1037.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157-1164.

Kornell, N. (2013, July 29). Some concerns on regulating scientists via preregistration. *Psychology Today*. Retrieved from https://www.psychologytoday.com/za/blog/everybody-is-stupid-except-you/201307/some-concerns-regulating-scientists-preregistration

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423-434.

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review, 62*(3), 221–230.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., ... & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815-818.

Pownall, M., Pennington, C. R., Norris, E., Juanchich, M., Smailes, D., Dr, Russell, P. S., … Clark, K. (2023, August 18). Evaluating the pedagogical effectiveness of study preregistration in the undergraduate dissertation: A Registered Report. https://doi.org/10.24072/pci.rr.100437

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E. J., & Aczel, B. (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, *9*(7), 211997.

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

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

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K. F., & Altman, D. G. (2010). Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Medicine*, *8*(1), 1-6.

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2), 220346.

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94-95.

TARG Meta-Research Group and Collaborators. (2022). Discrepancy review: A feasibility study of a novel peer review intervention to reduce undisclosed discrepancies between registrations and publications. *Royal Society Open Science*, *9*(7), 220142.

Toth, A. A., Banks, G. C., Mellor, D., O'Boyle, E. H., Dickson, A., Davis, D. J., ... & Borns, J. (2021). Study preregistration: An evaluation of a method for transparent reporting. *Journal of Business and Psychology*, *36*, 553-571.

Nederlandse samenvatting

Dit proefschrift draait om preregistratie, de werkwijze waarbij onderzoekers hun hypothesen, onderzoeksopzet en/of analyseplan publiceren voordat ze gegevens verzamelen of analyseren. Hoewel deze werkwijze al in de jaren 1950 werd voorgesteld als een nuttig instrument voor onderzoekers, is het pas in de jaren 2000 gangbaar geworden in de biomedische wetenschappen en in de jaren 2010 in de psychologie.

In de psychologie was de belangrijkste aanleiding voor de opkomst van preregistratie dat vele belangrijke resultaten niet werden gevonden in nieuwere studies met eenzelfde onderzoeksopzet (d.w.z. ze konden niet worden gerepliceerd). Dit leidde tot de zogenaamde replicatiecrisis, een staat van onzekerheid over welke bevindingen in het onderzoeksveld waar waren en welke onwaar. Deze staat van onzekerheid bracht veel psychologie-onderzoekers ertoe na te denken over de wetenschappelijke praktijken in het veld, hetgeen hielp bij het identificeren van mogelijke oorzaken en oplossingen voor de replicatiecrisis. Een van de vele voorgestelde oplossingen was preregistratie. Volgens Tom Hardwicke en Eric-Jan Wagenmakers heeft preregistratie voornamelijk tot doel (1) transparantie te vergroten en (2) *bias* door slechte onderzoekspraktijken (vaak gemeten door het aantal vals-positieve resultaten) te voorkomen. In Hoofdstukken 2 tot en met 5 presenteerde ik empirisch bewijs over de vraag of preregistratie in het veld van de psychologie deze twee hoofddoelen heeft bereikt. In Hoofdstukken 6 en 7 presenteerde ik twee *templates* die onderzoekers kunnen gebruiken om secundaire data-analyses en systematische reviews te preregistreren. Hieronder vind je een overzicht van alle gevonden inzichten uit deze zeven hoofdstukken.

In Hoofdstuk 2 onderzochten we de mate van selectieve rapportage van hypothesen in psychologisch onderzoek door de hypothesen in een set van 459 preregistraties te vergelijken met de hypothesen in de bijbehorende artikelen. We ontdekten dat meer dan de helft van de geregistreerde studies hypothesen weglieten (N = 224; 52%) of toevoegden (N = 227; 57%), en ongeveer een vijfde van de studies hypothesen met een veranderde richting bevatten (N = 79; 18%). We vonden slechts weinig studies met hypothesen die van primair naar secundaire belang werden gedegradeerd (N = 2; 1%) en geen studies met hypothesen die van secundair naar primair belang werden bevorderd. Dit kleine aantal kan echter te maken hebben met het feit dat het categoriseren van hypothesen als primair of secundair niet zo gebruikelijk is in de psychologie in vergelijking met vakgebieden zoals de biomedische wetenschappen. In totaal omvatte 60% van de studies ten minste één hypothese in een of meer van deze categorieën, wat wijst op een aanzienlijke *bias* bij het presenteren en selecteren van hypothesen door onderzoekers en mogelijk ook door reviewers/redacteuren. In tegenstelling tot onze verwachtingen vonden we onvoldoende bewijs dat toegevoegde hypothesen

en veranderde hypothesen een grotere kans hadden om statistisch significant te zijn dan niet-selectief gerapporteerde hypothesen. Voor de andere soorten selectieve hypotheserapportage hadden we mogelijk onvoldoende statistische *power* om de relaties te detecteren. Ten slotte vonden we dat replicatiestudies minder geneigd waren om selectief gerapporteerde hypothesen te bevatten dan oorspronkelijke studies. Samenvattend kunnen we zeggen dat selectieve rapportage van hypothesen problematisch vaak voorkomt in psychologisch onderzoek.

In Hoofdstuk 3 hebben we preregistratie beoordeeld op het beperken van *researcher degrees of freedom*, de vrijheid die onderzoekers hebben om tijdens hun onderzoek beslissingen te maken. We gebruikten een uitgebreid protocol om de produceerbaarheid van preregistraties te beoordelen (d.w.z. de mate waarin de studie op de juiste wijze kan worden uitgevoerd op basis van de informatie in de preregistratie) alsmede de consistentie tussen preregistratie en publicaties van 300 preregistraties van psychologiestudies. We ontdekten dat preregistraties vaak methodologische details missen en dat afwijkingen van preregistratieplannen zelden openbaar werden gemaakt. Bijvoorbeeld, slechts 22% - 35% van de afwijkingen voor de dataverzamelingsprocedure en ongeveer 15%-20% van de afwijkingen voor de uitsluitingscriteria en het statistisch model werden openbaar gemaakt. Het combineren van de resultaten van produceerbaarheid en consistentie benadrukt dat *bias* als gevolg van *researcher degrees of freedom* mogelijk blijft in veel preregistraties. Meer uitgebreide preregistratie *templates* leverden doorgaans meer produceerbare en dus betere preregistraties op. We vonden niet dat de effectiviteit van preregistraties in de loop van de tijd veranderde noch vonden we een verschil tussen originele studies en replicatiestudies. We vonden wel dat de operationalisaties van variabelen over het algemeen effectiever werden gepreregistreerd werden dan andere onderdelen van een studie. Inconsistenties tussen preregistraties en gepubliceerde studies deden zich voornamelijk voor bij gegevensverzamelingsprocedures, statistische modellen en uitsluitingscriteria.

De resultaten in Hoofdstukken 2 en 3 benadrukken dat onderzoekers hun studies vaak niet optimaal preregistreren. Met name het vergelijken van preregistraties en de bijbehorende artikelen was uitdagend omdat onderzoekers vaak de namen van variabelen wisselden, en inconsistente notatie of dubbelzinnige taal gebruikten. Illustratief voor deze problemen is het feit dat preregistraties en artikelen vaak in verschillende formats worden geschreven; preregistraties worden vaak gestructureerd op basis van preregistratie *templates*, terwijl artikelen vaak gestructureerd worden op basis van de vereisten van wetenschappelijke tijdschriften. Hoewel preregistratie momenteel niet zijn volledige potentieel benut, zou deze werkwijze nog steeds kunnen helpen om het aantal vals-positieve resultaten in de psychologische literatuur te verminderen.

Om dat te testen vergeleken we in Hoofdstuk 4 193 psychologiestudies die een Preregistration Challenge-prijs of Preregistration Badge hadden gewonnen met 193 vergelijkbare studies die niet waren gepreregistreerd. In tegenstelling tot onze theoretische verwachtingen en eerder onderzoek (Schäfer & Schwarz, 2019; Toth et al., 2021), vonden we niet dat gepreregistreerde studies een lager percentage positieve resultaten (Hypothese 1), kleinere effectgroottes (Hypothese 2) en minder statistische fouten (Hypothese 3) hadden dan niet-geregistreerde studies. Ter ondersteuning van onze Hypotheses 4 en 5 ontdekten we wel dat geregistreerde studies vaker *poweranalyses* bevatten en doorgaans grotere steekproefgroottes hadden dan niet-geregistreerde studies. Beiden zijn geassocieerd met een hogere onderzoekskwaliteit. Ten slotte lijken zorgen over de publiceerbaarheid en impact van gepreregistreerde studies ongegrond, aangezien gepreregistreerde studies niet langer nodig hadden om gepubliceerd te worden en op verschillende metingen van wetenschappelijke impact beter scoorden. Over het algemeen geven onze data aan dat preregistratie gunstige effecten heeft op het gebied van statistische *power* en impact, maar we vonden geen robuust bewijs dat preregistratie slechte onderzoekspraktijken en daarmee *bias* voorkomt.

In Hoofdstuk 5 voerden we twee vignetstudies uit om te onderzoeken hoe psychologische onderzoekers de resultaten interpreteren van een set van vier replicatiestudies die ofwel vooraf waren gepreregistreerd of niet. Slechts een klein percentage (Studie 1: 1,6%; Studie 2: 2,2%) van de deelnemers gebruikten de correcte methode van Bayesiaanse inferentie, terwijl veel van de deelnemers simpele en inaccurate methoden gebruikten op basis van het tellen van positieve en negatieve studies. De twee studies in dit hoofdstuk toonden aan dat veel psychologische onderzoekers het bewijs voor een theorie onderschatten als een of meer resultaten van een reeks replicatiestudies statistisch significant zijn. Dit wijst op de noodzaak van betere statistisch onderwijs. In beide studies ontdekten we daarnaast dat het geloof van de deelnemers in een theorie toenam met het aantal statistisch significante resultaten en dat het resultaat van een directe replicatie een sterker effect had op het geloof in de theorie dan het resultaat van een conceptuele replicatie. In Studie 2 ontdekten we bovendien dat het geloof van de deelnemers in de theorie lager was wanneer ze uitgingen van de aanwezigheid van *p-hacking*, maar dat het geloof in de theorie niet verschilde tussen gepreregistreerde en niet-gepreregistreerde replicatiestudies.

In Hoofdstukken 6 en 7 namen we een meer praktische benadering en ontwikkelden we twee preregistratie *templates*. Deze *templates* zijn relevant gezien het resultaat in Hoofdstuk 3 dat preregistratie *templates* helpen bij het schrijven van meer produceerbare preregistraties. In Hoofdstuk 6 presenteerden we een *template* dat specifiek gericht is op secundaire data-analyses, waarbij nieuwe analyses worden uitgevoerd met bestaande gegevens. Zo'n *template* is belangrijk omdat hypothesen en analyses van

onderzoekers mogelijk *gebiased* kunnen zijn door hun voorkennis van patronen in de data. De noodzaak van een juiste begeleiding op dit gebied is groter nu data steeds vaker openbaar worden gedeeld. In dit hoofdstuk presenteerden we naast het *template* ook een handleiding voor het gebruik van het *template* met opmerkingen en een uitgewerkt voorbeeld. Het hoofdstuk geeft aan dat het invullen van dergelijk *templates* haalbaar is, *researcher degrees of freedom* kan beperken, en onderzoekers mogelijk doelbewuster maakt in hun selectie en analyse van onderzoeksdata.

In Hoofdstuk 7 presenteerden we een algemeen preregistratie *template* voor systematische reviews dat kan worden gebruikt wanneer huidige *templates* niet toereikend zijn. Het *template* is speciaal ontworpen zodat het toepasbaar is over disciplines heen (d.w.z. psychologie, economie, recht, natuurkunde of elk ander vakgebied) en over verschillende soorten reviews (d.w.z. scoping review, review van kwalitatieve studies, meta-analyses, of andere vormen van review). De beoordeelde documenten kunnen dus onderzoeksrapporten omvatten, maar ook archiefdocumenten, jurisprudentie, boeken, gedichten, enz. Items werden geselecteerd en geformuleerd om brede toepasbaarheid te optimaliseren in plaats van specificiteit. Dit PRISMA 2020-compatibele *template* kan ook gebruikt worden gebruikt als er geen gespecialiseerd *template* of registratieplatform beschikbaar is. Bij het openen van dit *template* op de Open Science Framework-website worden gebruikers daarom eerst naar gespecialiseerde *templates* geleid als deze bestaan. Naast dit gebruik kan het *template* dienen als een startpunt voor het maken van registratie *templates* die aansluiten bij specifieke vakgebieden of soorten reviews.

De studies in dit proefschrift tonen duidelijk aan dat preregistratie zijn eerste doelstelling vervult: het vergroten van transparantie. Preregistratie maakte het mogelijk voor ons om onderzoeksplannen te beoordelen en deze te vergelijken met de gerealiseerde studies, wat onmogelijk is voor een groot aantal studies in de psychologische literatuur die niet zijn gepreregistreerd en waarvan de onderzoeksplannen vanaf het begin vaag of obscuur zijn. Echter, het werd ook duidelijk in dit proefschrift dat het moeilijk is om te beoordelen of de tweede doelstelling van preregistratie wordt vervuld, het voorkomen van *bias*. Het vergelijken van preregistraties met artikelen was over de hele linie een uitdaging omdat informatie over de studies vaak onvoldoende werd verstrekt. En wanneer deze informatie wel werd verstrekt, vonden we in veel gevallen sporen van *bias*. Kortom, er is ruimte voor verbetering met betrekking tot preregistratie in de psychologie. Ontwikkelingen die kunnen helpen bij het verbeteren van preregistratie zijn een groter gebruik van *registered reports*, waarbij *peer review* plaatsvindt vóór de dataverzameling, of de ontwikkeling van geschikte preregistratie *templates*.

Acknowledgements

Foremost, I would like to thank my supervisors Marjan, Marcel, and Jelte. I don't think I could have had a better trio to support me throughout these years. Marjan, you had a knack for checking in with me at exactly the right times, and you have an incredible eye for detail that I am sure has prevented dozens of mistakes in my dissertation. Marcel, your quick and elaborate feedback was indispensable, as was your ability to put things in perspective and relieve some of the stress that comes with doing a PhD. Jelte, your knowledge of the literature and your scientific writing advice made my paper simultaneously broader in scope and more in depth. I think the three of you complemented each other perfectly, but what you all had in common was a dedication to my personal and scientific well-being. Indicative of that is that I felt invigorated to tackle the challenges in front of me after every meeting we had. Moreover, I was often in awe of the statistical skills with which you were able to effortlessly transform the verbal into the mathematical and vice versa. You were instrumental in providing my dissertation with its thorough statistical backbone.

I also want to thank all my other collaborators for their insightful feedback, continuous patience, and enduring commitment. It is a long list but here it goes: Manon Enting, Myrthe de Jonge, How Hwee Ong, Franziska Rüffer, Martijn Schoenmakers, Andrea Stoevenbelt, Charlotte Pennington, Leone Verweij, Mahmoud Elsherif, Aline Claesen, Stefan Gaillard, Siu Kit Yeung, Jan-Luca Frankenberger, Kai Krautter, Jamie Cockcroft, Katharina Kreuer, Thomas Evans, Frédérique Heppel, Sarah Schoch, Max Korbmacher, Yuki Yamada, Nihan Albayrak-Aydemir, Shilaan Alzahawi, Alexandra Sarafoglou, Maksim Sitnikov, Filip Děchtěrenko, Sophia Wingen, Sandra Grinschgl, Helena Hartmann, Suzanne Stewart, Cátia de Oliveira, Sarah Ashcroft-Jones, Bradley Baker, Tsz Keung Wong, Linda Dominguez Alvarez, Sara Weston, Lorne Campbell, Bill Chopik, Rodica Damian, Pamela Davis-Kean, Andrew Hall, Jessica Kosie, Elliott Kruse, Jerome Elson, Stuart Ritchie, Katie Valentine, Anna van 't Veer, Gjalt-Jorn Peters, Caitlin Bakker, Rickard Carlsson, Nicholas Coles, Katie Corker, Gilad Feldman, David Moreau, Thomas Nordström, Jade Pickering, Amy Riegelman, Marta Topor, Nieky van Veggel, Mark Call, David Mellor, and Nicole Pfeiffer.

Then I want to thank my former colleagues at the Meta-Research Center in Tilburg and my current colleagues at QUEST in Berlin. In particular, I want to thank my roommates Esther, Leonie, and Delwen, who have made me feel at home away from home, and my current supervisor Daniel for his support in the final stage of my writing. I also want to thank Erdem and Aiden, who neatly represent the friendships I have made at the University of Amsterdam and Tilburg University. And I want to thank Anne, Noah, and Sarahanne, and all the other people that were part of PYMS for the spirited discussions

and the warm sense of community. PYMS has undoubtedly been one of the most rewarding experiences in my academic life.

A special thanks goes to my little brothers Abel and Kalle, and my friends Anand, Auke, Edo, Frank, Jan, Joost, Jorris, Mattia, Michiel, Mike, Ruben, Pieter, Sjoerd, Stijn, and many others. You all helped me get in and out of my comfort zone, made me feel both confident and humble, and were with me in good and bad moments. No matter the time or place, I know I can count on you.

Finally, I want to thank my parents who have shown me unconditional love for more than 34 years. Your practical and emotional support were invaluable, and seeing the pride in your eyes makes me feel incredibly proud as well. I am very happy that you can be at my defense together.

Thank you everyone!