Good research practices

Sarafoglou, A.

Publication date
2023

Document Version
Final published version

Alexandra Sarafoglou

Good Research Practices

Good Research Practices

Alexandra Sarafoglou

ISBN 978-94-6421-981-4
This publication is typeset in LaTeX using the Memoir class
Printed by Ipskamp Printing B.V., Enschede
Cover by Viktor Beekman

The research of this doctoral thesis received financial assistance from the Netherlands Organisation for Scientific Research (NWO; 406.17.568).

Good Research Practices

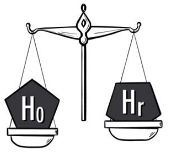
ACADEMISCH PROEFSCHRIFT
ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.-P. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel der Universiteit
op maandag 23 januari 2023, te 14.00 uur
door Alexandra Sophia Georgia Sarafoglou
geboren te Mannheim, Duitsland

Promotiecommissie

Promotor:	Prof. dr. E. M. Wagenmakers	Universiteit van Amsterdam
Copromotores:	Dr. J. M. Haaf	Universiteit van Amsterdam
	Dr. M. Marsman	Universiteit van Amsterdam
Overige leden:	Prof. dr. J. N. Rouder	University of California Irvine
	Prof. dr. B. G. Kuhlmann	University of Mannheim
	Prof. dr. D. Borsboom	Universiteit van Amsterdam
	Dr. D. Matzke	Universiteit van Amsterdam
	Dr. T. E. Hardwicke	Universiteit van Amsterdam
Faculteit:	Faculteit der Maatschappij- en Gedragswetenschappen	

Für meine Familie

Contents



Chapter 1

Introduction

1.1 A Personal Note on Popular Science

Anyone who's met me knows that I have a passion for fitness. During my PhD, I tried a variety of sports from general fitness and bodybuilding, to running, soccer, and have now arrived at Olympic weightlifting. In addition to sports, I also regularly see a nutritionist and I am part of the Weight Watchers group in Amsterdam. Accompanying my fitness journey is a plethora of books, blog posts, and social media content on training, nutrition, and self-improvement. This part of my private life is somewhat detached from my work life, and so I was surprised to discover how they intersect. That is, the longer I studied research methods and what constitutes good and bad scientific practice, the less I enjoyed the half-knowledge, Google research and baseless claims of self-proclaimed fitness experts. Over the years, I have also become painfully aware of the persistence of non-replicable psychological studies in popular science. In the following paragraphs I briefly describe some of the most striking examples from my personal experience.

My first anecdote refers to a book I read during the most work-intensive months of my doctoral studies. The book is called "Why We Sleep: Unlocking the Power of Sleep and Dreams" written by Matthew Walker (2017) and makes the following argument: adequate sleep is necessary to learn and work optimally, to avoid chronic illness, and to be emotionally balanced. In doing so, the book goes against the current grind culture that promotes that a few hours of sleep a night are sufficient and even necessary to be successful in one's full-time job, to have a rich family life, to be a top athlete, and to pick up two side-hustles at the same time.

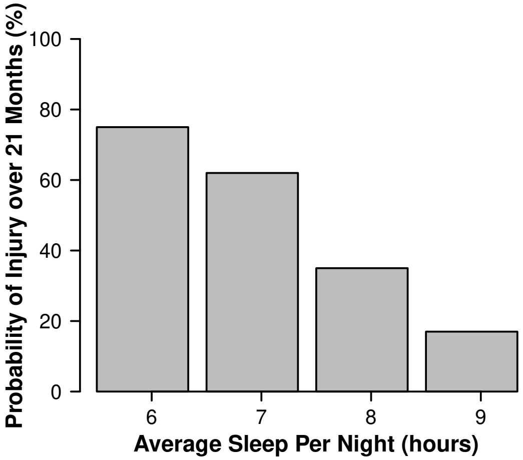


Figure 1.1: Relationship between sleep and injury risk for athletes as presented in Walker's book "Why We Sleep: Unlocking the Power of Sleep and Dreams". Data extracted from Figure 1 of the original manuscript.

The author stresses that "Why We Sleep" is scientifically accurate and should not be taken as popular science literature. Unfortunately, however, the book does

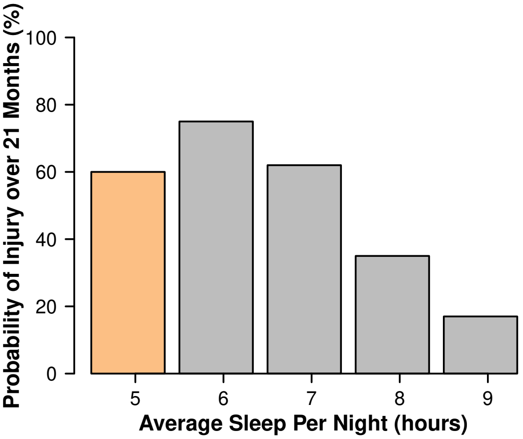


Figure 1.2: Relationship between sleep and injury risk for athletes as presented in the original article by Milewski et al. (2014). This graph features an additional column which is at odds with Walkers hypothesis: athletes who slept on average five hours, reported fewer injuries than athletes who slept six and seven hours.

not live up to this claim. In his sixth chapter, Walker (2017) argues that longer sleep duration increases athletic performance and lowers the risk of injury. To support his claim, Walker presents a graph showing the injury risk of athletes who sleep between six and nine hours a day. The result appears compelling: the less the athletes slept, the greater the likelihood of injury (see Figure 1.1). On the Joe Rogan podcast, Walker comments on this result: "the most surprising factor was injury risk: when they looked at athletes across the season, and they just plotted how frequently will they get injured and then they surveyed them 'how much sleep are you getting?' and they bucketed them into people getting nine hours, seven hours, six, five, four: it's a perfect linear relationship: The less sleep that you have, highered your injury risk" (36:40; Rogan, 2018).

When I checked the original publication cited by Walker, I learned that the authors (Milewski et al., 2014) did not actually test whether the risk of injury decreases the more athletes sleep. Instead, Milewski et al. (2014) first combined data from all athletes who slept less than eight hours and data from all athletes who slept eight hours or more; next they examined whether these two groups differed in their risk of injury. In other words, the analysis by Milewski et al. (2014) does not directly address the pertinent question involving a monotonic relation between sleep and risk of injury; all that follows from the analysis is that sleeping less than eight hours is statistically associated with a higher risk of injury than sleeping eight hours or more. The staircase pattern in Figure 1.1 does suggest a monotonic trend, but this was not statistically tested. Note that the original study does not provide error bars, nor information about how many athletes are in each bin.

More problematic, however, is that Walker draws this conclusion from data he had tinkered with (for a discussion, see Gelman, 2020 and Guzey, 2019). The findings from the original study by Milewski et al. (2014) are shown in Figure 1.2. Milewski et al. (2014) not only reported the injury risk of athletes with an average of six hours of sleep or more, but also with five hours of sleep. But instead of getting injured most often, the data shows that athletes who sleep five hours got injured *less often* than athletes who slept six or seven hours. Confronted with this inconvenient finding, Walker decided to simply crop the graph for his book. This example highlights the problem of non-transparent research reporting, but also the problem of testing hypotheses with inappropriate statistical methods.

The second anecdote relates to my effort to improve my relationship with food. To do so, I signed up to the Amsterdam Weight Watchers meetings at the beginning of my doctoral studies. Over the years, my coaches and the other group members taught me a lot about healthy eating, habit formation, and self-efficacy. However, I came to realize that a considerable amount of their information is –to put it mildly– no longer up to date. Every few months or so, when my Weight Watchers coach discussed tips and tricks to curb cravings, their catalog of recommendations seemed to be based exclusively on Brian Wansink's work. These included suggestions such as "use smaller plates to trick yourself into eating less" or "never go grocery shopping while hungry". Brian Wansink was what people would consider a *rockstar scientist*: his papers have been cited almost 38,000 times, he was involved in various food related programs administered by the United States Department of Agriculture (USDA), and his research was featured regularly in outlets such as *The New York Times* and *The Washington Post*.

That all changed in 2016, when Wansink published a blog post which led to his downfall. Originally, the blog post was meant to encourage aspiring researchers to seize opportunities. Instead, the blog post demonstrates a prime example of what constitutes shabby research practice. In his blog post, Wansink (2016) described two members of his lab: a *lazy* postdoctoral fellow and a *dedicated* PhD student. The *lazy* postdoc declined to continue working on a "failed study" which showed null results. The *dedicated* PhD candidate, on the other hand, spent her time more wisely. Encouraged by Wansink, she went on "deep data dives" with the goal of perhaps eliciting statistically significant results from the data after all. Eventually her fishing expeditions were rewarded: the PhD candidate was able to obtain the desired significant results leading to 5 publications. By contrast, the *lazy* postdoc left Wansink's lab after a year (and subsequently left academia) with only a quarter as many publications as the *dedicated* PhD student. This blog post, especially the encouragement to "massage" data to find interesting patterns (i.e., fudging) and then reverse-engineer matching hypotheses (i.e., HARKing; Kerr, 1998, see Figure 1.3), set off alarm bells among methodologists worldwide, prompting some of them to scrutinize Wansink's published articles more closely.

In subsequent years, a thorough examination of Wansink's articles revealed –among other things– conclusions that were not supported by the data, an extraordinary amount of misreporting (e.g., the so-called 'Pizzagate' affair, in which independent analysts discovered 150 discrepancies in 4 of his papers; van der Zee, Anaya, & Brown, 2017), incorrect statistical analyses, improper archiving, and refusal to share the original data. The investigations eventually led to the retraction of 18 Wansink articles (one article was retracted twice), with 15 other

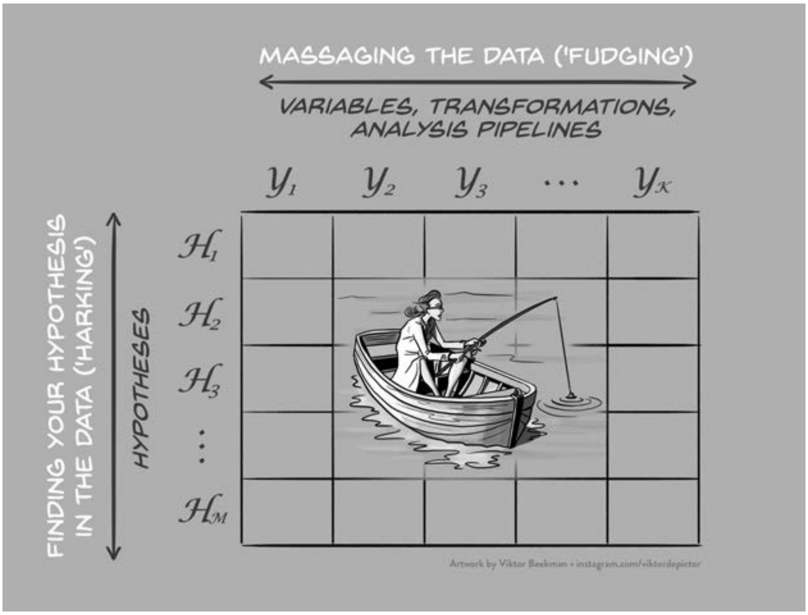

MASSAGING THE DATA ('FUDGING')
VARIABLES, TRANSFORMATIONS, ANALYSIS PIPELINES
y_1 y_2 y_3 ... y_K
FINDING YOUR HYPOTHESIS IN THE DATA ('HARKING')
HYPOTHESES
$\mathcal{H}_1$
$\mathcal{H}_2$
$\mathcal{H}_3$
...
$\mathcal{H}_M$
Artwork by Viktor Beekman • instagram.com/viktordepictor

Figure 1.3: A researcher fishing for significant results by either abusing different analysis pipelines until the result is significant (fudging) or by adapting the hypothesis to the data at hand (HARKing). Figure available at `https://www.bayesianspectacles.org/library/` under CC license `https://creativecommons.org/licenses/by/2.0/`.

articles having to be corrected, and several others receiving an expression of concern (Retraction Watch, 2022). In the meantime, Wansink had to step down from his position. Nevertheless, Wansink's work is still cited today and, in the case of Weight Watchers, is considered sound science that carries practical ramifications.

1.1.1 The Empirical Cycle

The reason "deep data dives" are considered bad research practice is that the reliability of empirical science rests on there being a sharp distinction between the *creative context of discovery* and the *statistical context of justification* (Reichenbach, 1938). This distinction prohibits researchers from "using the same data twice", that is, using them first to formulate a hypothesis and then again to test that hypothesis. Ideally, empirical research adheres to the empirical cycle (De Groot, 1956/2014) illustrated in Figure 1.4.

Researchers start from existing knowledge and data based on previous studies and published literature. The researcher can now speculate and explore in order to derive new hypotheses and concrete predictions. With the predictions at hand, the researcher then designs a new experiment, collects data, and executes a statistical analysis to test these predictions. By evaluating the results, the researcher determines whether or not the hypotheses receive support from the data, and the resulting conclusion is then in turn part of the scientific knowledge accumulation.

The problem arises when researchers do not adhere to the empirical cycle but decide to take a shortcut. That is, instead of testing new predictions on new data, the predictions are tested on the same data that inspired them. Here researchers are fooling themselves. Instead of making a scientific discovery, they are more likely to be interpreting a pattern that emerged from chance thus increasing the chance of misleading research claims (Simmons, Nelson, & Simonsohn, 2011). As Richard Feynman famously said: “The first principle is that you must not fool yourself–and you are the easiest person to fool”.

Researchers want to discover the truth. However, they also want to present convincing results that leave no room for doubt and publish papers that make interesting claims. Add to this the fact that although the researchers themselves have the greatest incentive to produce clean and significant effects, they are typically the ones who carry out all research steps themselves, from data collection to conducting the analysis to writing the manuscript. In psychology, we have reaped the fruits of these perverse incentives and researchers' reluctance to acknowledge uncertainty. Systematic, high-powered replication studies conducted over the past decade have shown that the field is in a “crisis of confidence” (Pashler & Wagenmakers, 2012) as, for the most part, only a disappointingly small percentage of studies can be replicated (Camerer et al., 2018; R. Klein, Ratliff, Vianello, Adams, et al., 2014; R. Klein et al., 2018; Open Science Collaboration, 2015).

1.1.2 Combating the Crisis of Confidence

I personally rate the crisis of confidence in psychology as luck in disguise (cf. Spellman, 2015 and Vazire, 2018). In fact, in our field the crisis cleared the way for ongoing extensive methodological reforms (Chambers, 2013, 2017; Kidwell et al., 2016; MacCoun & Perlmutter, 2015; Nosek et al., 2015, see e.g.,). The goal of these reforms is to improve research practices by increasing transparency, openness, and to establish good research practices that ensure researchers adhere to the empirical cycle.

Some of these reforms are already well established in the field. For instance, one reform which has quickly gained popularity is *preregistration* (Munafò et al., 2017; Nosek & Lindsay, 2018; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Through preregistration, researchers tie their hands to prevent (conscious or subconscious) significance chasing by detailing their research design, sample size, and analysis plan before they have collected the data. This prevents changes being made during data collection (e.g., more data being collected when results were not significant with the original number). When the data are available, the researchers execute the preregistered analysis, eliminating the confusion between hypothesis–generating and hypothesis–testing.

But despite its popularity, the method has also been criticized (e.g., Devezer, Navarro, Vandekerckhove, & Buzbas, 2020; Fiedler, 2018; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Szollosi et al., 2020). For instance, researchers are reluctant to adopt preregistration due to its lack of flexibility. Indeed, if one commits to a precise analysis plan, there is the risk that the planned analyses will not work on the data that is eventually collected. A solution to this problem may be *analysis blinding*, a method commonly applied in physics

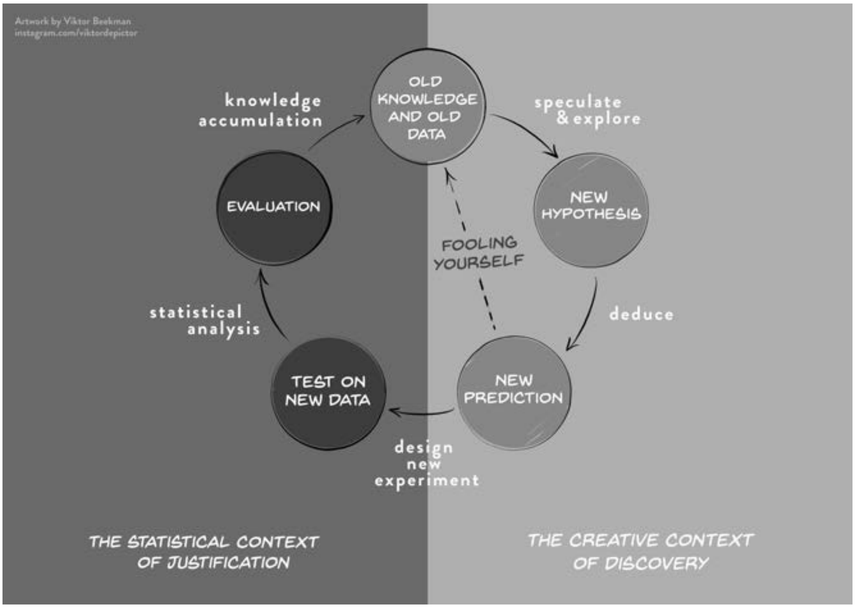


Figure 1.4: Empirical cycle and the distinction between the context of discovery and the context of justification. Researchers fool themselves when they test new predictions against old knowledge and data, setting up a feedback loop between analysis decisions and their results. Figure available at `https://www.bayesianspectacles.org/library/` under CC license `https://creativecommons.org/licenses/by/2.0/`.

but underappreciated in psychology and the social sciences (Dutilh, Sarafoglou, & Wagenmakers, 2019; MacCoun, 2020; MacCoun & Perlmutter, 2015; MacCoun & Perlmutter, 2018). With analysis blinding, researchers do not determine their analysis plan in detail before seeing the data. Instead, they develop their analysis strategy on "blinded" data for which a collaborator or independent researcher has removed all potentially biasing information (e.g., condition assignment). As with preregistration, analysis blinding breaks the feedback loop between analysis decisions and their results. An additional advantage is that analysis blinding retains the flexibility to account for unexpected peculiarities in the data.

Another promising development concerns the attitude towards hidden uncertainty in data analysis. There is an increasing shift towards acknowledging that there are multiple alternative statistical perspectives on data. That is, researchers might draw different conclusions even when answering the same research question based on the same data set. Following this shift, a number of multi-analyst studies have explored this uncertainty in more detail.

In concert with methodological reforms, increased attention has been paid to statistical innovations (e.g., Benjamin et al., 2018; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016). For instance, within Bayesian statistical framework, powerful methods have been proposed to help researchers formulate and test hypotheses about trends and patterns of effects (so-called ordinal hypotheses). These methods could be used, for instance, to test Walker's hypothesis that the risk of injury decreases the more you sleep.[1] In terms of addressing the crisis of confidence, the ability to test specific ordinal hypotheses is particularly important in replication research, in which researchers seek to confirm whether data from a replication study show the same trend as the original study. Furthermore, testing ordinal hypotheses ties in with the empirical cycle in two ways. By appropriately quantifying the predictions, ordinal hypotheses facilitate the derivation of concrete predictions based on theory (Haaf, Klaassen, & Rouder, 2019), and the associated statistical tests further improve the step from testing the data to evaluating the results.

The crisis of confidence has even spawned a new field of research called "meta-science" that conducts research on research. It validates reform ideas, develops new methods to improve science, and analyzes the acceptance of existing methods among researchers. My work is located at the intersection between meta-science and statistical research methods. Thus, the goal of my dissertation was to study interesting reform ideas but also to develop statistical methods that support researchers in their work. My efforts can be thematically divided into three parts, which will be summarized in the next section.

1.2 Chapter Outline

1.2.1 Part I: Revealing Hidden Uncertainty in Data Analysis

The first part of the dissertation examines current good research practices in psychological science. This part begins with a sobering realization: you do not need to be an expert to be able to predict whether a social science study replicates–Chapter 2 illustrates that laypeople too are able to predict replication success with above-chance performance. We suggest that laypeoples' predictions may be used to quantify intuitive plausibility of empirical effects and hence contribute to efficiently selecting studies for replication research.

Chapter 3 introduces the Many-Analysts Religion Project. In this project, we recruited 120 analysis teams to investigate (1) whether religious people self-report higher well-being, and (2) whether the relation between religiosity and self-reported well-being depends on perceived cultural norms of religion (i.e., whether it is considered normal and desirable to be religious in a given country). For the first research question, all but 3 teams reported positive effect sizes with credible/confidence intervals excluding zero. For the second research question, this was the case for 65% of the teams.

Chapter 4 contains our reflections and conclusions about the Many-Analysts Religion Project. We address the issue of theoretical specificity, highlight some more in-depth observations, discuss some methodological concerns, and reflect on our experience of organizing a many-analysts project.

Chapter 5 describes a survey study identifying the benefits and challenges of preregistration from the researcher's perspective. The study showed that prereg-

[1] I was curious how much evidence the complete and cropped data provided for Walker's hypothesis but unfortunately I was unable to perform an analysis since Milewski et al. (2014) did not publicly archive their data.

istration has benefits beyond safeguarding the adherence to the empirical cycle, including the improvement of the overall project quality. This survey, however, also illustrates some of the challenges that come with preregistration, such as the increase of the overall project duration and work related stress.

Chapter 6 introduces analysis blinding as an addition or possible alternative to preregistration. The chapter discusses how analysis blinding can be applied in experimental psychology. Specifically, it introduces different methods of analysis blinding, offers recommendations for blinding of popular experimental designs, and introduces the design for an online blinding protocol.

Following this idea, Chapter 7 compares the reported efficiency and convenience of preregistration and analysis blinding in the context of the Many-Analysts Religion Project. The recruited teams answered the same research questions based on the same data either preregistering their analysis or using analysis blinding. The study concludes that analysis blinding does not mean less work but approximately the same amount, but researchers can still benefit from the method since they can plan more appropriate analyses from which they deviate less frequently.

1.2.2 Part II: Multinomial Order-Restrictions

The second part of the dissertation discusses how theory-based knowledge can be quantified in statistical models and introduces statistical techniques to test ordinal hypotheses in the context of categorical data analysis.

Chapter 8 describes a Bayesian technique with which researchers can evaluate ordinal hypotheses concerning the distribution of multinomial proportions. Whenever researchers formulate ordinal hypotheses that entail expectations about increasing or decreasing trends they must rely on methods that are relatively inefficient and computationally expensive. To address this problem, we developed a bridge sampling routine that allows an efficient evaluation of these hypotheses for multinomial variables. An empirical example shows that bridge sampling outperforms current Bayesian methods in terms of accuracy and efficiency.

In order to maximize the accessibility of the proposed bridge sampling routine, we developed the user-friendly `R` package **multibridge** which is introduced in Chapter 9. The package implements the bridge sampling routine for multinomial variables and independent binomial variables. The chapter describes the core functions in **multibridge** and illustrates its use with two examples, one of which concerning the prevalence of statistical reporting errors across eight different psychology journals.

Chapter 10 applies the evaluation of ordinal hypotheses in the context of multinomial processing tree (MPT) models. In psychology, MPT models are used to test sophisticated theories on memory, judgement and decision making, and reasoning. The chapter highlights how researchers can refine their Bayesian MPT modeling practices by adequately capturing their theory in the model and testing their ordinal expectations.

1.2.3 Part III: Guidelines for Good Research Practices

The third part of the dissertation provides concrete suggestions on how to facilitate the uptake of good research practice among researchers. This part addresses this challenge on two levels: educating researchers and training students.

Chapter 11 presents the Transparency Checklist which allows researchers in social and behavioural sciences to improve and document the transparency of research reports. The initial set of items in the Transparency Checklist was evaluated by 45 behavioural and social science journal editors-in-chief and associate editors, as well as 18 open-science advocates. The final checklist spans the four study components: preregistration, methods, results and discussion as well as data, code and materials availability. Responses to the checklist items can be submitted along with a manuscript, providing reviewers, editors and, eventually, readers with critical information about the research process necessary to evaluate the robustness of a finding.

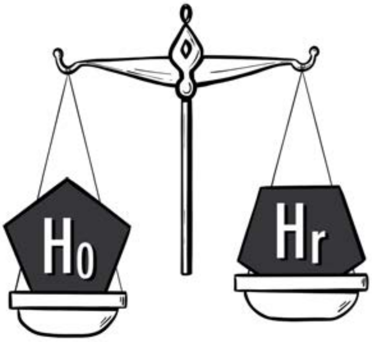
Chapter 12 discusses seven concrete statistical practices which embody the current aspirations in the social and behavioural sciences to increase transparency and reproducibility. These practices are (1) visualizing data; (2) quantifying inferential uncertainty; (3) assessing data preprocessing choices; (4) reporting multiple models; (5) involving multiple analysts; (6) interpreting results modestly; and (7) sharing data and code. We discuss the benefits and limitations of each practice and provide guidelines for its adoption.

The remaining two chapters show how the concepts of good research practices can be incorporated into the methodological training of students. Chapter 13 describes the content of the graduate course "Good Research Practices" which we have designed and taught at the University of Amsterdam. This course gives a general introduction into the crisis of confidence as well as recent methodological reforms proposed in psychological science, such as direct and conceptual replication studies, preregistration, and the public sharing of data, code, and analysis plans.

Chapter 14 presents a Bayesian research project that we conducted with undergraduate psychology students. This project aimed to (1) convey the basic mathematical concepts of Bayesian inference; (2) have students experience the entire empirical cycle including collection, analysis, and interpretation of data and (3) teach both the philosophy behind good research practices and the practical skills needed to apply them.

Part I

Revealing Hidden Uncertainty in Data Analysis



Chapter 2

Laypeople Can Predict Which Social-Science Studies Will Be Replicated Successfully

Abstract

Large-scale collaborative projects recently demonstrated that several key findings from the social science literature could not be replicated successfully. Here we assess the extent to which a finding's replication success relates to its intuitive plausibility. Each of 27 high-profile social science findings was evaluated by 233 people without a PhD in psychology. Results showed that these laypeople predicted replication success with above-chance performance (i.e., 58%). In addition, when laypeople were informed about the strength of evidence from the original studies, this boosted their prediction performance to 67%. We discuss the prediction patterns and apply signal detection theory to disentangle detection ability from response bias. Our study suggests that laypeople's predictions contain useful information for assessing the probability that a given finding will replicate successfully.

This chapter is published as Hoogeveen, S., Sarafoglou, A., & Wagenmakers, E. J. (2020). Laypeople can predict which social-science studies will be replicated successfully. *Advances in Methods and Practices in Psychological Science*, *3*, 267–285.

2.1 Introduction

Recent work has suggested that the replicability of social science research may be disturbingly low (Baker, 2016). For instance, several systematic high-powered replication projects demonstrated successful replication rates ranging from 36% (Open Science Collaboration, 2015), 50% (R. Klein et al., 2018), 62% (Camerer et al., 2018) to 85% (R. Klein, Ratliff, Vianello, Adams, et al., 2014). These low replication rates have been explained by several factors that operate at different levels. At the level of the scientific field as a whole, problems include publication bias (Francis, 2013) and perverse incentive structures (Giner-Sorolla, 2012). At the level of individual studies, problems concern low statistical power (Button et al., 2013; Ioannidis, 2005) and questionable research practices such as data-driven flexibility in statistical analysis (i.e., significance seeking; John, Loewenstein, & Prelec, 2012; Simmons et al., 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Here we focus on yet another problem that has recently been associated with poor replicability: the *a priori* implausibility of the research hypothesis (Benjamin et al., 2018; Ioannidis, 2005).

If the a priori implausibility of the research hypothesis is indicative of replication success, then replication outcomes can be reliably predicted based only on a brief description of the hypothesis at hand. Indeed, results from recent surveys and prediction markets demonstrated that researchers (i.e., experts) in psychology and related social sciences can anticipate replication outcomes with above-chance accuracy – as a group, experts correctly predicted the replication outcomes for 58%, 67%, and 86% of the studies included in the Reproducibility Project: Psychology, the Many Labs 2 project, and the Social Science Replication project, respectively (Camerer et al., 2018; Dreber et al., 2015; Forsell et al., 2018). These surveys and prediction markets involved forecasters with a PhD in the social sciences (e.g., psychology, economics). In addition, the forecasters had been provided with statistical information concerning the effect size in the original study, including p-values, effect sizes, and/or sample sizes. This raises two key questions about anticipated replicability: First, do forecasters need to be social science experts to predict replication outcomes with above-chance accuracy? Second, are forecasters' predictions driven by intuitions about empirical plausibility alone or also influenced by statistical information about the original effect?

In this study, our primary aim was to investigate whether and to what extent accurate predictions of replicability can be generated by people without a professional background in the social sciences (i.e., laypeople; people without a PhD degree in psychology) and without access to the statistical evidence obtained in the original study. Laypeople may be able to produce reliable evaluations of plausibility (and hence replicability) of research hypotheses, even without access to relevant statistical information or in-depth knowledge of the literature – after all, social science concerns itself with constructs that are often accessible and interesting to a lay audience (Milkman & Berger, 2014). Consequently, when presented with a non-technical description of a study's topic, operationalization and result, laypeople may well be able to produce accurate replicability forecasts. For example, consider a non-technical description of the research hypothesis by Kidd and Castano (2013):

> "Can reading literary fiction improve people's understanding of other people's emotions? Participants read a short text passage. In one group, the text passage was literary fiction. In the other group, the text passage was non-fiction. Afterwards, participants had to identify people's expressed emotion (e.g., happy, angry) based on images of the eyes only. Participants were better at correctly recognizing the emotion after reading literary fiction."

A general understanding of the concepts (e.g., literary fiction, emotions) and proposed relation between those concepts (e.g., reading literary fiction improves emotion recognition) may suffice to form intuitions about plausibility that match the (eventual) empirical evidence. The accuracy of such intuitions can be gauged by comparing laypeople's prediction against the empirical outcome – hence, for this study, we selected 27 high-profile findings that have recently been submitted to high-powered replication attempts (Camerer et al., 2018; R. Klein et al., 2018).

If laypeople can indeed make accurate predictions about replicability, these predictions may supplement theoretical considerations concerning the selection of candidate studies for replication projects. Given limited resources, laypeople's predictions concerning replicability could be used to define the subset of studies for which one can expect to learn the most from the data. In other words, researchers could use laypeople's predictions as input to assess information gain in a quantitative decision-making framework for replication (Hardwicke, Tessler, Peloquin, & Frank, 2018; MacKay, 1992). This framework follows the intuition that –for original studies with surprising effects (i.e., low plausibility) or small sample sizes (i.e., little evidence)– replications can bring about considerable informational gain (R. Klein, Ratliff, Vianello, Adams Jr, et al., 2014).

More generally, if even laypeople can to a large extent correctly pick out the unreplicable findings, this suggests that researchers should be cautious when conducting and eventually publishing studies with risky and counterintuitive hypotheses. Laypeople's adequate predictions of replicability may thus provide empirical support for a culture change that emphasizes robustness and 'truth' over novelty and 'sexiness' (Dovidio, 2016; Giner-Sorolla, 2012; Nosek, Spies, & Motyl, 2012). When extended to novel hypotheses, laypeople's skepticism may even serve as a 'red flag', prompting researchers to go the extra mile to convince their audience –laypeople and peers alike– of the plausibility of that particular research claim (e.g., by using larger samples, engaging in Registered Reports, setting a higher bar for evidence; see Benjamin et al., 2018; Chambers, 2013).

The secondary aim of the current study was to assess the extent to which the inclusion of information about the strength of the evidence obtained in the original study improves laypeople's prediction performance. In contrast to the expert prediction surveys by Camerer et al. (2018) and Forsell et al. (2018), we used Bayes factors rather than p-values and effect sizes to quantify the evidence in the original studies (Jeffreys, 1961; Kass & Raftery, 1995).

We preregistered the following expectations and hypotheses: First, we expected that, based on an assessment of the a priori plausibility of the research hypotheses at hand, (1a) laypeople can predict the replicability of empirical studies with above-chance accuracy, and (1b) laypeople's confidence is associated with the magnitude of the effects of interest in the replication study. The former would be reflected in a prediction accuracy rate above 50% and the latter in a positive correlation between people's confidence in replicability and the replication effect size. In addition, we hypothesized that (2) the inclusion of information on the strength of the original evidence (i.e., the Bayes factor) would improve prediction performance.

2.2 Disclosures

2.2.1 Data, materials, and preregistration

The current study was preregistered on the Open Science Framework by means of a time-stamped PDF; readers can access the preregistration, as well as all materials, reanalyses of the original studies, the anonymized raw and processed data (including relevant documentation for the data of ML2 and SSRP), and the R code to conduct all confirmatory and exploratory analyses (including all figures), in our OSF folder at: `https://osf.io/x72cy/`. Any deviations from the preregistration are mentioned in this chapter.

2.2.2 Supplemental Material

In the online Supplemental Material (`http://journals.sagepub.com/doi/10.1177/2515245920919667`) we provide additional details on the methods and additional exploratory analyses. Specifically, the supplemental material presents details on the Bayesian reanalyses of the original studies, the sampling plan, and the statistical models and prior specifications; includes tables with the descriptions (in English and Dutch) of all the original studies as presented to the participants; and reports two additional exploratory analyses. The first of these analyses concerns the accuracy of predictions derived from the Bayes factors alone, without human evaluation, and the second analysis is a Bayesian logistic regression model that includes random effects for both participants and studies.

2.2.3 Reporting

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

2.2.4 Ethical approval

The study was approved by the local ethics board of the University of Amsterdam and all participants were treated in accordance with the Declaration of Helsinki.

2.3 Methods

2.3.1 Participants

In total we obtained data from 257 participants, who were recruited from the online platform Amazon Mechanical Turk (n = 83), the online subject pool of first-year psychology students from the University of Amsterdam (n = 138), and social media platforms such as Facebook (n = 36). Participants from MTurk received a financial compensation for participation, first-year students from the University of Amsterdam received research credits, and participants from social media were given the opportunity to enter a raffle for a voucher from a Dutch webshop. After exclusions (see below), the final sample consisted of 233 participants, with 123 participants in the Description Only condition and 110 participants in the Description Plus Evidence condition.

2.3.2 Sampling Plan

Based on our sampling plan, we determined the minimum number of 103 observations per group to obtain strong evidence (i.e., a Bayes factor > 10) in favor of our hypothesis with a probability of 80%, assuming a medium effect size of $\delta = 0.5$, a default prior, and a study design that compares two independent groups (i.e., a t-test). As preregistered, data collection continued after the minimum number of participants was reached (i.e., 103 in each condition), until the pre-established data collection termination date of April 22nd, 2019.

2.3.3 Materials

Participants were presented with 27 studies, a subset of the studies included in the Social Sciences Replication Project (SSRP; Camerer et al., 2018) and the Many Labs 2 Project (ML2; R. Klein et al., 2018).

2.3.3.1 Study Selection Process

In the Description Plus Evidence condition, participants were provided with study descriptions accompanied by information on the strength of the evidence provided by the original study in the form of a Bayes factor. Therefore, one of the main criteria when selecting the studies was that the original analysis allowed for a Bayesian reanalysis using the Summary Stats module in JASP (JASP Team, 2021), that is, the main analysis should be conducted using a paired samples or independent samples t-test, a correlation test, or a binomial test.[1] Details about the reanalyses are provided in the Supplemental Material. We subsequently checked whether the proportion of successful vs. unsuccessful replications was similar to the proportions in the individual projects (i.e., 50% and 62%). This was the case; our subset included 14 successful and 13 unsuccessful replications (52%).

[1] For some studies, the original articles reported F-values derived from ANOVA designs, but as the crucial comparison was between only two groups, we converted the respective F-value to a t-value, which was then entered in the Summary Stats module in JASP.

2.3.3.2 Presentation of Studies

For each study, participants read a short description of the research question, its operationalization, and the key finding. The descriptions were inspired by those provided in SSRP and ML2, but rephrased to make them comprehensible for laypeople. In the Description Only condition, solely the descriptive texts were provided; in the Description Plus Evidence condition, the Bayes factor and its verbal interpretation (e.g., "moderate evidence") for the original study were added to the descriptions. The verbal interpretations were based on a classification scheme proposed by Jeffreys (1939) and adjusted by Lee and Wagenmakers (2013, p. 105). These verbal labels were added to assist the interpretation of the Bayes factors, since the concept of evidence ratios might be difficult or ambiguous for laypeople (Etz, Bartlema, Vanpaemel, Wagenmakers, & Morey, 2019). To prevent participants from reading up the replication outcomes of the original studies during the survey itself, we ensured that the descriptions did not contain identifying information, such as the names of the authors, the study titles, or any direct quotes. In addition to the 27 study descriptions, participants were also presented with one bogus item as an attention check. In the description of this item participants were instructed to answer "No" to the question whether the study will replicate and indicate a confidence of 75%. Participants from the Netherlands could choose to read the study descriptions in English or Dutch. The translation of the English study descriptions into Dutch were assisted by the online translation software DeepL (TechCrunch, 2019).

2.3.4 Procedure

The survey was generated using the online survey software Qualtrics (Qualtrics, 2021). Participants were randomly assigned to the Description Only or the Description Plus Evidence condition. First, participants read an explanation of the term 'replication' and its relevance in science: "*You will be asked whether you think that the described study will replicate. This means: if an independent lab will do this study again with a large number of participants, using the same materials, will they find convincing evidence for the same effect? If the effect really exists, it should be found by a different lab. However, it seems that not all studies can be replicated, because some results are based on coincidence, or poorly designed or executed studies.*" Participants in the Description Plus Evidence condition additionally received a short explanatory text of the Bayes factor, including the commonly used verbal interpretation categories for the strength of evidence (Lee & Wagenmakers, 2013, p. 105). The explanation of the Bayes factor was: "*A Bayes factor (BF) is the degree to which evidence is found for the existence of the effect, based on the data at hand. For instance, if BF = 2, the data suggest that it is 2 times more likely that the effect is present, than that there is no effect.*"[2]

[2] Unfortunately, this explanation fell prey to a prevalent misinterpretation of Bayes' rule (e.g., Wagenmakers, Etz, Gronau, & Dablander, 2018); the example describes the posterior odds (i.e., $\frac{p(\mathcal{H}_1|\text{data})}{p(\mathcal{H}_0|\text{data})}$) rather than the Bayes factor (i.e., $\frac{p(\text{data}|\mathcal{H}_1)}{p(\text{data}|\mathcal{H}_0)}$). When prior odds are assumed to be equal for the alternative and the null hypothesis –as is often assumed (e.g., Jeffreys, 1961)– the posterior odds equal the Bayes factor.

After the instructions, participants were presented with the 27 studies plus the bogus attention check study. Each study was presented and rated on a separate page. After reading the study description (and the Bayes factor plus verbal interpretation in one condition), participants could select a tick box to indicate that they did not understand that particular study description. Subsequently, they indicated whether they believed that this study would replicate or not (yes / no), and expressed their confidence in their decision on a slider ranging from 0 to 100. The order in which the studies were presented was randomized across participants.[3] Finally, at the end of the survey, participants were asked whether they were already familiar with the Many Labs 2 project and/or the Social Science Replication project.

2.3.5 Data Exclusions

As stated in our preregistration, we excluded participants (1) if they had a PhD in psychology (i.e., they qualified as experts rather than laypeople); (2) if they indicated that they did not understand more than 50% of the descriptions; (3) if they did not read the descriptions carefully (i.e., they failed the included attention check); or (4) if they were already familiar with the replication projects by Camerer et al. (2018) and/or R. Klein et al. (2018).

The current study applied a more stringent definition of experts than previous prediction survey studies (i.e., Camerer et al., 2018; Dreber et al., 2015; Forsell et al., 2018); whereas previous surveys defined ‘experts’ as researchers in psychology, ranging from graduate students to full professors, the current study defined experts as people with a PhD degree in psychology and hence classified graduate students as laypeople.[4] Participants who indicated to have a PhD in psychology were immediately redirected to the end of the survey and could not complete the actual study. As specified in our preregistration, participants passed the attention check if they answered as explicitly instructed: selecting “No” for the dichotomous replication question, and rating confidence in the interval between 70% and 80%. We excluded 3 participants because they indicated that they were familiar with the replication projects, and 22 participants because they failed the attention check. No participants indicated that they understood less than 50% of the study descriptions. In total, we excluded 1.6% (i.e., 99) of all predictions based on participants indicating that they did not understand the study description. 72% of participants (i.e., 167) understood all study descriptions.

2.3.6 Statistical Models

We constructed Bayesian (hierarchical) models to estimate and test the parameters of interest for each hypothesis. For all analyses the outcome measures were

[3] Due to a programming error, the study descriptions were not randomized for the $n = 12$ participants who were recruited from social media and selected to take the survey in Dutch.

[4] This discrepancy had no discernible influence on our conclusions; subsequent exploratory analyses suggested that the results did not change when excluding participants who were recruited via Amazon Mechanical Turk or social media platforms and who reported having studied psychology (at any level).

chosen based on what was most relevant and informative for answering the respective research questions. For the primary analysis we estimated accuracy rates $[0-1]$ as these afford the most intuitive and simple interpretation and are directly comparable with previous prediction survey studies. The experimental effect of Description Only vs. Description Plus Evidence was evaluated by means of Brier scores, because here the unit of interest was the individual prediction performance, which takes into account accuracy and confidence and is the most 'sensitive' measure for comparing people's performance across conditions. In the correlation analysis, the units of interest were the studies rather than participants, hence here we looked at the confidence ratings per study (aggregated across participants). All models and priors are described in detail in the Supplemental Material.

2.4 Results

2.4.1 Descriptive Pattern

Figure 2.1 displays participants' confidence ratings concerning the replicability of each of the 27 included studies, ordered according to the averaged confidence score. Positive ratings reflect confidence in replicability, and negative ratings reflect confidence in non-replicability, with -100 denoting extreme confidence that the effect would fail to replicate. Note that these data are aggregated across the Description Only and the Description Plus Evidence condition. The top ten rows indicate studies for which laypeople showed relatively high agreement that the associated studies would replicate. Out of these ten studies, nine replicated and only one did not (i.e., the study by C. Anderson, Kraus, Galinsky, & Keltner, 2012; note that light-grey indicates a successful replication, and dark-grey indicates a failed replication). The bottom four rows indicate studies for which laypeople showed relatively high agreement that the associated studies would fail to replicate. Consistent with laypeople's predictions, none of these four studies replicated. For the remaining 13 studies in the middle rows, the group response was relatively ambiguous, as reflected by a bimodal density that is roughly equally distributed between the negative and positive end of the scale. Out of these 13 studies, five replicated successfully and eight failed to replicate successfully. Overall, Figure 2.1 provides a compelling demonstration that laypeople are able to predict whether or not high-profile social science findings will replicate successfully. In Figure 2.2 and Figure 2.3 Laypeople's predictions are separately displayed for the Description Only and the Description Plus Evidence condition, respectively.

Figure 2.4 provides a more detailed account of the data for three selected studies. For the study in the top panel (i.e., Gneezy, Keenan, & Gneezy, 2014), most laypeople correctly predicted that the effect would successfully replicate; for the study in the middle panel (i.e., Tversky & Gati, 1978), laypeople showed considerable disagreement, with slightly over half of the participants incorrectly predicting that the study would replicate successfully; finally, for the study in the bottom panel (i.e., Shah, Mullainathan, & Shafir, 2012), most laypeople correctly predicted that the effect would fail to replicate.

Before conducting our preregistered confirmatory analyses, we first exploratorily investigated the relation between the Bayes factors of the *original studies* and

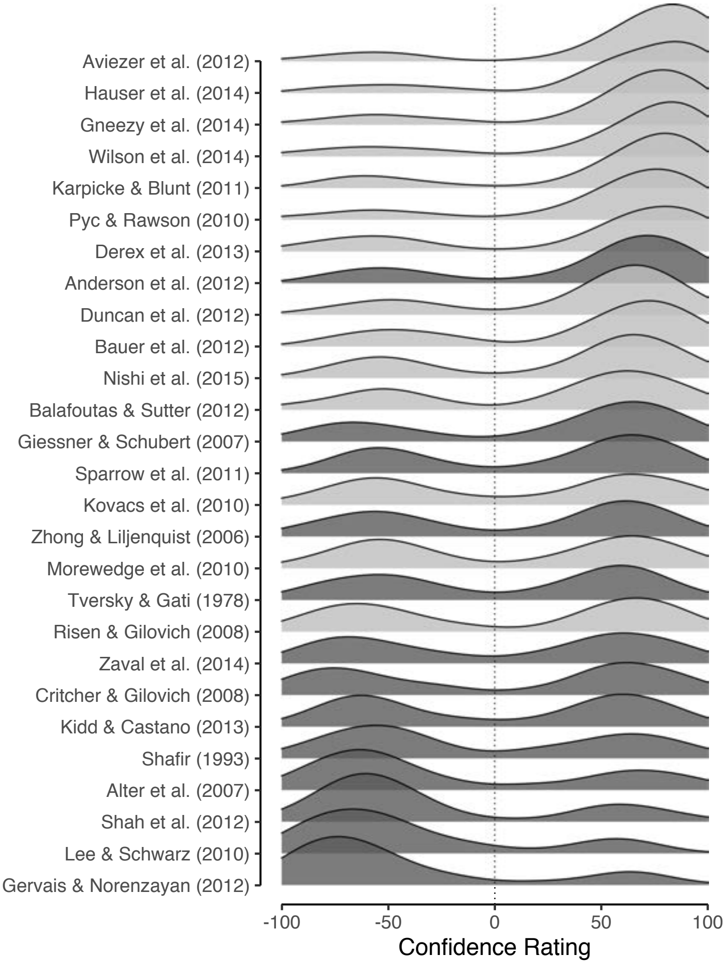


Figure 2.1: Laypeople's near unanimous judgments are highly predictive of replication outcomes. Light density distributions reflect studies that successfully replicated, dark grey distributions reflect studies that did not replicate. Confidence ratings are aggregated over both experimental conditions. Negative values reflect the 'does not replicate' prediction, and positive values the 'replicates' prediction.

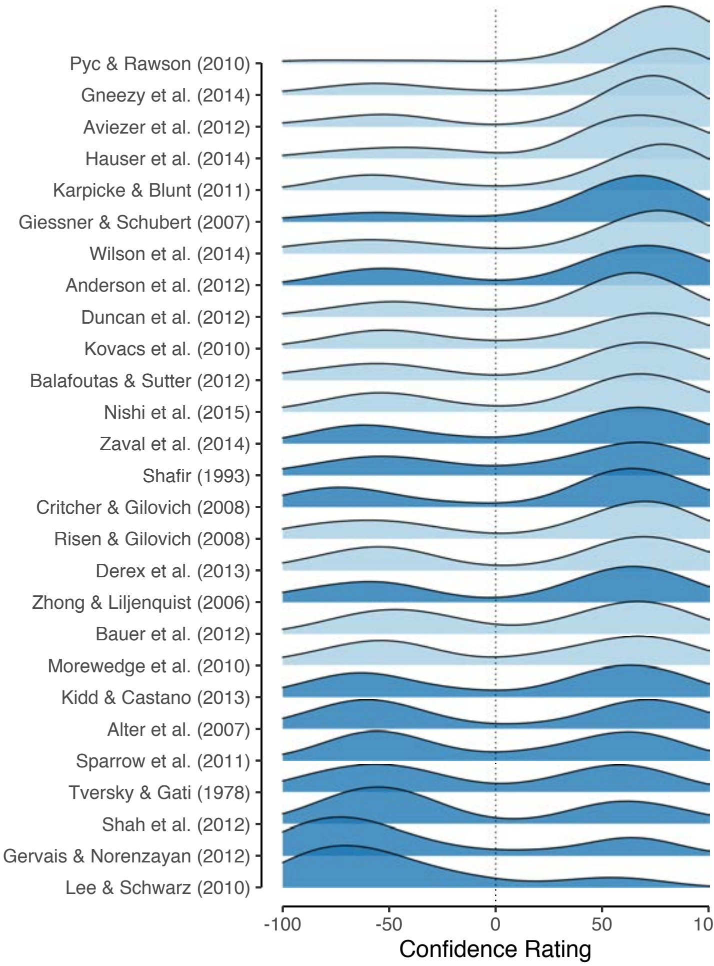


Figure 2.2: Distribution of participants' confidence ratings for each of the 27 studies, for the Description Only condition. The studies are ordered according to their average confidence ratings. Light shading indicates that a study was successfully replicated, and dark shading indicates that a study was not successfully replicated. Negative values indicate a prediction of replication failure, and positive values indicate a prediction of replication success.

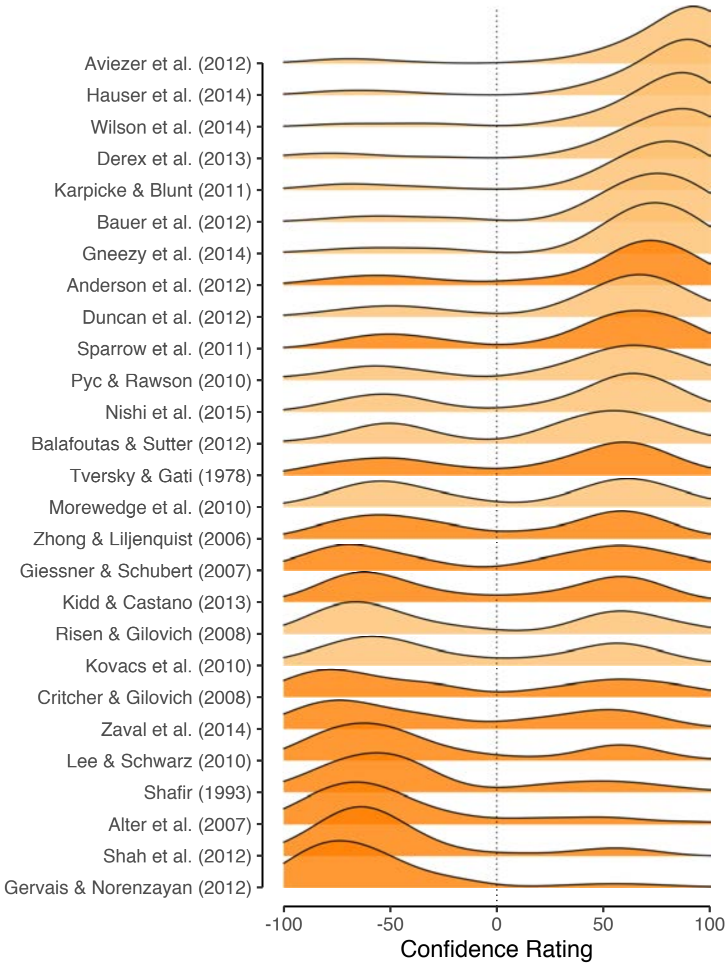


Figure 2.3: Distribution of participants' confidence ratings for each of the 27 studies, separately for the Description Plus Evidence condition. The studies are ordered according to their average confidence ratings. Light shading indicates that a study was successfully replicated, and dark shading indicates that a study was not successfully replicated. Negative values indicate a prediction of replication failure, and positive values indicate a prediction of replication success.

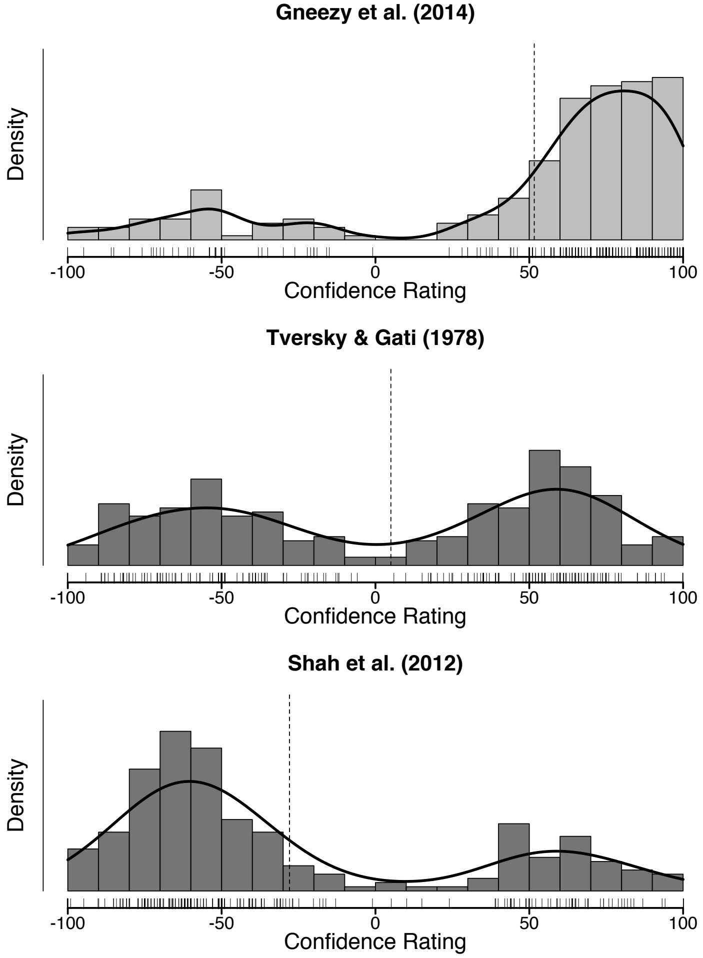


Figure 2.4: Histograms of confidence ratings for three studies for which laypeople were nearly unanimous in their belief that the study will either replicate (Gneezy et al., 2014, top panel) or will not replicate (Shah et al., 2012, bottom panel) or for which they are ambiguous (Tversky & Gati, 1978, middle panel). The vertical dotted line shows the average confidence rating for the respective study (i.e., group prediction).

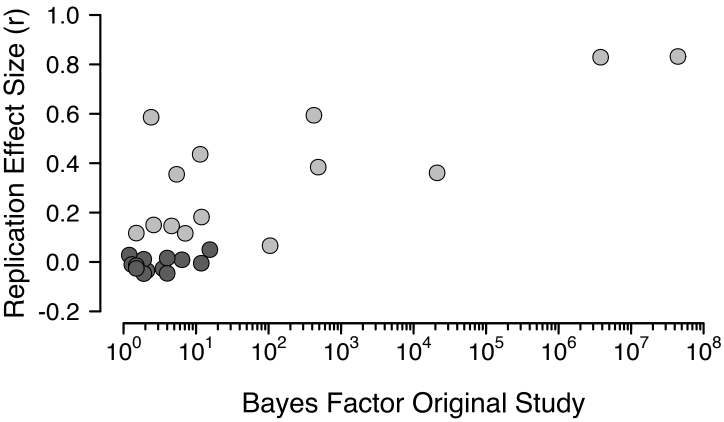


Figure 2.5: The evidence of the original studies (quantified by Bayes factors) is positively associated with replication effect sizes. The dark grey dots indicate the studies that did not replicate, the light grey dots indicate the studies that did replicate.

the effect sizes of the *replication studies*. To a large extent our study was based on the assumption that the Bayes factors of the original studies carry relevant information about replicability. To verify this claim we computed a Spearman correlation coefficient ρ between the log-transformed Bayes factors of the original studies and the standardized effect sizes of the replication studies expressed as correlation coefficients r.

The data provided overwhelming evidence in favor of a positive correlation ($\text{BF}_{+0} = 183$).[5] The median and 95% credible interval for the correlation coefficient ρ were $0.60\,[0.30, 0.77]$, indicating that the Bayes factors of the original studies indeed conveyed useful information (see Figure 2.5).

2.4.2 Preregistered Analyses

2.4.2.1 Quality Check

As preregistered, we implemented a quality check for the data that served as prerequisite for our confirmatory analyses. We considered the data inappropriate for subsequent analyses in case the data provided strong evidence for the hypothesis that overall laypeople performed *worse* than chance level when predicting the replicability of empirical studies. An accuracy rate that is worse than chance level (i.e., less than 50%) indicates that participants either did not understand or follow the instructions correctly, or misinterpreted the presented information (i.e., the description of the study and the Bayes factor).

We tested the restricted hypothesis $\mathcal{H}_{r1}$ that the overall accuracy of laypeople is smaller than 50%, that is $\mathcal{H}_{r1} : \omega < 0.5$, where ω is the mode of the Beta

[5] The subscripts on the Bayes factor to refer to the hypotheses being compared, with the first and second subscript referring to the one-sided hypothesis of interest and the null hypothesis, respectively.

distribution for the group-level accuracy rate. This hypothesis was tested against the encompassing hypothesis $\mathcal{H}_e$ which lets ω free to vary, that is $\mathcal{H}_e : \omega \sim \text{Beta}(1, 1)$. The Bayes factor in favor for the encompassing hypothesis, BF_{er1}, was computed using the encompassing prior approach (Klugkist, Kato, & Hoijtink, 2005). The evidence for the encompassing hypothesis was estimated to approach "infinity", that is $\text{BF}_{er1} = \infty$, which means that the data passed the quality check.[6]

2.4.2.2 Difference in prediction performance between conditions

For the confirmatory analyses, we first investigated whether there was a difference between the two study conditions. Specifically, we evaluated whether or not the inclusion of the Bayes factor for the original effect increased prediction performance as measured by individual Brier scores (Brier, 1950). The Brier score takes into account both the accuracy and the indicated (un)certainty of the prediction; highly certain correct predictions are rewarded and highly certain incorrect predictions are punished, relative to uncertain predictions. As preregistered, individual Brier scores were log-transformed to account for skewness in the distribution of Brier scores.

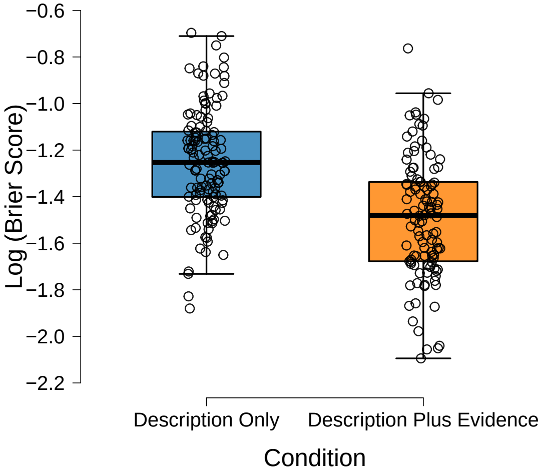
We conducted a Bayesian independent samples t-test with the log Brier score as dependent variable and the condition assignment as grouping variable. The hypothesis of interest states that the Brier scores of participants in the Description Plus Evidence condition are lower than the Brier scores of participants in the Description Only condition, with lower scores indicating better prediction performance. This one-sided default alternative hypothesis was specified as effect size δ for the difference being smaller than zero, that is $\mathcal{H}_- : \delta < 0$. The hypothesis was tested against the null hypothesis $\mathcal{H}_0$ that the effect size is exactly zero, that is $\mathcal{H}_0 : \delta = 0$. The results reveal overwhelming evidence that laypeople in the Description Plus Evidence condition outperform laypeople in the Description Only condition, $\text{BF}_{-0} = 1.0 \times 10^{10}$. The median of the effect size distribution is -0.96, with a 95% credible interval of $[-0.68, -1.23]$ (see Figure 2.6 for a boxplot of the data as well as the prior and posterior distribution of the effect size δ).

2.4.2.3 Group accuracy per condition

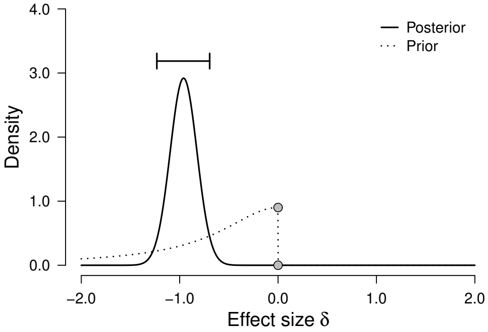
To investigate whether laypeople can adequately predict replication outcomes, we tested whether the group-level accuracy rates[7] are above chance level, that is, higher than 50%. Here, we only considered the accuracy of predictions regardless of raters' confidence. We applied a Bayesian hierarchical model to analyze the accuracy data. For each condition separately, we then tested the restricted hypotheses that accuracy rate ω (i.e., the mode of the group-level distribution) was

[6] When using the encompassing prior approach, we can obtain a Bayes factor estimated to be "infinite" if no posterior samples are in accordance with the restricted hypothesis.

[7] Note that group-level accuracy refers to the accuracy for the 'average' individual, which is estimated in a hierarchical model. A hierarchical model has the benefit that it shrinks individual estimates towards the group-level mean, thereby reducing the influence of extreme cases. Note, however, that the estimated group-level accuracy differs from the accuracy of the group as a collective (the latter being simply the aggregate across people per study).



Boxplot of log Brier scores per condition.



Prior and posterior distribution of population effect size δ.

Figure 2.6: The data and distribution of effect size δ of the Brier scores show that laypeople who received both the study descriptions and information about the strength of the evidence in the original study performed better than laypeople who received the study descriptions only. Figures created in JASP (JASP Team, 2021).

higher than chance for laypeople in the the Description Only condition (denoted as $\mathcal{H}_{r2}$), and for laypeople in the Description Plus Evidence condition (denoted as $\mathcal{H}_{r3}$), that is, $\mathcal{H}_{r2}, \mathcal{H}_{r3} : \omega > 0.5$. The hypotheses $\mathcal{H}_{r2}$ and $\mathcal{H}_{r3}$ were tested against the null hypothesis $\mathcal{H}_0$ stating that ω should be exactly equal to 0.5, which would indicate chance level performance: $\mathcal{H}_0 : \omega = 0.5$.

The data provide extreme support for the restricted hypothesis that laypeople in the Description Only condition perform better than chance, $\text{BF}_{r20} = 3.4 \times 10^8$.

The median and 95% credible interval for the parameter ω are $0.58\,[0.57, 0.60]$, which implies a 58% accuracy rate for laypeople in the Description Only condition at the group level. The data also provide extreme support for the restricted hypothesis that laypeople in the Description Plus Evidence condition perform above chance level, $\mathrm{BF}_{r30} = 2.8 \times 10^{24}$. The median and 95% credible interval for the parameter ω are $0.67\,[0.65, 0.69]$, implying a 67% accuracy rate for laypeople in the Description Plus Evidence condition at the group level. The non-overlapping credible intervals of the two conditions corroborate the results from the independent samples *t*-test on the Brier scores; accuracy is higher in the Description Plus Evidence condition than in the Description Only condition. The distributions of both groups of laypeople are displayed in Figure 2.7.

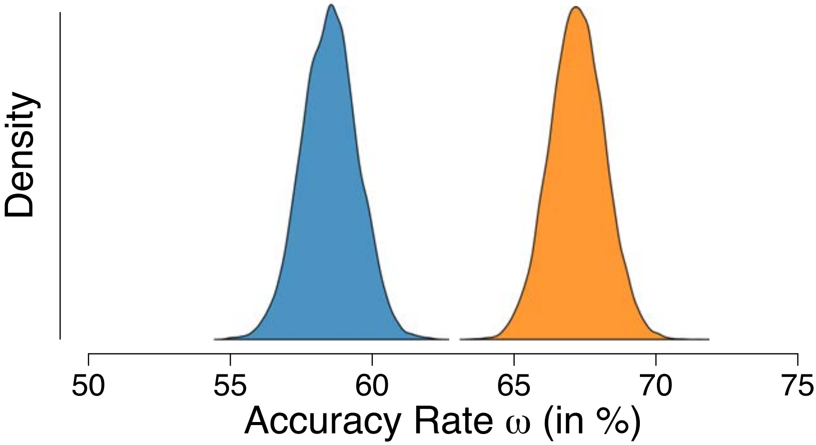
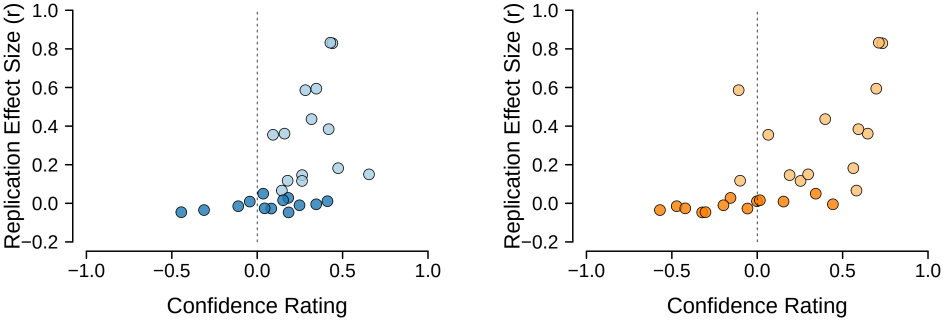


Figure 2.7: Accuracy rates of laypeople in both conditions. Posterior distributions of the group-level accuracy rate for laypeople in the Description Only condition are depicted in blue and those of laypeople in the Description Plus Evidence condition are depicted in orange.

2.4.2.4 Correlation between laypeople's confidence and replication effect size

In addition to the analysis of laypeople's binary predictions of replicability, we assessed whether the confidence with which people make their decisions is indicative of the size of the effect observed in the replication studies (cf. Camerer et al., 2018). In other words, we tested whether laypeople are more certain about their decisions if the replication effect size is large, and become less certain (i.e., more certain about non-replicability) as the underlying replication effect size approaches zero. The replication effect sizes were retrieved from Camerer et al. (2018) and R. Klein et al. (2018). The data are plotted in Figure 2.8, displayed per condition.



Description Only condition

Description Plus Evidence condition

Figure 2.8: Relationship between the average confidence rating per study and the replication effect size for the Description Only condition and the Description Plus Evidence condition. The dotted line represents the cutoff between perceived confidence in successful replication (i.e., positive values), and the perceived confidence in failed replication (i.e., negative values). The dark dots indicate the studies that did not replicate, and the light dots indicate the studies that did replicate.

We used a Bayesian Spearman correlation (van Doorn, Ly, Marsman, & Wagenmakers, 2018) to test the null hypothesis (i.e., $\mathcal{H}_0 : \rho = 0$) against the one-sided restricted hypothesis that the correlation coefficient ρ is positive, for both the Description Only condition (i.e., $\mathcal{H}_{r4} : \rho > 0$), and the Description Plus Evidence condition (i.e., $\mathcal{H}_{r5} : \rho > 0$). The data provide extreme evidence for the restricted hypothesis $\mathcal{H}_{r4}$ of a positive correlation between the average confidence ratings of laypeople and the replication effect sizes in both the Description Only ($\text{BF}_{r40} = 231$) and the Description Plus Evidence condition ($\text{BF}_{r50} = 7126$). For the Description Only condition the median and 95% credible interval for the distribution of the Spearman correlation coefficient ρ are $0.60\,[0.31, 0.76]$. For the Description Plus Evidence condition the median and 95% credible interval for the distribution of ρ are $0.76\,[0.56, 0.87]$. Note that for studies that did not replicate, the effect sizes -by definition- cluster around zero. Although the Spearman correlation coefficient is a rank-based measure, the correlation should still be interpreted with caution.

2.4.3 Exploratory Analyses

2.4.3.1 Disentangling discriminability and response bias

According to signal detection theory (SDT; Green & Swets, 1966; Tanner Jr & Swets, 1954), binary decisions are driven by two main components: the ability to distinguish between the response options (discriminability) and the a priori tendency to prefer one option over the other (response bias). In an exploratory analysis, we applied SDT to decompose laypeople's predictions into discriminability and bias. Here, the *discriminability* relates to the degree to which replicable and unreplicable studies are distinguishable, which is influenced by characteristics of the stimuli (i.e., information provided about the studies) and by raters' underlying ability (i.e., individual prediction skills). The *bias* reflects laypeople's overall tendency towards either predicting that a given study will replicate or predicting that it will not replicate, regardless of the information about the respective study. These parameters were estimated by applying a Bayesian hierarchical equal-variance Gaussian SDT model Lee and Wagenmakers (p. 164 2013).

Figure 2.9 shows the group-level posterior distributions of the discriminability and bias parameters based on the replication predictions, separately for the two conditions. Larger values for discriminability (bottom panel) indicate higher ability to distinguish replicable from unreplicable findings. Consistent with the Brier score analysis reported above, the discriminability parameters show a clear difference between conditions; people in the Description Plus Evidence condition (blue in the figure) are better at separating replicable studies from unreplicable studies than people in the Description Only condition (orange in the figure). The enhanced discriminability for the Description Plus Evidence condition is also visualized in the top panels of Figure 2.10, which shows that the separation between the distribution for replicable and unreplicable studies is larger for the Description Plus Evidence condition than for the Description Only condition. For the bias parameter, the difference between conditions is less pronounced; the negative values for bias (Figure 2.9, right panel) indicate that all laypeople in our sample tended to overestimate replicability (i.e., they displayed a bias towards saying 'the study replicates'). This bias also becomes clear in the top panels of Figure 2.10: in both conditions, the adopted criterion is located to the left of the optimal criterion.

The Receiver Operating Characteristic (ROC) curve is often used to interpret the parameter values of the SDT. This curve reflects the proportion of hits (i.e., replication successes that were deemed replicable) and false alarms (i.e., replication failures that were deemed replicable) as a function of all possible levels of bias, given the estimated discriminability. The further the curve moves away from the diagonal (i.e., chance level), the better the classification performance. The derived Area Under the Curve (AUC) metric is used to quantify the information captured by the ROC curve; it reflects the probability that a given stimulus (i.e., study) is correctly classified (i.e., replication successess as replicable and replication failures as unreplicable). We created the ROC curves for laypeople's prediction performance in both conditions as derived from the estimated discriminability (disregarding the estimated bias). The ROC curves in the lower panels of Figure 2.10 again show that the ratio between hits and false alarms was better for people in the Description Plus Evidence condition compared to people in the Description Only condition. This is also quantified by the associated AUC metric; the median and 95% credible interval were $0.62\,[0.60, 0.65]$ for the Description Only condition and $0.74\,[0.72, 0.77]$ for the Description Plus Evidence condition.

Together, the SDT model indicates that access to the statistical evidence predominantly affected discriminability rather than bias. This suggests that the evidence (i.e., the Bayes factor) provided information that enhanced laypeople's ability to correctly distinguish between replicable and unreplicable studies, rather than making them simply more skeptical across the board. Note that we did not conduct any tests, but solely estimated the discriminability and bias parameters per condition, as well as the associated AUC metrics.

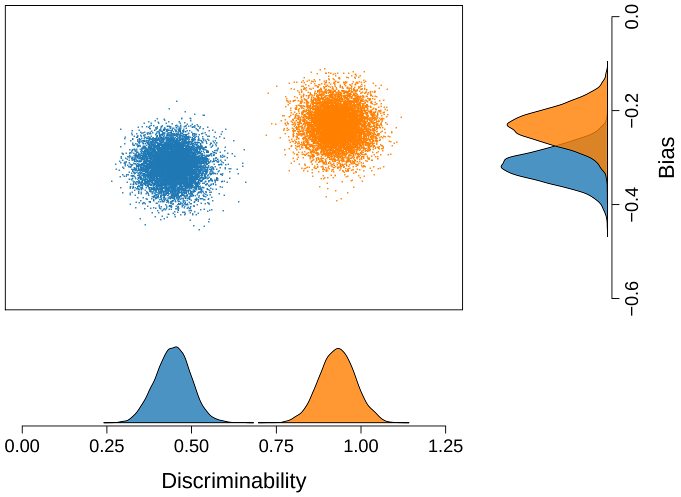
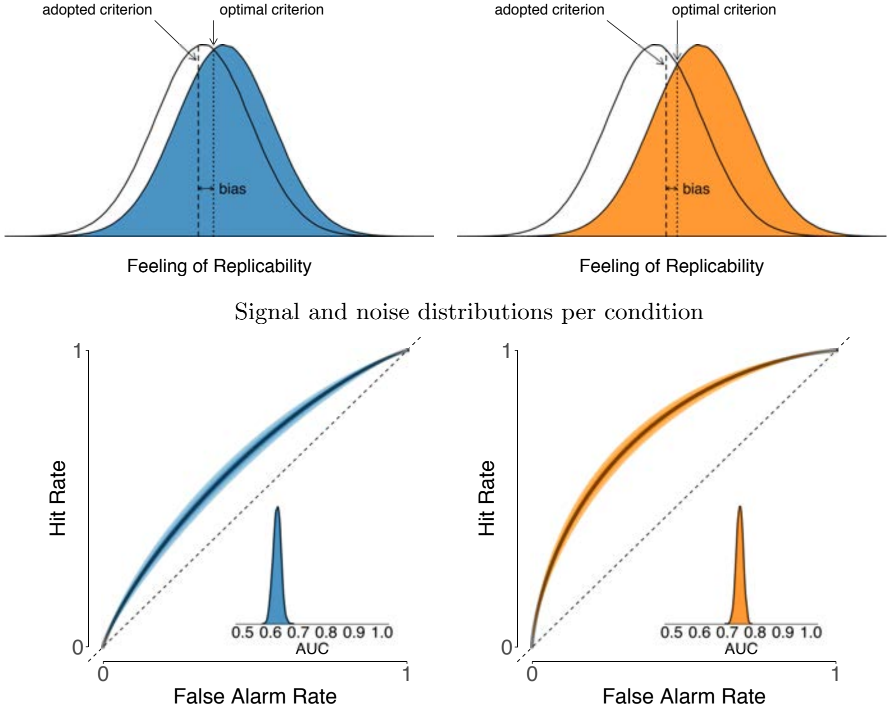


Figure 2.9: Laypeople in both conditions are biased towards predicting that a given study will replicate (as indicated by the posterior distributions of the bias parameter in the right panel). In addition, the posterior distributions of the discriminability parameter in the bottom panel show that laypeople in the Description Plus Evidence condition (orange) have a higher ability to correctly discriminate replicable from unreplicable studies than laypeople in the Description Only condition (blue).

2.4.3.2 Estimating prediction accuracy of experts

In a second exploratory analysis, we applied a Bayesian hierarchical model to generate the posterior distributions of the accuracy rates for the experts' predictions that were measured by Camerer et al. (2018) and Forsell et al. (2018) for the SSRP and ML2 project, respectively. Experts in the SSRP project showed the highest accuracy rate; they were able to correctly predict almost three quarters of the studies, that is, $0.72\,[0.69, 0.74]$. The median accuracy rate of the experts in the ML2 project was 0.65 with a credible interval of $[0.62, 0.68]$. Both expert and non-expert accuracy distributions (expressed as percentages) are presented in Figure 2.11. The figure suggests that the prediction accuracy of laypeople who were provided with a description and Bayes factor of the original study, is at least as good if not better than the prediction accuracy of experts who anticipated outcomes of the ML2 project (and who were also provided with statistics of the original study).

It is important to note, however, that the performance of experts and laypeople may not be completely comparable, as the included studies are only partly overlapping for the different populations (participants in the current study rated 17 studies from the SSRP and 10 from ML2). Unintentionally, the subset drawn from the SSRP included 12 out of 17 studies that replicated successfully, whereas the subset drawn from ML2 included only 2 out of 10 studies that replicated success-



Signal and noise distributions per condition

Group-level ROC curves per condition

Figure 2.10: The top two panels demonstrate that the separation between the noise distribution (white) and signal distribution (colored) is larger for the Description Plus Evidence condition (top right panel; orange) than for the Description Only condition (top left panel; blue). The dashed lines indicate the criteria adopted by the forecasters and the dotted lines indicate the optimal criteria. In the bottom panels, the group-level ROC curves with the 95% credible interval and the posterior distributions of the Area Under the Curve (AUC) metric similarly indicate that laypeople in the Description Plus Evidence condition have a better trade-off between hits and false alarms. The dashed lines indicate chance-level performance. Figure based on Selker et al. (2019).

fully. Because of these unequal proportions, that are also not representative for the respective projects, we estimated accuracy rates for the full set of studies rated by the experts in each project, rather than only the subsets that we presented to laypeople.

2.5 Discussion

The present study showed that laypeople without a professional background in the social sciences are able to predict replicability with above-chance accuracy, even when provided solely with study descriptions. This suggests that intuitions

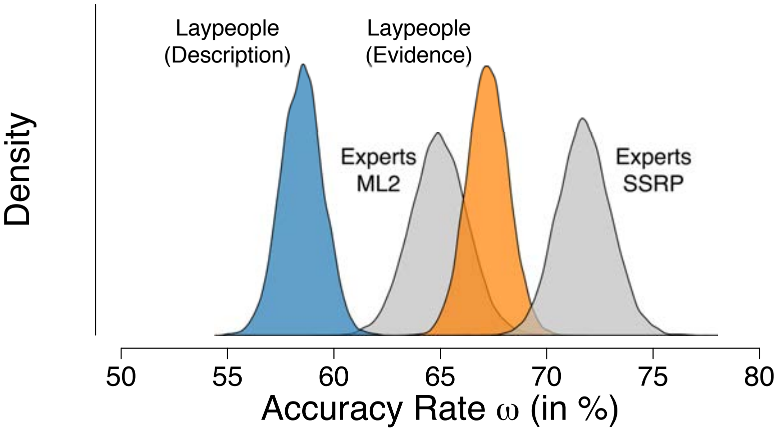


Figure 2.11: Accuracy rates of laypeople and experts. Posterior distributions of the group-level accuracy rates for laypeople in the Description Only condition are displayed in blue and for laypeople in the Description Plus Evidence condition in orange. Posterior distributions of the group-level accuracy rates for experts in the Many Labs 2 Project and in the Social Sciences Replication Project are displayed in grey.

about the plausibility of the targeted effects carry information about the likelihood of a successful replication outcome. Prediction accuracy further increased with access to the statistical evidence (i.e., the Bayes factor) for the original study. In addition to accuracy in binary predictions, laypeople were able to derive a sense of the magnitude of the targeted effects from the descriptions, as indicated by the correlation between raters' confidence in replicability and replication effect size. Again, inclusion of information on the original evidence amplified the relation between confidence ratings and replication effect sizes.

The notion that intuitive plausibility of scientific effects may be indicative of replicability is not novel (nor counterintuitive). The Open Science Collaboration (2015), for instance, already suggested that non-surprising studies are more replicable than highly surprising ones. Wilson and Wixted (2018) built on the data from the Open Science Collaboration (2015) replication project and found that lower prior odds for the crucial effects explained the difference between replicability rates in social and cognitive psychology; social psychological studies contained more risky but potentially groundbreaking effects compared to cognitive psychological studies. The authors suggest that the key factor influencing prior odds of an effect is "established knowledge, acquired either from scientific research or from common experience (e.g., going without sleep makes a person tired)" (Wilson & Wixted, 2018, p. 191). Our study sought to identify exactly this underlying feature of unreplicable studies derived from the latter source of knowledge, which we called "intuitive plausibility", "surprisingness", or "unexpectedness". Our results provided empirical support for the suggestion that intuitive (i.e., non-surprising) studies are more replicable than highly surprising studies, in the sense that replicable studies are in fact deemed more replicable by a naive group of laypeople.

In principle, we expect our results to generalize to most people, provided that the instructions, explanation of replicability, and study descriptions are written in plain language, avoiding technical terms. It is possible that prediction accuracy may rise with increased expertise, for instance, graduate students may on average outperform people without any expertise in social sciences. However, previous prediction studies showed that weighting experts' predictions based on self-reported topical expertise did not improve average prediction accuracy, suggesting that at least knowledge about a particular study's topic may be irrelevant (Dreber et al., 2015; Forsell et al., 2018).

An obvious downside is that generating predictions from laypeople narrows the pool of studies that are suited for prediction surveys; complex psychophysics experiments or fMRI studies may indeed not be comprehensible for laypeople and be better evaluated by experts. However, for the majority of social science studies and related disciplines (e.g., economics) targeting laypeople rather than experts may be advantageous in terms of availability, accessibility, and the possibility to include previously published studies (the results of which experts may already be familiar with or simply look up). A further prerequisite is that the evaluated replication studies should be of high quality (e.g., preregistered, high-powered, featuring manipulation checks, et cetera) to ensure the validity of the accuracy assessment. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

A final side-note on the generalizability of the findings concerns the wider implications and scope of the results. Although participants in our study strongly overestimated overall replicability, they still believed that approximately 20% of the studies would not replicate. This does not necessarily imply, however, that they will distrust the results of 1 in 5 studies they encounter in the media.

The presentation of Bayes factors in the Description Plus Evidence condition could be interpreted as demand characteristics; the quantitative marker plus verbal label may have steered participants' judgments towards the correct conclusions. In the current scenario, it may be practically and theoretically difficult to distinguish between demand characteristics and information given to participants. We do not deny that people may have developed strategies to derive their predictions directly from the value of the Bayes factors. In fact, we assumed that they did so. Although one may argue that this setup creates a confound, one can also conceive it as a demonstration of the benefits of Bayes factors: they constitute a simple metric that can effectively convey information about a study's evidential value. This is not a direct argument for Bayes factors over frequentist p-values and/or effect sizes per se; in fact, we expect that the inclusion of frequentist statistics may similarly enhance laypeople's prediction performance.

We acknowledge that replication outcomes cannot be equated with the 'truth'. Although the projects by Camerer et al. (2018) and R. Klein et al. (2018) were high-powered and followed detailed preregistration protocols, the replication outcomes are not definitive or irrefutable. Moreover, there currently exists no consensus on which decision rule is superior for determining replication success (Cumming, 2008; Open Science Collaboration, 2015; Simonsohn, 2015; Verhagen & Wagenmakers, 2014). We categorized studies into 'successfully replicated' and 'failed to replicate' following the primary replication criteria used in the SSRP and the ML2 project, which were based on finding a significant effect in the same direction as the original study. However, it should be noted that R. Klein et al. (2018) and Camerer et al. (2018) report additional indicators to evaluate replicability that result in slightly different categorizations of replication success. The replication outcomes should thus not be regarded as reflective of the absolute truth, but rather of the current, tentative state of knowledge.

Along the same lines, laypeople's predictions should also not be equated with the truth. Although clearly above chance level, the prediction accuracy rates of 58% and 67% as found for laypeople in the Description Only and the Description Plus Evidence condition, respectively, are far from perfect. One reason for laypeople's moderate prediction success may arise from their tendency to overestimate the replicability of empirical findings; relative to the bleak reality of the current replication rate in psychological science, laypeople are optimists. This pattern becomes evident from Figure 2.1 and is corroborated by the signal detection analysis indicating that laypeople demonstrate a bias toward saying that a given study will replicate. Notably, the optimistic perspective does not seem to be unique to laypeople; experts similarly overestimated replicability in Dreber et al. (2015), Camerer et al. (2016) and Forsell et al. (2018), though not in Camerer et al. (2018). The biased responding may allow for possibilities to boost prediction accuracy; the area under the curve metric indicated that if laypeople adopted the optimal unbiased criterion, i.e., if they were more conservative, then accuracy may be enhanced to 62% for predictions based on verbal descriptions only and 74% based on descriptions plus evidence in the original study. This suggestion is speculative but could be assessed in future research, for instance by manipulating expectations of baseline replicability rates.

Nevertheless, we believe laypeople's predictions are more informative than is captured by the estimated accuracy rates. This is exemplified by the prediction pattern as displayed in Figure 2.1. The pattern suggests that there is a group of studies for which laypeople as a collective were divided (characterized by the symmetrical bi-modal distribution) and a group for which they were in agreement (i.e., the top and bottom rows of the figure). For those studies for which laypeople were nearly unanimous, the predictions were highly accurate. Moreover, as the figure shows, when laypeople as a group predicted that a particular study would fail to replicate, it failed to replicate.

These results emphasize that the scientific culture of striving for newsworthy, extreme, and sexy findings is indeed problematic, as counterintuitive findings are the least likely to replicate. This also relates to the aphorism that "extraordinary (i.e., intuitively implausible) claims require extraordinary evidence". Many studies included in our sample were considered implausible and thus would have required highly compelling evidence to establish the effects. However, the pattern of Bayes factors in Figure 2.5 shows that many original findings were based on weak initial evidence; of the included studies, 37% (10 studies) yielded a Bayes factor lower than 3, evidence that is "not worth more than a bare mention" according to Jeffreys' 1939 criteria. The combination of low intuitive plausibility and weak initial evidence is remarkable and arguably worrisome, especially in the light of the low replication rates in social science. To account for the extraordinary nature of a claim, researchers should adjust the prior probability of the respective alternative hypothesis and the null hypothesis. In the Bayesian framework, this means that a higher Bayes factor is necessary to conclude that the effect is present; in the frequentist framework, a lower p-value is necessary to reject the null hypothesis (cf. Benjamin et al., 2018).

The notion of prediction surveys and markets as a valuable component of replication research seems to be gaining momentum. The Replication Market platform (`https://www.replicationmarkets.com`), for instance, invites researchers as well as the general public to predict and bet on $3,000$ studies associated with the SCORE project (`https://www.darpa.mil/program/systematizing-confidence-in-open-research-and-evidence`). Although these predictions yield valuable insights, we naturally do not advocate to replace replication studies with judgments of the general public – nor with those of experts. Rather, people's predictions may be used to provide a quick snapshot of expected replicability. This can facilitate the replication process by informing the selection of to-be-replicated studies. The uni- versus bimodality of the distribution of replication predictions by laypeople may for instance steer researchers' confidence in whether the predictions are more or less reliable, respectively. Additionally, the relative ordering of laypeople's confidence in replicability for a given set of studies may provide estimations of the relative probabilities of replication success.

If a replicator's goal is to purge the literature of unreliable effects, he or she may start by conducting replications of the studies for which replication failure is predicted by naive forecasters. Alternatively, if the goal is to clarify the reliability of studies for which replication outcomes are most uncertain, one could select studies for which the distribution of the expected replicability is characterized by a bi-modal shape. As such, prediction surveys may serve as 'decision surveys', instrumental in the selection stage of replication research (cf. Dreber et al., 2015). These informed decisions could not only benefit the replicator, but also optimize the distribution of funds and resources for replication projects. This idea could easily be extended to assessing prior plausibility of a proposed and yet to be empirically investigated hypothesis in a systematic fashion, similar to the social science prediction platform (DellaVigna & Vivalt, 2019). An interesting application would be to use these assessments in conjunction with large collaborative research efforts such as the Psychological Science Accelerator (Moshontz et al., 2018). As such, laypeople's predictions may not only contribute to replication research, but also inform the prior plausibility of novel studies.

Chapter 3

A Many-Analysts Approach to the Relation Between Religiosity and Well-being

Abstract

The relation between religiosity and well-being is one of the most researched topics in the psychology of religion, yet the directionality and robustness of the effect remains debated. Here, we adopted a many-analysts approach to assess the robustness of this relation based on a new cross-cultural dataset (N = 10,535 participants from 24 countries). We recruited 120 analysis teams to investigate (1) whether religious people self-report higher well-being, and (2) whether the relation between religiosity and self-reported well-being depends on perceived cultural norms of religion (i.e., whether it is considered normal and desirable to be religious in a given country). In a two-stage procedure, the teams first created an analysis plan and then executed their planned analysis on the data. For the first research question, all but 3 teams reported positive effect sizes with credible/confidence intervals excluding zero (median reported $\beta = 0.120$). For the second research question, this was the case for 65% of the teams (median reported $\beta = 0.039$). While most teams applied (multilevel) linear regression models, there was considerable variability in the choice of items used to construct the independent variables, the dependent variable, and the included covariates.

This chapter is in press as The MARP Team. (2022). A Many-Analysts Approach to the Relation Between Religiosity and Well-being. *Religion, Brain & Behaviour.*

3.1 Introduction

The relation between religion and well-being has been a topic of debate for centuries. While Freud considered religion a “universal obsessional neurosis” and Nietzsche called Christianity “the greatest misfortune of humanity”, the recent scientific literature has painted a more positive picture of religion’s effect on (mental) health (e.g., Gebauer et al., 2017; George, Ellison, & Larson, 2002; Koenig & Larson, 2001; Plante & Sherman, 2001; Seybold & Hill, 2001; Thoresen, 1999; Zimmer et al., 2016). Individual religiosity has, for instance, been related to less depression (T. B. Smith, McCullough, & Poll, 2003), more happiness (Abdel-Khalek, 2006; Lewis & Cruise, 2006), higher life satisfaction (Lim & Putnam, 2010), and even lower mortality (Ebert, Gebauer, Talman, & Rentfrow, 2020; Stavrova, 2015). At the same time, the robustness, universality, and methodological specificity of the religion–well-being relation remains an outstanding question. In this project, we adopted a many-analysts approach to investigate two research questions using a new large cross-cultural dataset featuring $N = 10{,}535$ participants from 24 countries. Specifically, we recruited 120 teams to conduct analyses in order to answer the following two research questions: (1) “Do religious people self-report greater well-being?”, and (2) “Does the relation between religiosity and self-reported well-being depend on perceived cultural norms regarding religion?”. In the subsequent sections, we will first introduce our theoretical framework, dataset, and the many-analysts approach, before describing the key results with respect to the stated research questions and the varying approaches taken by the many-analysts teams. A general discussion of the project and the results is included in the closing article (Hoogeveen, Sarafoglou, van Elk, & Wagenmakers, in preparation).

3.2 Theoretical Background

The literature on the psychology of religion is replete with positive correlations between (self-rated) religiosity and mental health (Abdel-Khalek, 2006; George et al., 2002; Koenig & Larson, 2001; Plante & Sherman, 2001; Seybold & Hill, 2001; T. B. Smith et al., 2003; Thoresen, 1999; Zimmer et al., 2016; see Koenig, 2009 for a review). At the same time, meta-analyses indicate that the relation between religion and well-being is often small (around $r = .1$; Bergin, 1983; Hackney & Sanders, 2003; Koenig & Larson, 2001). In addition, it has been argued that positive associations are found only for particular measures and operationalizations of these constructs (Hackney & Sanders, 2003; Poloma & Pendleton, 1989). A recent meta-analysis of longitudinal studies reported that, out of eight religiosity/spirituality measures, only participation in public religious activities and the importance of religion were statistically significantly related to self-rated mental health, which was operationalized as distress, life satisfaction, well-being, and quality of life (Garssen, Visser, & Pool, 2020).

Furthermore, the type of religiosity (i.e., intrinsic vs extrinsic; positive vs. negative religious coping) and religious status (religious vs. uncertain) appear to moderate the relationship between religion and mental well-being (T. B. Smith et al., 2003; Villani, Sorgente, Iannello, & Antonietti, 2019). For instance, extrinsic religious orientation (i.e., when people primarily use their religious community as a social network, whereas personal religious beliefs are secondary) and negative religious coping (i.e., when people have internal religious guilt or doubts) have been shown to be negatively related to well-being (Abu-Raiya, 2013; Weber & Pargament, 2014). Yet other research suggests that it is precisely the social aspect of religious service attendance and congregational friendships that explains how religiosity is positively associated with life satisfaction (Lim & Putnam, 2010). Moreover, the direction of the religiosity–mental health relation remains unclear; while engaging in religious activities might make people happier, people with better mental health might also be more likely to engage in public, social events.

Additionally, there is large variability in the extent to which religion is ingrained in culture and social identity across the globe (Kelley & de Graaf, 1997; Ruiter & van Tubergen, 2009). Accordingly, when investigating the association between religiosity and well-being, it may be necessary to take into account the cultural norms related to religiosity within a society. Being religious may contribute to self-rated health and happiness when being religious is perceived to be a socially expected and desirable option (Diener, Tay, & Myers, 2011; Ebert et al., 2020; Gebauer et al., 2017; Stavrova, 2015; Stavrova, Fetchenhauer, & Schlösser, 2013). This makes sense from the literature on person-culture fit (Dressler, Balieiro, Ribeiro, & Santos, 2007): a high person-culture fit indicates good agreement between one's personal values and beliefs and the beliefs that are shared by one's surrounding culture. A fruitful way to measure cultural norms is through the shared, intersubjective perception of the beliefs and attitudes that are prevalent in a society (Chiu, Gelfand, Yamagishi, Shteynberg, & Wan, 2010; Zou et al., 2009). Intersubjective norms of religiosity, for instance, refer to the shared perception of the importance of religion within a society or culture. Rather than expressing the importance of religious beliefs and behaviors in one's own personal life, intersubjective norms of religiosity (henceforth: cultural norms of religiosity) uncover the perceived importance of religious beliefs and behaviors for the average person within their culture. Religious individuals may be more likely to benefit from being religious when their convictions and behaviors align with perceived cultural norms. For countries in which religion is more trivial or even stigmatized, the relation between religiosity and well-being may be absent or even reversed. Relatedly, in secular countries, religion might be practiced relatively often by minority groups, which has been shown to attenuate the positive association between religious involvement and well-being (Hayward & Elliott, 2014; Huijts & Kraaykamp, 2011; May & Smilde, 2016; Okulicz-Kozaryn, 2010).

3.3 A Many-Analysts Approach

In the current project, we aim to shed light on the association between religion and well-being and the extent to which different theoretically- or methodologically-motivated analytic choices affect the results. To this end, we initiated a many-analysts project, in which several independent analysis teams analyze the same dataset in order to answer a specific research question (e.g., Bastiaansen et al., 2020; Boehm et al., 2018; Botvinik–Nezer et al., 2020; Silberzahn et al., 2018; van

Dongen et al., 2019). A many-analysts approach has been proposed as a way to mitigate the influence of individual-researcher biases (e.g., confirmation bias by the proponent of a theory or disconfirmation bias by the skeptic), especially since the analysis teams are not typically invested in the outcome. More generally, a many-analysts study is arguable less vulnerable to publication bias toward publishing only significant rather than null results, which may lower the (unconscious) tendency toward p-hacking by individual analysts. A many-analysts approach can balance out the effects of researcher bias while still allowing for expertise-based analytic decisions such as reasonable preprocessing steps, variable exclusion, and model specification. As such, it enables one to assess the robustness of outcomes and quantify variability based on theory-driven analysis decisions and plausible statistical models. Specifically, we believe that the more consistent the results from different analysis teams are, the more confident we can be in the conclusions we draw from the results. A many-analysts approach may be preferable to an exhaustive multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016) that might simply include the full spectrum of options, including those that are theoretically and methodologically unrealistic.

The idea of inviting different analysis teams to answer the same research question using the same data is relatively novel (Silberzahn & Uhlmann, 2015; see Aczel et al., 2021 for general guidelines); we are aware of three papers in neuroscience (Botvinik–Nezer et al., 2020; Fillard et al., 2011; Maier-Hein et al., 2017), one in microeconomics (Huntington-Klein et al., 2021), and eight in psychology, three of which pertain to cognitive modeling (Boehm et al., 2018; Dutilh, Annis, et al., 2019; Starns et al., 2019) while the remaining five are from other fields of psychology (Bastiaansen et al., 2020; Salganik, Lundberg, Kindel, Ahearn, Al-Ghoneim, et al., 2020; Schweinsberg et al., 2021; Silberzahn et al., 2018; van Dongen et al., 2019). Most similar to the current work are the projects that applied a many-analysts approach to perform statistical inference on the relation between two variables, such as skin color and red cards in soccer (Silberzahn et al., 2018), scientist gender and verbosity (Schweinsberg et al., 2021), or amygdala activity and stress (van Dongen et al., 2019). While the exact focus of previous many-analysts projects varied (e.g., experience sampling, fMRI preprocessing, predictive modeling, proof of the many-analysts concept), the take-home messages were rather consistent: all papers showed that different yet equally justifiable analytic choices result in very different outcomes, sometimes with statistically significant effects in opposite directions (e.g., Schweinsberg et al., 2021; Silberzahn et al., 2018). In addition, it has proved difficult to pinpoint the exact sources of variability due to the fact that analytic approaches differed in many respects simultaneously (e.g., exclusion criteria, inclusion of covariates etc.). Nevertheless, the outcomes of these previous projects suggest that choices of statistical model (Silberzahn et al., 2018), statistical framework (van Dongen et al., 2019), (pre)processing software (Botvinik–Nezer et al., 2020), and the variables themselves (Schweinsberg et al., 2021) exert substantial effects on the results and conclusions.

We believe a many-analysts approach is uniquely suited to address various concerns in the study of religion and well-being. First, the relation between religion and health has been researched for decades with hundreds of qualitative reports, cross-sectional and longitudinal studies, and even randomized controlled trials with religious/spiritual interventions for mental health issues (Captari et al., 2018; J. I. Harris et al., 2018; Koenig, Al-Zaben, & VanderWeele, 2020; Rosmarin, Pargament, Pirutinsky, & Mahoney, 2010). Yet new studies keep emerging (e.g., M.-C. Chang et al., 2021; Luo & Chen, 2021; Simkin, 2020) and the debate seems far from settled (see for instance the recent special issue in the International Journal for the Psychology of Religion; van Elk, 2021). Second, both 'religion' and 'well-being' are broad and multifaceted constructs that are sensitive to different measures and operationalizations, which might result in both quantitatively and qualitatively different conclusions (Hackney & Sanders, 2003; Poloma & Pendleton, 1989). Third, the standard way to assess robustness of an effect or association is often through meta-analysis, but the fragmentation of the literature on the religion–health link and methodological heterogeneity between studies challenge the use and validity of meta-analyses in this domain (Koenig, Hill, Pirutinsky, & Rosmarin, 2021). In general, meta-analyses may suffer from several drawbacks such as publication bias and sensitivity to arbitrary methodological choices (e.g., different meta-analytic techniques can result in different conclusions; de Vrieze, 2018; van Elk et al., 2015). Moreover, the estimated effect sizes in meta-analyses might be as much as three times larger than in preregistered multiple-site replication studies (Kvarven, Strømland, & Johannesson, 2020). Fourth, the discussion on the potential health-benefits of religion has been muddied by concerns about researcher interests and biases. That is, it has been argued that scholars of religion might be biased by their own (religious) beliefs (Ladd & Messick, 2016; Swigart, Anantharaman, Williamson, & Grandey, 2020; Wulff, 1998) or by the fact that a substantial amount of research in the science of religion is funded by religiously-oriented organizations such as the John Templeton Foundation (Bains, 2011; Wiebe, 2009).[1] Inviting independent analysts from various backgrounds including but not restricted to religious studies attenuates this potential concern. Moreover, in addition to quantifying variability, with a sufficiently large number of analysis teams one can also investigate factors that might explain observed variability, such as those related to theoretical or methodological expertise and prior beliefs (Aczel et al., 2021).[2]

In addition to the theoretical rationale for using a many-analysts approach to answer the research questions at hand, we also consider the current dataset particularly appropriate for such an approach. That is, the complexity of the data allows for many justifiable choices for the operationalization of the variables and the statistical approach to be employed. While the questions posed to the participants in the cross-cultural study could no longer be changed, the specific method of derivation for the religiosity and well-being scores was at the discretion of the many analysts. At the same time, the research questions and data structure (cross-sectional correlational data) were sufficiently intuitive and manageable to inspire many researchers in the fields of (social) psychology, religious studies,

[1] Ironically, so is the present project.

[2] Note that we acknowledge that another important problem in the literature on religion and well-being concerns the issue of causality. However, as our project uses non-experimental cross-sectional data, this issue cannot immediately be addressed in the current study (but see Grosz, Rohrer, & Thoemmes, 2020; Rohrer, 2018 for a perspective on causal inference in non-experimental studies).

health science, and general methodology to propose an analysis.

Finally, we believe that our project involves a combination of elements that extend existing many-analysts work. First, we collected new data for this project with the aim to provide new evidence for the research questions of interest, as opposed to using an existing dataset that has been analyzed before. Second, we targeted both researchers interested in methodology and open science, as well as researchers from the field of the scientific study of religion and health to encourage both methodologically sound and theoretically relevant decisions (see the section 'Analysis teams'). Third, in comparison to previous many-analysts projects in psychology, the current project includes a lot of teams (i.e., 120 vs. 4, 12, 14, 17, 27, 29, and 70 teams, though note that a machine learning project included 160 analyst teams; Salganik, Lundberg, Kindel, Ahearn, Al-Ghoneim, et al., 2020). Fourth, we applied a two-step procedure that ensured a purely confirmatory status of the analyses: in stage 1, all teams first either completed a preregistration or specified an analysis pipeline based on a blinded version of the data. After submitting the plan to the OSF, teams received the real data and executed their planned analyses in stage 2 (see Sarafoglou, Hoogeveen, & Wagenmakers, 2022 for more details on and an empirical investigation of preregistration vs. data blinding based on the present data). Fifth, the many-analysts approach itself was preregistered prior to cross-cultural data collection (see osf.io/xg8y5), although the details of the processing and analysis of the many-analysts data were not preregistered.

3.4 The Dataset

The dataset provided to the analysts featured data from 10,535 participants from 24 countries collected in 2019. The data were collected as part of the cross-cultural religious replication project (see also Hoogeveen et al., 2021; Hoogeveen & van Elk, 2018). The dataset contained measures of religiosity, well-being, perceived cultural norms of religion, as well as some demographic items. The full dataset, the data documentation file, and original questionnaire can be found on the OSF project page (osf.io/qbdce/).

Participants Participants were recruited from university student samples, from personal networks, and from (demographically representative) samples accessed by panel agencies and online platforms (MTurk, Kieskompas, Sojump, TurkPrime, Lancers, Qualtrics panels, Crowdpanel, and Prolific). Participants were compensated for participation by financial remuneration, the possibility for a reward through a raffle, course credits, or received no compensation. Everyone aged 18 years or above could participate.[3]

Participants were required to answer all multiple choice questions, and hence there were no missing data (except for 36 people who did not provide a numeric age and 995 people who chose not to answer the item on sexual satisfaction, as this was the only item for which participants were not required to provide an answer.) The countries were convenience-sampled (i.e., through personal networks), but

[3] Note that we did not exclude the 19 participants who indicated they were younger than 18 (but some of the analysis teams did exclude these participants).

were selected to cover six continents and include different ethnic and religious majorities. The final sample included individuals who identified as Christian (31.2%), Muslim (6.1%), Hindu (2.9%), Buddhist (2.0%), Jewish (1.0%), or were part of another religious group (2.9%). Finally, 53.9% of participants did not identify with any religion. See Tables B1 and B2 in the online appendix for the full descriptive statistics of the dataset.

Measures Personal religiosity was measured using nine standardized self-report items taken from the World Values Survey (WVS; World Values Survey, 2010), covering religious behaviors (institutionalized such as church attendance and private such as prayer/meditation), beliefs, identification, values, and denomination. The well-being measure consisted of 18 self-report items from the validated short version of Quality of Life scale, as used by the World Health Organization (WHOQOL-BREF; WHOQOL Group, 1998). Included items cover general health and well-being, as well as the domains of physical health, psychological health and social relationships. Specific items evaluated: the quality of life in general, and satisfaction of overall health (general); pain, energy, sleep, mobility, activities, dependence on medication, and work capability (physical domain); life enjoyment, concentration, self-esteem, body-image, negative feelings, and meaningfulness (psychological domain); as well as personal relationships, social support, and sexual satisfaction (social domain). In addition to the raw scores for each item, we also provided an overall mean, as well as three means per subscale, following the calculation instructions in the WHOQOL-BREF manual. Cultural norms of religiosity were measured with two items assessing participants' perception of the extent to which the average person in their country considers a religious lifestyle and belief in God/Gods/spirits important (Wan et al., 2007). Finally, demographics were measured at the individual level (i.e., age, gender, level of education, subjective socioeconomic status (SES), and ethnicity) whereas GDP per capita (current US$, World Bank Group, 2017), sample type (e.g., university students, online panels), and means of compensation (e.g., course credit, monetary reward) were determined at the country/sample level. Items were reverse-coded when applicable. Personal religiosity items were additionally rescaled to the 0-1 range to make them contribute equally to an average religiosity score since the items were measured on different scales (e.g., a 1-8 Likert scale or a 'yes/no' item, which was coded as 'no'=0 and 'yes'=1).[4] GDP was provided as a raw value as well as standardized at the country level.

3.5 Disclosures

3.5.1 Data, materials, and preregistration

At the start of this project we did not envision a particular statistical analysis to be executed across the reported results from the individual teams, and therefore we did not preregister any statistical inference procedure. However, at an earlier

[4] When teams indicated that they preferred the raw data, we provided the function to back-transform the data.

stage, we did preregister our own hypotheses regarding the research questions that were posed to the analysis teams (see osf.io/zyu8c/). This preregistration also anticipates the many-analysts approach, yet does not specify the exact details of the project. In this preregistration document, we indicated that the analysis teams would first receive a blinded version of the data, but we later decided that half of the teams would work with blinded data and the other half would write their own preregistration (see Sarafoglou et al., 2022). Note that we did not include our own estimated effect sizes in the results as shown below. Our results, however, do corroborate the overall pattern of results from the analysis teams. Interested readers can access our preregistered analysis of the research questions on the OSF (osf.io/vy8z7/).

All documents provided to the analysis teams (dataset, documentation, questionnaire), as well as the administered surveys, the anonymized raw and processed data (including relevant documentation), and the R code to conduct all analyses (including all figures), can be found on the project page on the OSF (osf.io/vy8z7/). Identifying information (such as names, email-addresses, universities) was removed from all free-text answers. See also Table 3.2 for an overview of all resources. Online appendices can be accessed via https://osf.io/9kpfu/.

3.5.2 Reporting

We report how we determined our sample size, all data exclusions, and all manipulations in the study. However, it should be noted that this project also involved an empirical evaluation of analysis blinding, which is reported in another paper (i.e., Sarafoglou et al., 2022). Here, we only describe measures relevant to the theoretical research questions and the many-analysts approach. The description of the remaining measures that were only used for the experimental analysis proposal manipulation can be found in Sarafoglou et al. (2022).

3.5.3 Ethical approval

The study was approved by the local ethics board of the University of Amsterdam (registration number: 2019-PML-12707). All participants were treated in accordance with the Declaration of Helsinki. See the online appendix for details on the ethical approval for the cross-cultural data collection.

3.6 Methods

3.6.1 Analysis Teams

The analysis teams were recruited through advertisements in various newsletters and email lists (e.g., the International Association for the Psychology of Religion (IAPR), International Association for the Cognitive Science of Religion (IACSR), Society for Personality and Social Psychology (SPSP), and the Society for the Psychology of Religion and Spirituality (Div. 36 of the APA)), on social media platforms (i.e., blogposts and Twitter), and through the authors' personal network. We invited researchers of all career stages (i.e., from doctoral student to

Table 3.1: Career Stages and Domains of Expertise Featured in the 120 Analysis Teams.

	Percentage of teams
Career Stages	
Doctoral Student	54 (45 %)
Post-doc	45 (37.50 %)
Assistant Professor	32 (26.67 %)
Associate Professor	26 (21.67 %)
Full Professor	20 (16.67 %)
Domains of Expertise	
Social Psychology	43 (35.83 %)
Cognition	28 (23.33 %)
Methodology and Statistics	25 (20.83 %)
Religion and Culture	25 (20.83 %)
Psychology (Other)	19 (15.83 %)
Health	17 (14.17 %)

Note. Teams may include multiple members of the same position and in the same domain.

full professor). Teams were allowed to include graduate and undergraduate students in their teams as long as each team also included a PhD candidate or a more senior researcher. Initially, $N = 173$ teams signed up to participate in the many-analysts project. From those teams, $N = 127$ submitted an analysis plan and $N = 120$ completed the project. The members from each analysis team were offered co-authorship on the main manuscript. No individual researcher or team was excluded from the study.

The number of analysts per team ranged from 1 to 7, with most teams consisting of 1 (41%) or 2 (33%) analysts (median = 2). The different career stages and domains of expertise featured in the analysis teams are given in Table 3.1. In addition, Figure 3.1 shows the self-rated collective knowledge about the topic of religion and well-being and about methodology and statistics. As becomes evident, most of the analysis teams had more methodological and/or statistical expertise than substantive expertise; 80% of the teams reported considerable expertise with regard to methods and statistics compared to 31% with regard to religion and well-being, 19% compared to 17% was neutral, and 3% compared to 50% reported little to no knowledge, respectively.

3.6.2 Sampling Plan

For a separate component of the project (see Sarafoglou et al., 2022), the preregistered sample size target was set to a minimum of 20 participating teams, which was based on the recruited analysis teams in the many-analysts project from Silberzahn et al. (2018). However, we did not set a maximum number of participating

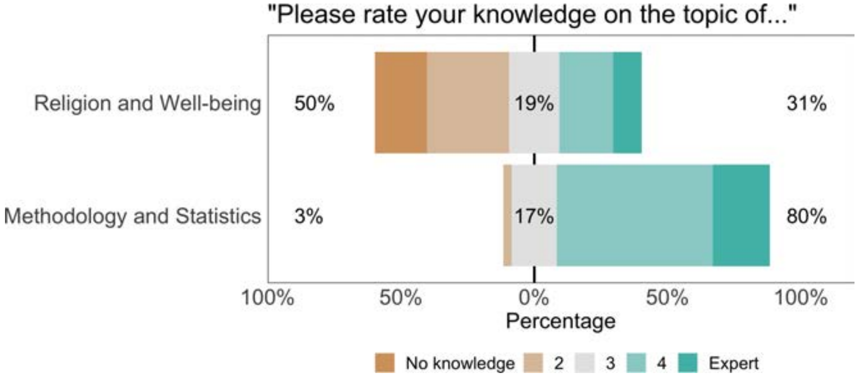


Figure 3.1: Responses to the survey questions on self-rated topical and methodological knowledge. The top bar represents the teams' answers about their knowledge regarding religion and well-being and the bottom bar represents the teams' answers about their knowledge regarding methodology and statistics. For each item, the number to the left of the data bar (in brown/orange) indicates the percentage of teams that reported little to no knowledge. The number in the center of the data bar (in grey) indicates the percentage of teams that were neutral. The number to the right of the data bar (in green/blue) indicates the percentage of teams that reported (some) expertise.

teams. The recruitment of analysis teams was ended on December 22, 2020.

3.6.3 Materials

3.6.3.1 Surveys

The analysts received three surveys, here referred to as the pre-survey, the mid-survey, and the post-survey. In the pre-survey, participating teams indicated the career stages and domains of expertise featured in their team, self-rated their (collective) theoretical and methodological knowledge (self-reported; 5-pt Likert scale), and anticipated the likelihood of the effects of interest (7-pt Likert scale). In the mid-survey, teams were asked about the experienced effort, frustration, workload in hours spent on the project, and the extent to which this workload was lower or higher than expected for the analysis planning phase (i.e., stage 1; 7-pt Likert scales). In the post-survey, the teams provided the results of their analyses and again indicated their experiences during the analysis executing phase (i.e., stage 2). Specifically, per research question, teams were asked about their statistical approach, the operationalization of the independent variable(s) and dependent variable(s), included covariates, analytic sample size, (unit of) effect size, *p*-value or Bayes factor, and additional steps they took for the analysis. Furthermore, for both research questions, the teams gave a subjective conclusion about the evidence for the effect (i.e., "good evidence for a relation", "ambiguous evidence", or "good evidence against a relation"), about the practical meaningfulness/relevance of the effect (based on the data; "yes" or "no"), and indicated again the likelihood of the effects of interest (on a 7-pt Likert scale). Additionally, teams indicated the appropriateness of their statistical approach (7-pt Likert scale), the suitability of the dataset for answering each research question (7-pt Likert scale), and whether or not they deviated from their planned analysis. In case this last question was answered affirmatively, they specified with regard to which aspects they deviated (i.e., hypotheses, included variables, operationalization of the independent variable(s), operationalization of the dependent variable(s), exclusion criteria, statistical test, statistical model, direction of the effect). Finally, teams again reported the experienced effort, frustration, workload in hours and the extent to which this workload was lower or higher than expected for stage 2 (on 7-pt Likert scales).

3.6.4 Procedure

After signing up, participating teams received a document outlining the aim of the project, the timeline, a short theoretical background with respect to the research questions, and a description of the dataset. Then, after completing the pre-survey, teams could access the full data documentation, the questionnaire as presented to the participants of the cross-cultural study, and either a blinded version of the data or a preregistration template, depending on which condition they had been assigned to. Teams could then design their analysis and upload their documents on their own team page on the OSF (deadline: December 22nd, 2020). The project leaders ‘froze’ the stage 1 documents and sent the link to the mid-survey. Upon completion of this survey, teams automatically received access to the real data. They could execute and upload their final analysis scripts on the OSF until February 28th, 2021. Teams were encouraged to also upload a document summarizing their results, but this was not mandatory. Finally, all teams completed the post-survey. See Table 3.2 for an overview of the procedure.

3.7 Results

Here, we report the key results of the project. Specifically, we evaluate the teams’ reported effect sizes and their subjective conclusions about the research questions (i.e., the primary results). In addition, we provide descriptive results about the many-analysts aspect (i.e., the secondary results: variability in analytic approaches, included variables, and the teams’ experiences across the two different stages). Finally, we assessed whether or not the reported effect sizes are related to subjective beliefs about the likelihood of the research questions.

3.7.1 Primary Results

Teams could report any effect size metric of their choosing, but we noted that we preferred a beta coefficient (i.e., a fully standardized coefficient; z-scored predictors and outcomes) to allow for a comparison between teams. As we correctly anticipated that (1) most teams would conduct linear regression analyses (see Table 3.3) and (2) both the (scale of the) independent and dependent variables might vary across teams, we considered a beta coefficient the most suitable effect

Table 3.2: Overview of Project Stages and Resources.

Process	Link
Stage 1	
Recruitment and sign-up	osf.io/hpd6b
Pre-survey	osf.io/kgqze
Access to data documentation, questionnaire and either of:	
a) preregistration form	osf.io/a5ent
b) blinded data	osf.io/ktvqw
Design analysis and upload plan	OSF team pages
Mid-survey	osf.io/kgqze
Stage 2	
Access to data	osf.io/6njsy
Execute analysis and upload script (optional: + report)	OSF team pages
Post-survey	osf.io/kgqze
Lead team: summarize and write-up key results	
Invite analysis teams to write commentary	

Note. See osf.io/vy8z7 for an overview of all team pages.

size metric. Note that our request for beta coefficients as effect size metrics may have affected the teams' choice of statistical model and encouraged them to use regression models that generate beta coefficients. For teams that did not provide a (fully) standardized coefficient, we recalculated the beta based on the respective team's analysis script whenever possible. Specifically, for (multilevel) linear regression models we used the `effectsize` package or the `jtools` package to extract standardized coefficients in R. For analyses in SPSS and non-standard models in R, we standardized the data manually prior to executing the analyses. Finally, many teams reported multiple effect sizes, as they either separately considered multiple predictors (e.g., religious beliefs and religious behaviors) and/or multiple dependent variables (e.g., psychological well-being and physical well-being). In that case, we asked the teams to provide us with one primary effect size they considered most relevant to answer the research question or to select one randomly. In the online appendix, we additionally list (1) effect sizes for the different subscales of the well-being measure as reported by the teams and (2) effect sizes from teams that could not provide a beta coefficient (e.g., machine learning models).

3.7.1.1 Research Question 1: "Do religious people self-report higher well-being?"

We were able to extract 99 beta coefficients from the results provided by the 120 teams that completed stage 2.[5] As shown in Figure 3.2, the results are remarkably

[5] One team misinterpreted the scoring of the items and hence miscoded the direction of the effect. As they subsequently also based their subjective conclusions on the incorrect results,

consistent: all 99 teams reported a positive beta value, and for all teams the 95% confidence/credible interval excludes zero. The median reported beta is 0.120 and the median absolute deviation is 0.036. Furthermore, 88% of the teams concluded that there is good evidence for a positive relation between religiosity and self-reported well-being. Notably, although the teams were almost unanimous in their evaluation of research question 1, only eight of the 99 teams reported combinations of effect sizes and confidence/credibility intervals that matched those from another team (i.e., four effect sizes were reported twice). Do note that in contrast to the unanimity in results based on the beta coefficients, out of the 21 teams for whom a beta coefficient could not be calculated, 3 teams reported evidence against the relation between religiosity and well-being: 2 teams used machine learning and found that none of the religiosity items contributed substantially to predicting well-being and 1 team used multilevel modeling and reported unstandardized gamma-weights for within- and between-country effects of religiosity whose confidence intervals included zero (see the online appendix).

Figure 3.3 displays the average prior and final beliefs about the likelihood of the hypothesis. Researchers' prior beliefs about religiosity being positively related to self-reported well-being were already high (M = 4.90 on the 7-point Likert scale), but were raised further after them having conducted the analysis (M = 5.49 on the 7-point Likert scale). Specifically, before seeing the data, 72% of the teams considered it likely that religiosity is related to higher self-reported well-being. This percentage increased to 85% after having seen the data, while 11% were neutral and 3% considered it unlikely. Finally, 75% of teams indicated the relation to be relevant or meaningful based on these data.

3.7.1.2 Research Question 2: "Does the relation between religiosity and self-reported well-being depend on perceived cultural norms of religion?"

Out of the 120 teams who completed stage 2 we were able to extract 101 beta coefficients for research question 2. As shown in Figure 3.4 the results for research question 2 are more variable than for research question 1; 97 out of 101 teams reported a positive beta value and for 66 teams (65%) the confidence/credible interval excluded zero. The median reported effect size is 0.039 and the median absolute deviation is 0.022. Furthermore, 54% of the teams concluded that there is good evidence for an effect of cultural norms on the relation between religiosity and self-reported well-being. Again, most reported effect sizes were unique; only 3 out of the 101 reported combination of effect size and confidence/credible intervals appeared twice.

Figure 3.5 shows the researchers' average prior and final beliefs about the likelihood of the second hypothesis. As for research question 1, prior beliefs about the hypothesis were rather high. However, in contrast to research question 1, conducting the analysis lowered beliefs about the likelihood of hypothesis 2. Specifically, before seeing the data, 71% of the teams considered it likely that the relation between religiosity and self-reported well-being depends on perceived cultural norms

we excluded the reported effect sizes, subjective evaluation, and prior+final beliefs about the likelihood of the hypotheses for this team.

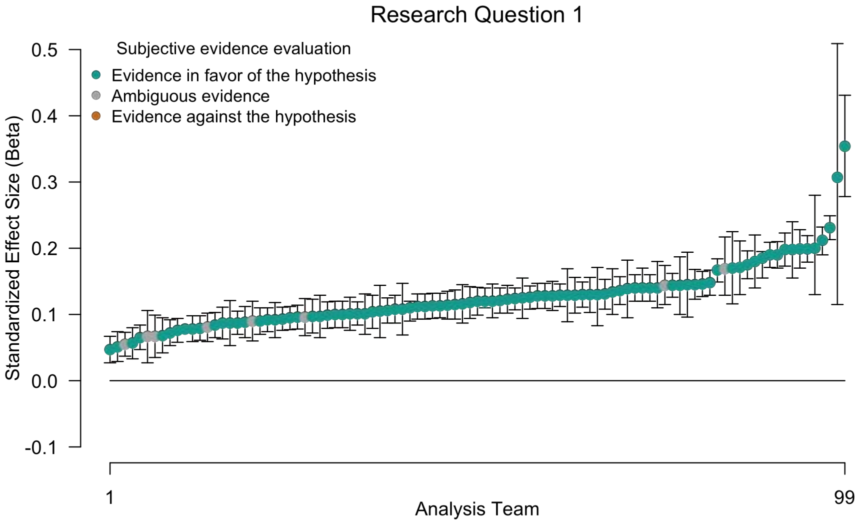


Figure 3.2: Beta coefficients for the effect of religiosity on self-reported well-being (research question 1) with 95% confidence or credible intervals. Green/blue points indicate effect sizes of teams that subjectively concluded that there is *good evidence for a positive relation* between individual religiosity and self-reported well-being, grey points indicate effect sizes of teams that subjectively concluded that *the evidence is ambiguous*, and brown/orange points indicate effect sizes of teams that subjectively concluded that there is *good evidence against a positive relation* between individual religiosity and self-reported well-being. The betas are ordered from smallest to largest.

of religion. This percentage dropped to 54% after having seen the data, while 19% were neutral and 27% considered it unlikely. Finally, only about half of the teams (49%) indicated the effect of cultural norms to be relevant or meaningful based on these data.

3.7.2 Secondary Results

In addition to evaluating the overall results for the two main research questions, we also assessed perceived suitability of the data and analytic approaches, variability in analytical approaches (i.e., statistical models), variable inclusion, and teams' experiences during the two stages of the project.

3.7.2.1 Perceived Suitability of Dataset

At the end of the project, all teams reported how suitable they found the current dataset for answering the research questions. As shown at the top of Figure 3.6, most teams considered the data (very) suitable for answering the research questions: for research question 1, 86% found the data suitable, 8% neutral, and 6% unsuitable; for research question 2, 70% found the data suitable, 19% neutral, and 11% unsuitable.

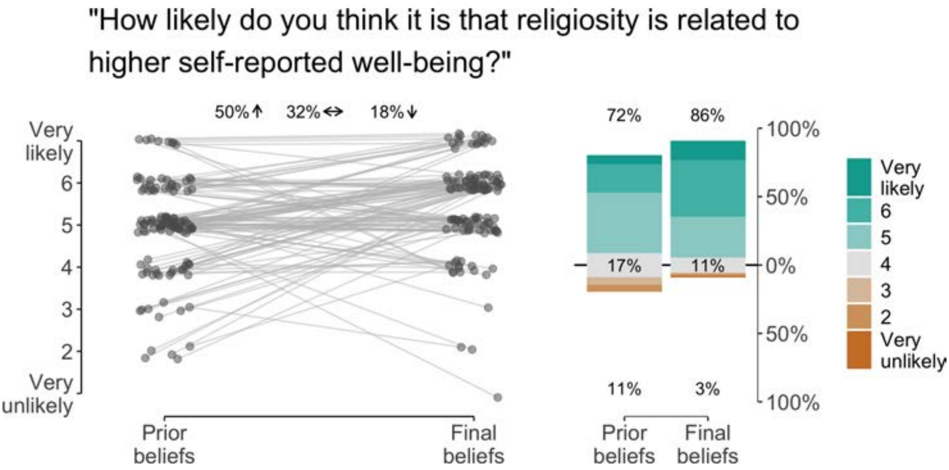


Figure 3.3: Responses to the survey questions about the likelihood of hypothesis 1. The left side of the figure shows the change in beliefs for each analysis team. Fifty percent of the teams considered the hypothesis somewhat more likely after having analyzed the data than prior to seeing the data, 18% considered the hypothesis less likely after having analyzed the data, and 32% did not change their beliefs. Likelihood was measured on a 7-point Likert scale ranging from 'very unlikely' to 'very likely'. Points are jittered to enhance visibility. The right side of the figure shows the distribution of the Likert response options before and after having conducted the analyses. The number at the top of the data bar (in green/blue) indicates the percentage of teams that considered the hypothesis (very) likely, the number in the center of the data bar (in grey) indicates the percentage of teams that were neutral, and the number at the bottom of the data bar (in brown/orange) indicates the percentage of teams that considered the hypothesis (very) unlikely.

3.7.2.2 Analytic Approaches

Table 3.3 displays the different statistical approaches used in the project, as well as the percentage of teams that employed the respective approach. While a total of 25 different statistical methods was mentioned, (multilevel) linear regression was clearly the dominant approach. Specifically, 34% of the teams used linear regression, another 45% used multilevel linear regression, and the remaining 21% used a different approach.

In general, teams were confident that their chosen statistical approach was appropriate for analyzing the research questions; as shown at the bottom of Figure 3.6, 89% of the teams indicated to be (very) confident, 4% was neutral, and 7% was not (at all) confident.[6]

[6] Note that out of the 8 teams reporting not being confident, 2 did not submit a final analysis and 2 did not provide a usable effect size.

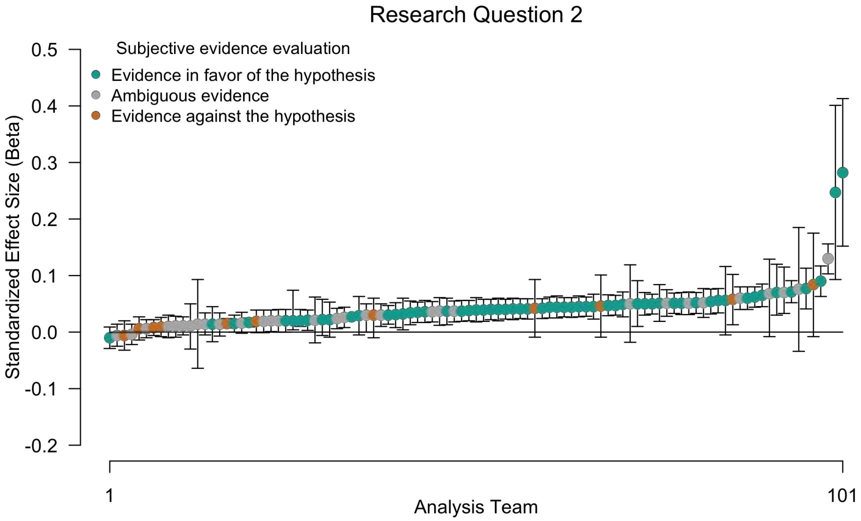


Figure 3.4: Beta coefficients for the effect of cultural norms of the relation between religiosity and self-reported well-being (research question 2) with 95% confidence or credible intervals. Green/blue points indicate effect sizes of teams that subjectively concluded that there is *good evidence for the hypothesis* that the relation between individual religiosity and self-reported well-being depends on the perceived cultural norms of religion, grey points indicate effect sizes of teams that subjectively concluded that *the evidence is ambiguous*, and brown/orange points indicate effect sizes of teams that subjectively concluded that there is *good evidence against the hypothesis* that the relation between individual religiosity and self-reported well-being depends on the perceived cultural norms of religion. The betas are ordered from smallest to largest.

3.7.2.3 Variable Inclusion

For each team we coded which of the items provided in the dataset were included as (1) dependent variable, (2) independent variable, and (3) covariates in the analysis for each research question.[7]

Dependent Variable The subjective well-being measure consisted of three subscales (psychological, physical, social), as well as two general items. In the dataset, we provided responses for all 18 individual items as well as an overall mean and one mean for each of the three subscales. Teams could decide to either use any of the provided averages or combine specific items themselves (e.g., take the mean, median, sum). In addition, some teams conducted a factor analysis and used one or multiple extracted factors as the dependent variable. In this case, we coded which items were used as input for the factor analysis. Figure 3.7 shows the included items as dependent variable aggregated over all teams for research question

[7] Please see the document 'variable mapping' on the OSF (osf.io/qbdce/) for how the items correspond to the item names in the datafile.

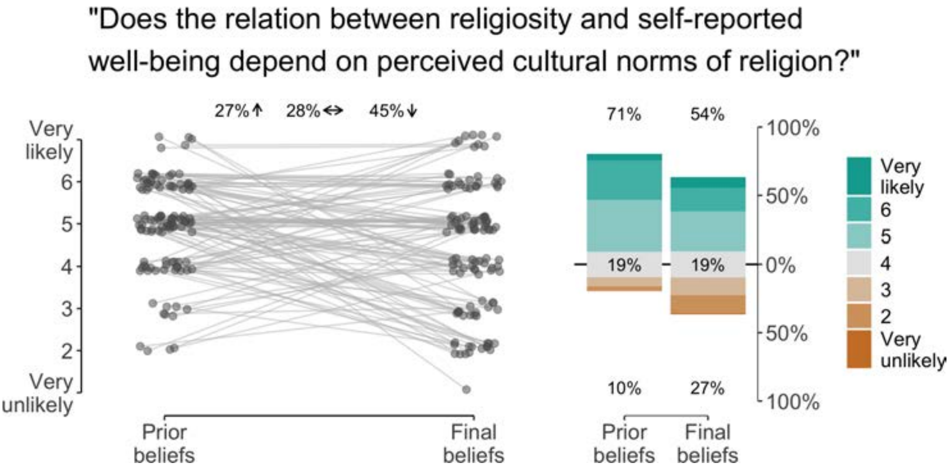


Figure 3.5: Responses to the survey questions about the likelihood of hypothesis 2. The left side of the figure shows the change in beliefs for each analysis team. Twenty-seven percent of the teams considered the hypothesis somewhat more likely after having analyzed the data than prior to seeing the data, 45% considered the hypothesis less likely having analyzed the data, and 28% did not change their beliefs. Likelihood was measured on a 7-point Likert scale ranging from 'very unlikely' to 'very likely'. Points are jittered to enhance visibility. The right side of the figure shows the distribution of the Likert response options before and after having conducted the analyses. The number at the top of the data bar indicates the percentage of teams that considered the hypothesis (very) likely, the number in the center of the data bar (in grey) indicates the percentage of teams that were neutral, and the number at the bottom of the data bar (in brown/orange) indicates the percentage of teams that considered the hypothesis (very) unlikely.

1 and research question 2. For research question 1, the most frequently used items are *enjoying life* and *meaningfulness* (included by over 43% of the teams). Note that all but four teams used the same dependent variable for research question 1 and 2.[8] In the online appendix, we show the included items separately for each team.

Independent Variable The religiosity measure consisted of 9 primary items on response scales ranging from dichotomous to 8-points and the cultural norms of religiosity measure consisted of two items on a 5-point scale. Averages were not provided in the dataset, but could be created by the teams themselves. Figure 3.8 shows the included items as independent variable aggregated over all teams for research question 1 and research question 2. In online appendix, we show the included items separately for each team.

[8] Two of the four teams that did not use the same dependent variable for research question 1 and 2 only conducted an analysis for research question 1.

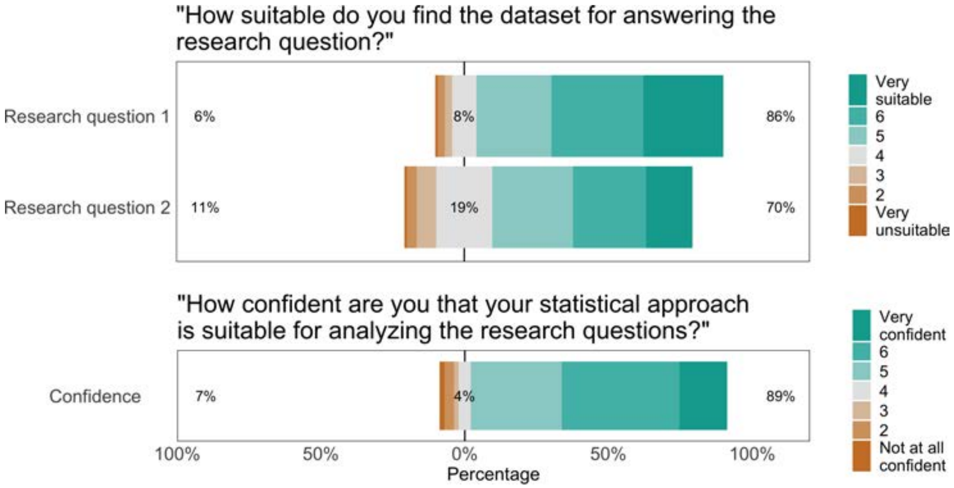


Figure 3.6: Responses to the survey questions about the suitability of the dataset for answering the research questions (top) and the teams' confidence in their analytic approach (bottom). For question 1, the top bar represents the teams' answers with respect to research question 1 and the bottom bar represents the teams' answers for research question 2. For each item, the number to the left of the data bar (in brown/orange) indicates the percentage of teams that considered the data (very) unsuitable / were not (at all) confident in their approach. The number in the center of the data bar (in grey) indicates the percentage of teams that were neutral. The number to the right of the data bar (in green/blue) indicates the percentage of teams that considered the data (very) suitable / were (very) confident in their approach.

For research question 1 (i.e., the relation between religiosity and self-reported well-being), over 75% of the teams operationalized the independent variable by including the items *frequency of service attendance, belief in God/Gods, frequency of prayer, belief in afterlife, personal importance of a religious lifestyle*, or *personal importance of belief in God.* The remaining three religiosity items were used less frequently: 70% of the teams included the item *religious status (religious/not religious/atheist)* and *spirituality*, while only 50% included *religious membership.*

For research question 2 (i.e., the effect of perceived cultural norms on the relation between religiosity and self-reported well-being), all but four teams used the interaction term between their chosen religiosity measure and their chosen cultural norms measure as the independent variable.[9] More teams operationalized cultural norms using the item *importance of a religious lifestyle in their country* (93%) than *importance of belief in God/Gods in their country* (89%). Here again, over 75% of the teams operationalized the independent variable by including the items *frequency of service attendance, belief in God, frequency of prayer, belief*

[9] The four teams that did not use an interaction in their evaluation of research question 2 either used the main effect of cultural norms on well-being or the main effect of religiosity on well-being (while controlling for cultural norms).

Table 3.3: Analytic Approaches Taken by the Analysis Teams.

Analytic Approach	Percentage of teams
Multilevel Linear Regression	45/128 (35.16 %)
Linear Regression	36/128 (28.12 %)
Bayesian Multilevel Linear Regression	7/128 (5.47 %)
Structural Equation Model	6/128 (4.69 %)
ANOVA	5/128 (3.91 %)
T-test	4/128 (3.12 %)
Bayesian Linear Regression	3/128 (2.34 %)
Path Analysis	3/128 (2.34 %)
Bayesian Multilevel Ordinal Regression	2/128 (1.56 %)
Ordinal Logistic Regression	2/128 (1.56 %)
ANCOVA	1/128 (0.78 %)
Bayesian Additive Regression Trees	1/128 (0.78 %)
Bayesian ANOVA	1/128 (0.78 %)
Bayesian Multilevel Structural Equation Model	1/128 (0.78 %)
Correlation	1/128 (0.78 %)
Machine Learning	1/128 (0.78 %)
Meta-Analysis	1/128 (0.78 %)
Mixed-Effects ANOVA	1/128 (0.78 %)
Moderated Generalized Linear Regression	1/128 (0.78 %)
Multilevel Structural Equation Model	1/128 (0.78 %)
Multiverse Analysis	1/128 (0.78 %)
Multiverse Of Multilevel Linear Regression	1/128 (0.78 %)
Network Analysis	1/128 (0.78 %)
Non-linear Regression	1/128 (0.78 %)
Non-parametric Partial Correlation	1/128 (0.78 %)

Note. Some teams reported multiple statistical approaches.

in afterlife, personal importance of a religious lifestyle, or *personal importance of belief in God*, whereas the items *religious status (religious/not religious/atheist)* and *spirituality* were included by about 70% and 68% of the teams, respectively; only 52% of the teams included *religious membership.* Note that almost all teams used the same religiosity measure for research question 1 and research question 2.

Covariates Teams were free to include as covariates in their models any of the measured demographic variables (e.g., age, socio-economic status), country-level variables (e.g., gross domestic product – GDP) or sample characteristics (e.g., general public or student sample, means of compensation). Figure 3.9 displays the included items as covariates aggregated over all teams for research question 1 and research question 2. The most frequently included covariates are *age* (59%), *socio-economic status* (55%), *gender* (53%), and *education* (50%). Note that per team the choice of covariates was largely equal across research questions, with the

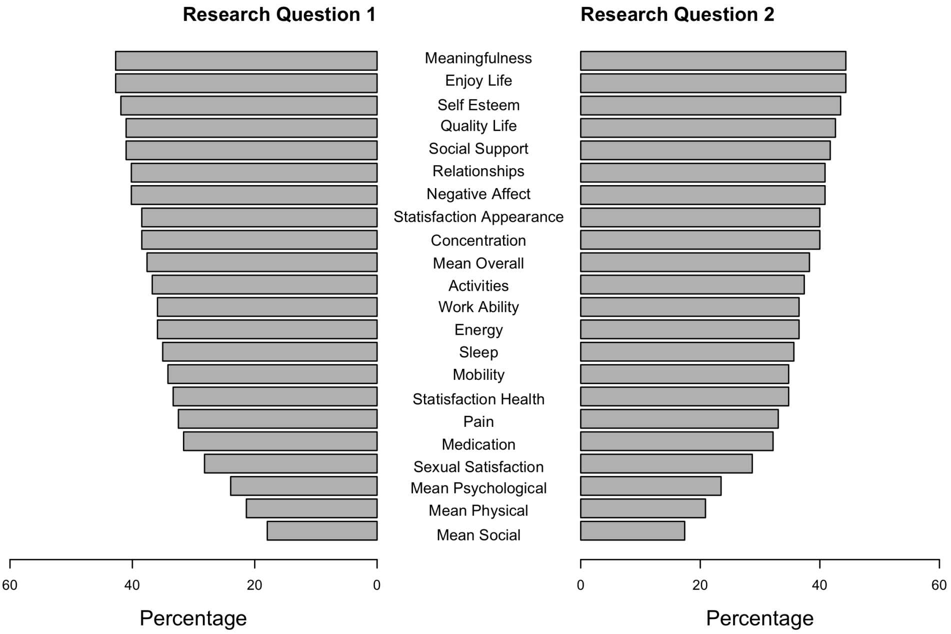


Figure 3.7: Items included as dependent variables for research question 1 (on the left) and research question 2 (on the right). Note that the averages for the well-being subscales ('Mean Psychological', 'Mean Social', 'Mean Physical'), as well as the overall average ('Mean Overall') were provided by the MARP team.

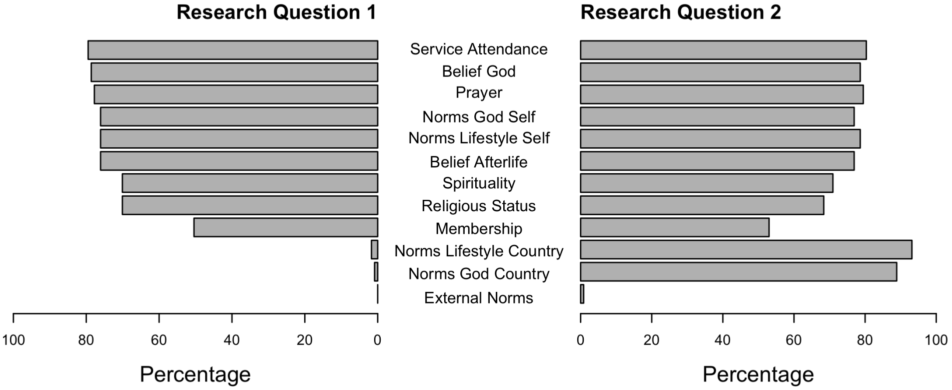


Figure 3.8: Items included as independent variables for research question 1 (on the left) and research question 2 (on the right).

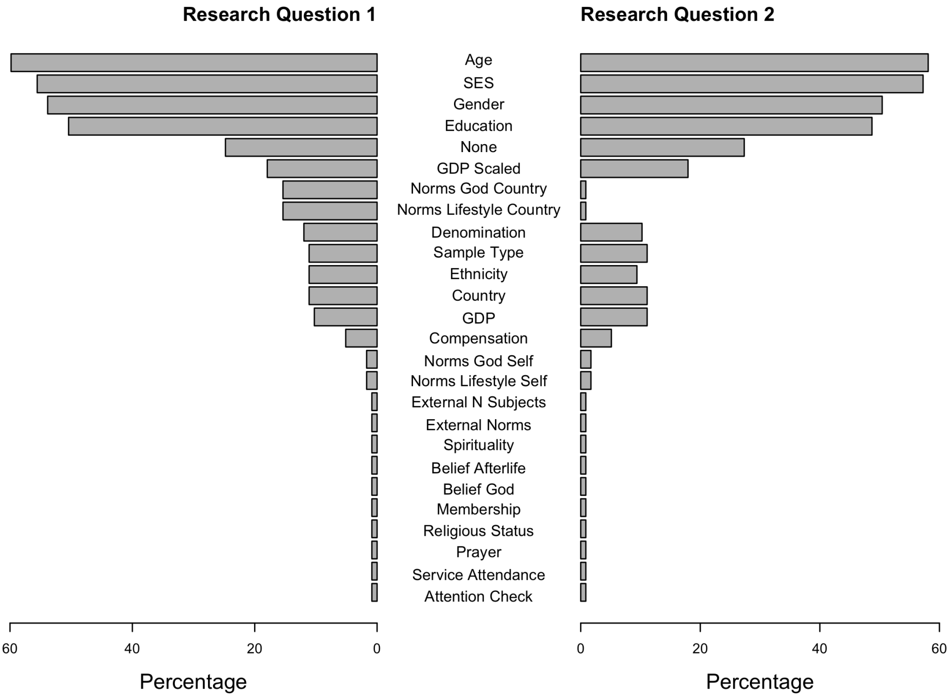


Figure 3.9: Items included as covariates for research question 1 (on the left) and research question 2 (on the right). Variables indicated as 'external' refer to covariates that are based on data not provided by the MARP team.

exception that the cultural norms items were occasionally added as covariates for research question 1 while they were part of the independent variable for research question 2.

3.7.2.4 Teams' Experiences

Although most teams indicated that effort was (very) high, the majority also reported that frustration was (very) low and that they spent as much time as anticipated (see Figure 3.10). That is, in stage 1, 55% of the teams reported (very) high effort, 17% were neutral, and 28% reported (very) low effort. For stage 2, 48% of the teams reported (very) high effort, 18% were neutral, and 34% reported (very) low effort. In stage 1, 17% of the teams reported (very) high frustration, 23% were neutral, and 60% reported (very) low frustration. In stage 2, 18% of the teams reported (very) high frustration, 17% were neutral, and 65% reported (very) low frustration. The median time spent on the analyses was 8 hours for both stages, although the range was quite wide: 1 to 80 hours for stage 1 and 30 minutes to 140 hours for stage 2. Most teams anticipated as much time as they spent: 51% for stage 1 and 52% for stage 2. In stage 1, 36% spent (much) more time than anticipated and 13% spent (much) less time. In stage 2, 33% spent (much) more time than anticipated and 15% spent (much) less time.

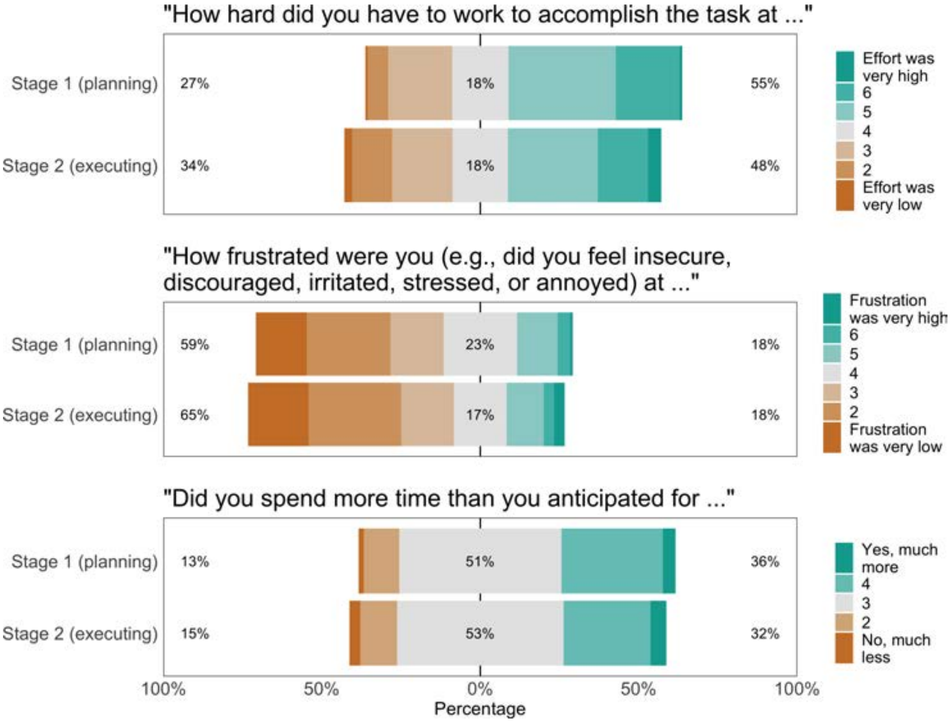


Figure 3.10: Responses to the survey questions about effort (top), frustration (middle), and workload (bottom). For each question, the top bar represents the teams' answers about stage 1 (planning) and the bottom bar represents the teams' answers about stage 2 (executing). For each item, the number to the left of the data bar (in brown/orange) indicates the percentage of teams that considered effort/frustration/workload (very) low. The number in the center of the data bar (in grey) indicates the percentage of teams that were neutral. The number to the right of the data bar (in green/blue) indicates the percentage of teams that considered effort/frustration/workload (very) high.

3.7.2.5 Correlation between Effect Sizes and Subjective Beliefs

Following Silberzahn et al. (2018) we explored whether the reported effect sizes were positively related to subjective beliefs about the plausibility of the research question *before* and *after* analyzing the data. This hypothesis was tested against the null-hypothesis that there is no relation between reported effect sizes and subjective beliefs. As the subjective beliefs were measured on a 7-point Likert scale, we used a rank-based Spearman correlation test with a Uniform[0, 1] prior (van Doorn, Ly, Marsman, & Wagenmakers, 2020).

For research question 1, we obtained strong evidence *against* a positive relation between prior beliefs about the plausibility of the research question and the reported effect sizes: $\text{BF}_{+0} = 0.03$; $\text{BF}_{0+} = 30.34$, $\rho_s =$ -0.21, 95% credible interval [-0.37, -0.04]. In addition, we found moderate evidence against a positive relation between posterior beliefs about the plausibility of the research question and the reported effect sizes: $\text{BF}_{+0} = 0.31$; $\text{BF}_{0+} = 3.18$, $\rho_s = 0.10$, 95% credible interval [-0.08, 0.27].

For research question 2, we found moderate evidence against a positive relation between prior beliefs about the plausibility of the research question and the reported effect sizes: $\text{BF}_{+0} = 0.12$; $\text{BF}_{0+} = 8.55$, $\rho_s = 0.01$, 95% credible interval [-0.16, 0.18]. For the posterior beliefs, however, we obtained strong evidence in favor of a positive relation between posterior beliefs about the plausibility of the research question and the reported effect sizes: $\text{BF}_{+0} = 67.39$, $\rho_s = 0.33$, 95% credible interval [0.15, 0.46].

To further investigate changes in belief over the course of the project, we assessed the correlation between the reported effect sizes and the change in belief (i.e., the difference between posterior and prior beliefs for both research questions). For research question 1, there was basically no evidence for or against a positive relation between effect size and change in belief: $\text{BF}_{+0} = 1.81$, $\rho_s = 0.18$, 95% credible interval [0.01, 0.33]. For research question 2 on the other hand, we obtained moderate evidence that effect sizes were positively related to change in subjective belief about the plausibility of the hypothesis: $\text{BF}_{+0} = 9.88$, $\rho_s = 0.24$, 95% credible interval [0.07, 0.39].

These results regarding prior beliefs provide no indication that expectations and confirmation bias influenced the teams' results. For the posterior beliefs, on the other hand, it seems that the teams updated their beliefs about the plausibility of research question 2 based on the results of their analyses. Note, however, that based on the scatterplot in Figure 3.11D, we should not put too much weight on this finding, as it may be partly driven by two outliers. For research question 1, the updating of beliefs may not have happened because prior beliefs about research question 1 were already in line with the outcomes, i.e., most teams expected and reported evidence for a positive relation between religiosity and well-being, with little variation between teams.

Finally, we assessed whether reported effect sizes were related to self-reported expertise. Here, we used a Uniform$[-1,\ 1]$ prior and an undirected test. This hypothesis was tested against the null-hypothesis that reported effect sizes and self-reported expertise were not related. For research question 1, we found moderate evidence against a correlation between effect sizes and methodological knowledge ($\text{BF}_{10} = 0.13$; $\text{BF}_{01} = 7.80$, $\rho_s = 0.03$, 95% credible interval [-0.17, 0.21]) and weak evidence against a correlation between effect sizes and theoretical knowledge ($\text{BF}_{10} = 0.48$; $\text{BF}_{01} = 2.09$, $\rho_s =$ -0.16, 95% credible interval [-0.31, 0.03]). For research question 2, we again obtained moderate evidence against a relation between effect sizes and methodological knowledge ($\text{BF}_{10} = 0.12$; $\text{BF}_{01} = 8.00$, $\rho_s = 0.02$, 95% credible interval [-0.17, 0.20]) and moderate evidence against a correlation between effect sizes and theoretical knowledge ($\text{BF}_{10} = 0.16$; $\text{BF}_{01} = 6.41$, $\rho_s =$ -0.08, 95% credible interval [-0.24, 0.09]). See Figure 3.12 for scatterplots of the data.

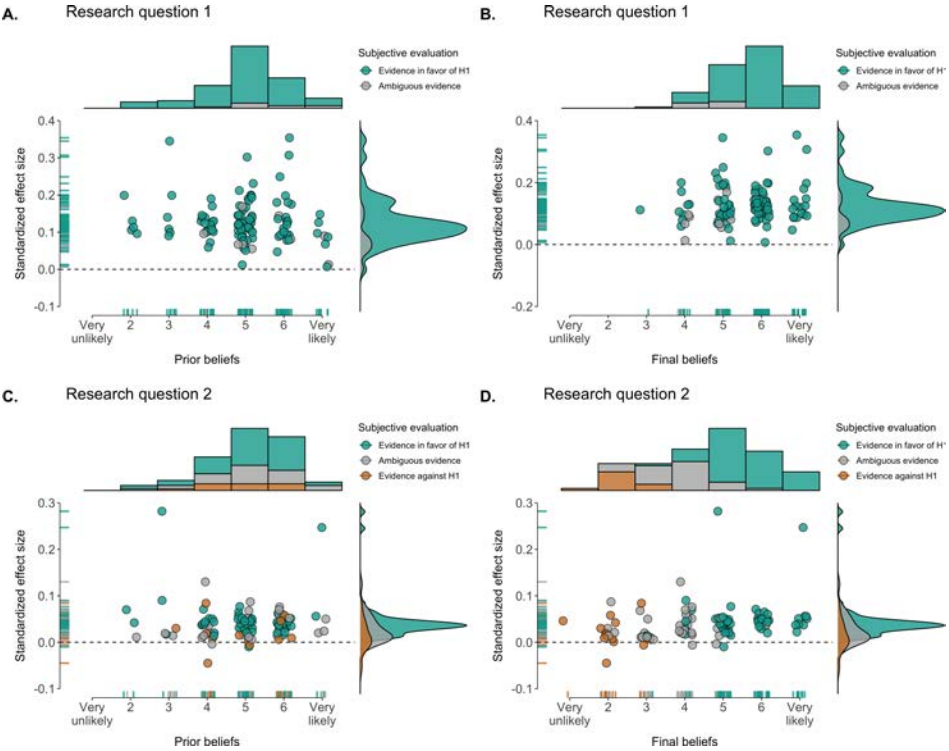


Figure 3.11: Reported effect sizes (beta coefficients) and subjective beliefs about the likelihood of the hypothesis. **A.** shows the relation between effect size and prior beliefs for research question 1, **B.** shows the relation between effect size and final beliefs for research question 1, **C.** shows the relation between effect size and prior beliefs for research question 2, and **D.** shows the relation between effect size and final beliefs for research question 2. Points are jittered on the x-axis to enhance visibility. The dashed line represents an effect size of 0. The data are separated by subjective evaluation of the evidence; green/blue points reflect the conclusion that there is good evidence for the hypothesis, grey points reflect the conclusion that the evidence is ambiguous, and brown/orange points indicate the conclusion that there is good evidence against the hypothesis. Histograms at the top represent the distribution of subjective beliefs and the density plots on the right represent the distribution of reported effect sizes.

3.8 Summary

In the current project, 120 analysis teams were given a large cross-cultural dataset (N = 10,535, 24 countries) in order to investigate two research questions: (1) “Do religious people self-report higher well-being?” and (2) “Does the relation between religiosity and self-reported well-being depend on perceived cultural norms of religion?”. In a two-stage procedure, the teams first proposed an analysis and then executed their planned analysis on the data.

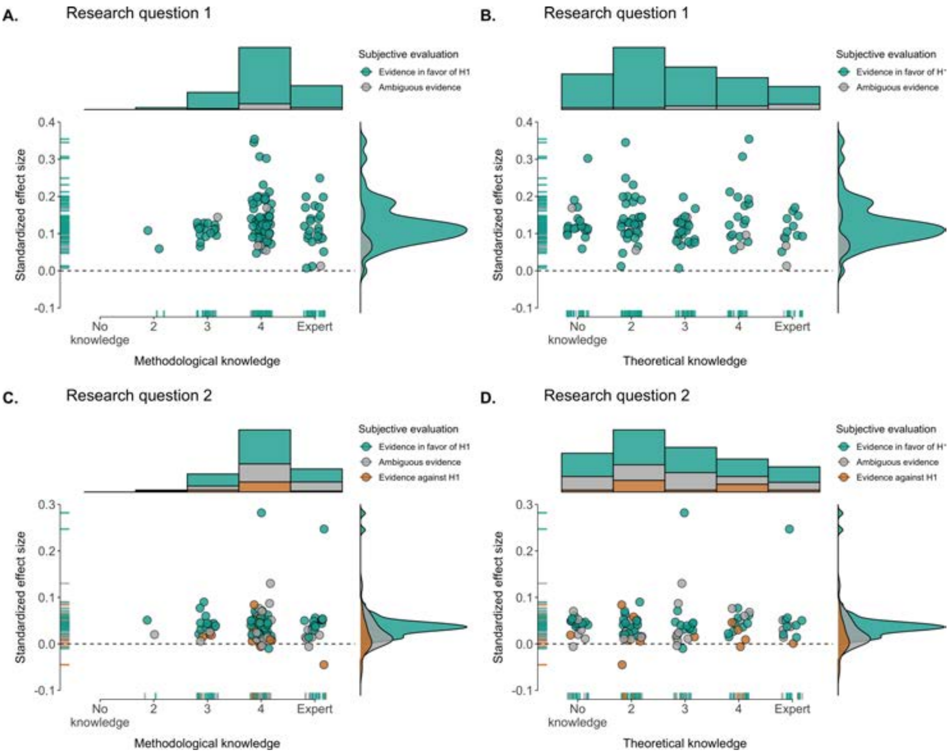


Figure 3.12: Reported effect sizes (beta coefficients) and self-reported team expertise. **A.** shows the relation between effect size for research question 1 and methodological knowledge, **B.** shows the relation between effect size for research question 1 and theoretical knowledge, **C.** shows the relation between effect size and for research question 2 and methodological knowledge, and **D.** shows the relation between effect size for research question 2 and theoretical knowledge. Points are jittered on the x-axis to enhance visibility. The dashed line represents an effect size of 0. The data are separated by subjective evaluation of the evidence; green/blue points reflect the conclusion that there is good evidence for the hypothesis, grey points reflect the conclusion that the evidence is ambiguous, and brown/orange points indicate the conclusion that there is good evidence against the hypothesis. Histograms at the top represent the distribution of reported expertise and the density plots on the right represent the distribution of reported effect sizes.

Perhaps surprisingly in light of previous many-analysts projects, results were fairly consistent across teams. For research question 1 on the relation between religiosity and self-reported well-being, all but three teams reported a positive effect size and confidence/credible intervals that exclude zero. For research question 2, the results were somewhat more variable: 95% of the teams reported a positive effect size for the moderating influence of cultural norms of religion on the association between religiosity and self-reported well-being, with 65% of the confidence/credible intervals excluding zero. While most teams used (multilevel) linear regression, there was considerable variability in the choice of items used to construct the independent variable, the dependent variable, and the included covariates.

A further discussion of these results including limitations and broader implications, as well as a reflection on the many-analysts approach is covered in the closing article (Hoogeveen et al., in preparation). There, we also address the commentaries written by some of the analysis teams.

Chapter 4

Many-Analysts Religion Project: Reflection and Conclusion

This chapter is in press as Hoogeveen, E. J., Sarafoglou, A., van Elk, M., & Wagenmakers, E. J. (2022). Many-Analysts Religion Project: Reflection and Conclusion. *Religion, Brain & Behaviour.*

4.1 Introduction

In the main article on the Many-Analysts Religion Project (MARP) the results of the 120 analysis teams were summarized by taking each team's reported effect size and subjective assessment of the relation between religiosity and well-being, and the moderating role of cultural norms on this relation (The MARP Team, 2022). The many-analysts approach allowed us to appraise the uncertainty of the outcomes, which has been identified as one of the pillars of good statistical practice (Wagenmakers et al., 2021). A downside of this approach, however, is that a fine-grained consideration of the details and nuances of the results becomes difficult. Summaries of the individual approaches are documented in the teams' OSF project folders, but time and space did not permit the inclusion of details on each of the individual analysis pipelines in the main article.

However, we believe the scope of the project and the effort of the analysis teams justifies highlighting some more in-depth observations. Here, we aim to address these supplementary findings, taking the points raised in the 17 commentaries written by various participating analysts as a guideline. We identified three overarching themes in the commentaries and our own experiences. First, there was a need for more focus on theoretical depth and specificity. We refer to this aspect as "zooming in". Second, multiple commentaries reflected on the broader implications of our results, elaborating on robustness and (the limits of) generalizability. We refer to this aspect as "zooming out". Third, several commentaries addressed the appropriateness of the analysts' chosen statistical models given the MARP data.

In the following sections, we will first zoom in and address the issue of theoretical specificity. We will then zoom out and discuss to what extent the MARP results are robust and can be generalized. Subsequently, we discuss some methodological concerns, mostly related to the structure of the data. Finally, we will reflect on our experience of organizing a many-analysts project and highlight some lessons learned.

4.2 Zooming In: Theoretical Specificity

The broad setup of the project inspired some analyst teams to dive deeper into the data themselves in order to offer more nuanced interpretations and test additional hypotheses (e.g., Atkinson et al.; Murphy and Martinez; Pearson, Lo, and Sasaki; E. Smith; Vogel et al.). Others, however, criticized the lack of specificity and questioned whether the current setup has led to valid results. Specifically, some authors argued that the broad formulation of the MARP research questions allowed for different interpretations, thereby contributing to analytic flexibility and undesirable heterogeneity (Edelsbrunner et al.; Krypotos, Klein, and Jong; Murphy and Martinez). For instance, the first research question "Do religious people report higher well-being" might be understood as a causal effect or an observational effect, which also has consequences for the inclusion of covariates (Edelsbrunner et al.). The authors called for more specific research questions in terms of the type of effect, the structure of the data, and the level of analysis that is of focal interest.

This concern was echoed by Murphy and Martinez, who argued that it is more meaningful to ask which specific behaviors benefit certain well-being markers for a specific population (e.g., “Does belief in God lead to a more meaningful life, when controlling for the influence of socioeconomic status?”).

Similarly, Bulbulia emphasized the need for researchers to clearly specify the outcome, the exposure, the contrasts, and the study design, in order to address the causal questions of interest. Bulbulia showed that model-free inferences might lead to implausible conclusions, such as that anxiety reduces service attendance. Instead, the author demonstrates the advantage of the application of causal modelling that yields alternative interpretations which are supported both by the data and existing theories of religion (i.e., service attendance buffers anxiety). We believe this approach to causal inference for observational data is an important future direction and think the workflow outlined by Bulbulia may serve as an example.

At the same time, other analysts suggested that the setup of the project was in fact too constrained. For instance, Vogel et al. argued that our request to provide only one effect size per research question may have led different teams to converge toward the same operationalizations. Specifically, this setup may have implicitly encouraged teams to focus on the broadest operationalizations possible and discouraged teams to investigate the multifaceted nature of both religiosity and well-being.

We acknowledge that the broad specification of the research questions may have caused some confusion and/or promoted the use of the global indices instead of specific items for the teams’ analyses. However, the lack of specificity was to some extent intentional. Precisely because of the multifaceted nature of religiosity and well-being and the different operationalizations found in the literature, we did not want to restrict the researchers’ interpretation of these constructs (beyond the limits of what the dataset contained). And indeed, the MARP results were largely robust against the different analytic choices, suggesting that the exact operationalization does not matter for the robustness of the general relationship. At the same time, as pointed out in the commentaries, this approach leaves open which aspects of religiosity specifically contribute to which aspects of well-being.

Here, we highlight some notable examples of more in-depth observations that provide insight into the specificity of the religion–well-being relationship. First, based on the follow-up analyses carried out by 19 teams, it appears that religiosity is most strongly related to psychological well-being, followed by social well-being and not so much to physical well-being. Vogel et al. found that two items of the physical well-being subscale, namely ‘pain’ and ‘dependence on medical treatment’, were in fact negatively related to religiosity. Atkinson et al. similarly showed that these two items and ‘mobility’ were not predicted by religiosity. Second, E. Smith distinguished between the role of cultural norms at the individual and at the country level: they found no moderation of cultural norms of religion at the individual level (i.e., “individuals who see their country as more religious than other individuals in the same country do not benefit more from being religious”) but a strong effect at the country-level (i.e., “individuals in countries that are on average perceived as more religious benefit more from being religious than individuals in countries where religion is less normative”). Third, Pearson et al. further investigated the cultural match hypothesis, by assessing to what extent the cultural dimension of tightness-looseness and multiculturalism moderate the influence of cultural norms on the relation between individual religiosity and well-being. Drawing on additional country-level data, they found that the influence of religiosity on psychological well-being may be greater when people perceive their country to be more religious, but more so when that country is culturally tighter. Fourth, Murphy and Martinez showed that two theoretically defensible choices of operationalizing religiosity (e.g., Paloutzian, 2017) did not result in significantly different outcomes; there was no difference in effect sizes between using a composite measure of beliefs, practices, values, and identification or a single-item self-identification measure (i.e., religious, non-religious, or atheist).

4.3 Zooming Out: Generalizability and Robustness

We believe that the comprehensiveness of the MARP data, which featured a large number of participants, countries, and religious denominations, leads to conclusions that are generalizable to other populations (e.g., new samples from the included countries, samples from other countries). Moreover, the variety of statistical strategies and the consistency of the main results suggest that the outcomes are robust against statistical decisions made by a different sample of analysis teams.

In addition, Atkinson et al. discussed how generalizability can be explored within a certain analysis, for instance by either including an extensive random effects structure or by applying cross-validation techniques. The authors found that the results were overall stable, but also report some limits on generalizability. That is, religiosity was not related to pain, medical dependence, and mobility (as noted by Vogel et al. as well). Furthermore, including the covariates age, socioeconomic status, and education were necessary to optimize the model fit across different partitions of the data.

Two commentaries discussed the promise of multiverse analyses as an alternative way to assess uncertainty and robustness (Hanel and Zarzeczna; Krypotos et al.). When conducting a multiverse analysis, a research team does not execute one analysis to the data set, but rather the set of all plausible analysis pipelines. The main advantage of multiverse analyses over the many-analysts approach is that they allow for a systematic investigation over the entire decision space, without relying on the involvement of many different researchers.

At the same time, a multiverse still requires theoretically-influenced decisions as typically only one aspect (e.g., variable construction) can be systematically varied while others are fixed (e.g., statistical model and data preprocessing). This restriction is due to both limits on interpretability and practical feasibility (i.e., it takes too much time and processing power to include the entire range of all combinations). The analysis reported by Hanel and Zarzeczna illustrates the limits of a multiverse. The authors examined the effects of *all possible* operationalizations of well-being and religiosity on the results, totaling more than $260,000$ analysis pipelines. Not only were certain aspects of the analysis fixed (e.g., a simple correlation was used without covariates), but the authors also executed the analysis on only a subset of the data because analysing the entire data set was too time consuming. A notable outcome of the multiverse analysis was that the well-being item measuring meaningfulness had the strongest impact on the results, which resonates well with the observations from Vogel et al.).

A promising avenue might be to combine the advantages of multiverse analysis and the many-analysts approaches (i.e., comprehensiveness and theoretical + methodological expertise) in a hybrid format. Instead of a full multiverse that may include implausible paths, Krypotos et al. proposed that an expert panel decides on theoretically motivated restrictions on the analyses and the aspects that require systematic investigation. We believe that this approach could be beneficial for many-analysts projects for which (1) the research question has no strong theoretical boundaries in terms of the operationalization of variables and modeling approach (thus resulting in a multitude of possible analyses), (2) the goal is to investigate the impact of specific items (e.g., covariates) on the relationship, or (3) the pool of qualified analysts is relatively small.

Another method to investigate the relative impact of specific items was discussed by van Lissa. The author applied machine learning techniques to identify the strongest predictors of well-being in the MARP data. They found that socioeconomic status strongly outperformed religiosity as a predictor for well-being; a result that is consistent with that of another team that applied machine learning.[1] The goal of the MARP was not to optimise predictions but to explore a theory and replicate evidence for an existing framework. However, we believe that machine learning techniques, in addition to the interpretation of effect sizes and the subjective judgments of the teams, could be a useful tool in future studies, for instance in determining which features (e.g., what aspects of religiosity) predict well-being best.

In addition to investigating the robustness and generalizability of the current dataset, Himawan, Martoyo, Himawan, Aditya, and Suwartono reviewed whether the MARP results apply to other contexts. Specifically, they provided insight into the results with respect to the Indonesian population. In the same spirit, Islam and Lorenz offered a suggestion to further extend future projects: many analysts analysing many data sets. In such an approach, analysts would be provided with data collected from different projects. This way, generalizability across measures and samples can be assessed. Alternatively, such external data could complement the MARP data. For instance, Islam and Lorenz explored the inclusion of external data on religious majorities as a covariate or moderator in the analysis on the MARP data. (They found no effect, suggesting that well-being does not depend on the match between one's own religion and that of the majority in one's country.)[2] This approach is worth pursuing in future many-analysts projects on the topic of religion and well-being: since there are many large-scale surveys covering both constructs, this seems a feasible endeavor.

4.3.1 Methodological Appropriateness

Several commentaries focused on methodological and statistical appropriateness of the models used in the MARP given the structure of the data. For instance,

[1] See `https://osf.io/w8954/` for their analysis.

[2] This approach was also taken by Team 138 who used an external variable to operationalize 'cultural norms' for research question 2 `https://osf.io/jafx6/`.

Schreiner et al. point out that measurement invariance is an important precondition for cross-cultural comparisons between any construct of interest, a view shared by Ross, Sulik, Buczny, and Schivinski.[3] Specifically, Schreiner et al. showed that the religiosity construct does not have the same factor structure across all countries, potentially invalidating a statistical analysis of the relation between religiosity and well-being.

Furthermore, Balkaya-Ince and Schnitker highlight the nested structure in the MARP data and therefore strongly advocate the use of multilevel regression models. Several commentaries, on the other hand, question their appropriateness of ordinary multilevel linear regression models due to the distributional properties of the items. That is, Schreiner et al. emphasize that categorical variables, as used in the MARP, should not be treated as continuous scores and added to an average score. They advise future projects to avoid providing precomputed means, as that may (unjustifiably) encourage teams to use continuous measures where categorical items are used. This concern is echoed by Lodder, who illustrate that the results from the regression approaches in MARP might be misleading because the ordered categorical items violate the normality assumption, in this case underestimating the size of the effect. Finally, McNamara agree that Likert scale data –such as those in the MARP– should in principle not be treated as continuous. However, they argue that the MARP results show that in practice, it may not matter whether or not Likert data are treated as ordinal or interval, as the results largely converged regardless of applying ordinal or linear models.

The fact that subjective analytic decisions did not qualitatively change the conclusions is informative in itself; whether a single-item or composite religiosity measure was used, whether a country's religious majority was accounted for, whether the non-dependence of countries was taken into account, or the fact that participants were from different countries in the first place, whether items were treated as categorical or continuous, it appears that across all these defensible strategies, the results largely converged. That is, for research question 1, all but 3 teams reported positive effect sizes with credible/confidence intervals excluding zero and for the second research question, this was the case for 65% of the teams. This is not to say that these decisions do not matter in principle–as scientists we need to think critically about both theoretical and statistical assumptions when conducting research. However, we believe that there is no "Best Model" but rather many plausible alternative analytic approaches, each with their own theoretical and statistical limitations.

4.4 Future Directions

Over the course of the project, we as the MARP core team have also gained important insights into the organisation of a many-analysts project. We were pleased that the preregistration and analysis blinding components were well-received and

[3] Ross et al. challenged us to check how many teams did check for measurement invariance/-construct validity. A quick scan through the submissions identified seven teams that mentioned investigating measurement invariance, one of which concluded that their intended analyses could not be carried out as the assumption of measurement invariance was violated.

appreciated by the teams (see Sarafoglou et al. (2022) for the comparison of analysis blinding and preregistration in the MARP). The teams used OSF templates for their preregistrations; future many-analysts projects whose analysis teams exclusively use `R` may also opt for more elaborate preregistration techniques using the `R` package `WORCS` (van Lissa, Peikert, & Brandmaier, 2021). `WORCS` allows analysis teams to (1) create a reproducible draft manuscript, (2) incorporate a version control system for their manuscripts, and (3) document all dependencies required software for a particular project (van Lissa).

A complex but critical aspect of orchestrating a many-analysts project is how to best evaluate the outcomes. We asked the analysts to provide us with one effect size measure per research question, but did not specify the type of effect size. Rather, we allowed them to submit the effect size measure that naturally followed from their analyses, since we did not want to influence the teams in their analytic approach.

To make our results interpretable we then transformed these effect sizes into standardized regression coefficients where possible. However, van Assen, Stoevenbelt, and van Aert showed that in some cases this might lead to nonsensical effect size estimates (though not necessarily in the MARP). Rather than combining (transformed) effect size measures, the authors propose to summarise the results differently, for instance, by focusing on the sign of the effect size, evidence against the hypotheses (p-values) and evidence in favour of the hypotheses (e.g., Bayes factors). Our main concern with this approach is that neither p-values nor Bayes factors quantify the size of the effect.

While we acknowledge the drawbacks of transforming effect sizes, we currently do not see a better alternative for this standard practice. Yet we underscore that there is much to be gained in research on how to best summarize results from different studies/analytic approaches, especially as meta-science projects are becoming more common. Future studies might focus on either resolving problems with respect to transforming effect sizes, creating a standardized output measure (e.g., similar to a "number needed to treat" approach in medicine), or designing a well-founded measure for subjective assessment of effect sizes.

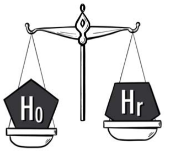
When planning the MARP, we have long considered whether the quality of the analyses should be reviewed, since it may suffer from a lack of theoretical or methodological knowledge, or from a reduced sense of ownership by the analysis teams as argued in Ross et al.. For these reasons, Silberzahn et al. (2018) evaluated the quality of the submitted analyses in a kind of peer review system. A quality control could also be established in other ways, for instance, by letting topical and methodological experts assess the submissions. These assessments can be implemented at the proposal stage (i.e., the experts act as consultants) or at the end of the project. In the latter case, the results could be weighted according to their quality, so that higher quality analyses have a greater impact on the final results (e.g., when computing the mean effect size). One problem with this approach is the subjectivity that is introduced: as apparent in the main article and in the comments on the methodological appropriateness, analysts have strong and sometimes conflicting opinions about which analysis method is best to answer the research questions. Another problem with this approach is the additional effort and time demanded from both the analysis teams and the organizing team, which might lead to delays and (presumably) a smaller number of teams starting or completing the project. Ultimately, in the MARP we assumed that all teams have principled arguments for choosing their specific analytic approach. However, this is not a general guideline; each many-analysts project must evaluate the pros and cons of implementing a quality control. Researchers interested in planning a many-analysts project will find other helpful guidance in the recently published article by Aczel et al. (2021).

4.5 Concluding Remarks

The main finding of the MARP is that religiosity and well-being are positively associated. This relation was established in a strictly confirmatory manner and seems robust against a plethora of different analytic decisions and strategies. In addition, the positive relation between individual religiosity and well-being appears stronger when religion is perceived to be normative in a particular country than when it is perceived as less normative. This moderating effect of cultural norms of religion was found consistently in the same direction, but appears less robust than the main association between religiosity and well-being.

Many-analysts approaches are relatively new to the social sciences and we hope that they will become more widely adopted in the coming years. We believe the two main merits of a many-analysts approach are that it provides (1) an indication of the robustness of the effect of interest, and (2) a concrete demonstration of the variety of theoretical angles and statistical strategies that may be added to researchers' toolboxes. We would recommend the many-analysts approach especially for much-debated research questions that are tested using a fairly straightforward design (e.g., simple associations or effects from an existing theory instead of complex cognitive models for a new hypothesis).

We consider the MARP a positive example of team science and would like to thank the analysis teams for their efforts. In fact, we are intrigued by the creative contributions of the teams exploring different aspects of religiosity and well-being beyond our imposed research questions. We hope the MARP can serve as an inspiration for future many-analysts projects.



Chapter 5

A Survey on How Preregistration Affects The Research Workflow: Better Science But More Work

Abstract

The preregistration of research protocols and analysis plans is a main reform innovation to counteract confirmation bias in the social and behavioral sciences. While theoretical reasons to preregister are frequently discussed in the literature, the individually experienced advantages and disadvantages of this method remain largely unexplored. The goal of this exploratory study was to identify the benefits and challenges of preregistration from the researcher's perspective. To this aim, we surveyed 355 researchers, 299 of whom had used preregistration in their own work. The researchers indicated the experienced or expected effects of preregistration on their workflow. The results show that experiences and expectations are mostly positive. Researchers in our sample believe that implementing preregistration improves or is likely to improve the quality of their projects, and that preregistration makes it easier to avoid questionable research practices. Criticism of preregistration is primarily related to the increase in work-related stress and the overall duration of the project. The majority of researchers with experience in preregistration reported that the benefits outweigh the challenges. However, the majority of researchers without preregistration would not consider preregistration for future projects or recommend the practice to colleagues. Our interpretation of the results is that preregistration can have positive side-effects as it adds an extra preparatory step in researchers' workflow, thus requiring researchers to think through the theoretical and practical aspects of their project.

This chapter has been submitted for publication as Sarafoglou, A., Kovacs, M., Bakos, B. E., Wagenmakers, E. J., & Aczel, B. (2021). A survey on how preregistration affects the research workflow: Better science but more work. Available as *PsyArXiv preprint*: `https://psyarxiv.com/6k5gr`.

5. A Survey on How Preregistration Affects The Research Workflow: Better Science But More Work

> A physicist had a horseshoe hanging on the door of his laboratory. His colleagues were surprised and asked whether he believed that it would bring luck to his experiments. He answered: “No, I don’t believe in superstitions. But I have been told that it works even if you don’t believe in it.”
>
> Jones (1973, p. 14)

5.1 Introduction

Over the past decade, the social sciences have undergone a methodological metamorphosis. In order to increase the quality and credibility of confirmatory empirical research, both journals and researchers have adopted a series of methodological reform measures (Spellman, 2015; Spellman, Gilbert, & Corker, 2018). Among these reform measures, preregistration is arguably the most consequential. The preregistration of empirical studies entails the specification of the research design, the hypotheses, and the analysis plan before data is collected and analyzed. Preregistration protects the confirmatory status of the reported results by preventing biases –such as confirmation bias and hindsight bias– from contaminating the statistical analysis (Munafò et al., 2017; Wagenmakers et al., 2012).

The concept of preregistration is not new; as early as 1878, Peirce (1878b, p. 476) established three rules to guarantee that a hypothesis leads to a probable result, the first rule being that a hypothesis should be explicitly stated before data are collected to test its truth. In some research areas, such as medical clinical trials, preregistration has long become scientific routine. For instance, in the world’s highest impact journal, the *New England Journal of Medicine*, the registration of clinical trials is a prerequisite for publication.

In the last ten years, preregistration has also found its way into psychological science. In fact, preregistration has become so widespread that some believe that it is on its way to becoming the norm (Nosek & Lindsay, 2018). The number of preregistrations has increased at “unprecedented and accelerating rates” (Nosek & Lindsay, 2018, p. 19), online repositories have been created to store preregistrations (e.g., the Open Science Framework (OSF; `https://osf.io`) and `AsPredicted.org`), and several journals recognize preregistered studies with badges (Kidwell et al., 2016). In addition, over 300 journals now offer the Registered Reports format as a submission option, allowing authors to integrate preregistration with the peer-review process (Chambers, 2013; Nosek & Lakens, 2014; `https://osf.io/rr/`).

In the course of its rapid spread, however, the effectiveness of preregistration has been repeatedly questioned. When discussing ways to combat the crisis of confidence, critics have argued that too heavy an emphasis is being placed on methodological reforms (e.g., Fiedler, 2018; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Szollosi et al., 2020). Specifically, methodological reforms such as preregistration should not be viewed as a silver bullet: if predictions were derived from weak theories, even the application of the most rigid methodologies will not lead to reliable scientific findings. For instance, if theories do not adequately define the conditions under which a particular phenomenon is observed, it remains unclear whether a non-significant result constitutes evidence against the theory or whether the chosen operationalizations were inappropriate (Oberauer & Lewandowsky, 2019). Thus, preregistration critics have argued that instead of focusing primarily on the prevention of questionable research practices, the discussion on how to improve psychological science should be dominated by topics such as theory development, good experimental designs, and the proper statistical modelling of theoretical predictions (Fiedler, 2018; Oberauer & Lewandowsky, 2019).

In defense of preregistration, van 't Veer and Giner-Sorolla (2016) argued that while preregistration might not *directly* improve theory development, preregistration will help shift the research focus away from the evaluation of a consistent and statistically significant pattern of results and toward the assessment of theory and methods. In addition, van 't Veer and Giner-Sorolla (2016) argue that preregistration may lead to positive side-effects that improve the overall quality of the scientific product. For instance, since all team members need to approve and scrutinize the hypotheses, methods, and analyses before data collection, study preregistration would improve the collaboration within the team and therefore yield more carefully thought-out research plans. However, it is still unclear whether or to what extent researchers actually perceive preregistered studies to be of higher quality than non-preregistered studies. On the one hand, Alister, Vickers-Jones, Sewell, and Ballard (2021) found that researchers reported that they would be more confident that a finding would replicate when the original authors had adhered to open science practices such as preregistration. On the other hand, a study by Field et al. (2020) found only ambiguous evidence that researchers trust in preregistered empirical findings more than non-preregistered ones.

It has been argued that the scrutiny associated with preregistration might even harm certain aspects of the research workflow. For instance, preregistration can be effortful and time-consuming (e.g., Nosek & Lindsay, 2018; van 't Veer & Giner-Sorolla, 2016). Open research practices were also found to have a small but statistically significant association with work pressure (Gopalakrishna et al., 2021). As recognized by Nosek et al. (2019) "[p]reregistration requires research planning and it is hard, especially contingency planning. It takes practice to make design and analysis decisions in the abstract, and it takes experience to learn what contingencies are most important to anticipate. This might lead researchers to shy away from preregistration for worries about imperfection" (p. 817). Note that other researchers have claimed the exact opposite, namely that preregistration is easy (Wagenmakers & Dutilh, 2016) and that the Registered Report format saves time (Field et al., 2020).

To date there does not exist an empirical assessment about the experiences and expectations that researchers have concerning the impact of preregistration on their workflow. This study seeks to chart the perceived benefits and drawbacks of preregistration to learn what motivates researchers to adopt this practice and possibly also what prevents researchers from adopting it. At the same time, researchers' past experiences with preregistration may be informative for pragmatic would-be adopters.

This study concerns two groups of researchers: those who published both preregistered studies and non-preregistered studies and those who only published non-preregistered studies.

5.2 Disclosures

5.2.1 Data, Materials, and Preregistration

The current study was preregistered on the Open Science Framework; in our project folder, readers can access the preregistration, as well as all materials for both the pilot and the main survey, the contact database used for the main survey, the anonymized raw and processed data (including relevant documentation), and the R code to conduct all analyses (including all figures; see Table 5.1 for an overview of URLs for the different resources). In our datasets, identifying information such as names and affiliations of the respondents were removed. Any deviations from the preregistration are mentioned in this manuscript. Note that we removed email addresses from the contact database for privacy reasons.

5.2.2 Reporting

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

5.2.3 Ethical Approval and Participant Compensation

The study was approved by the local ethics board of the University of Amsterdam (registration number: 2019-PML-11423) and of the Eotvos Lorand University (registration number: 2019/17). All participants were treated in accordance with the Declaration of Helsinki. Researchers who participated in the survey were given the opportunity to enter a raffle for a voucher from a webshop of their choice.

5.3 Methods

5.3.1 Pilot Study And Creating Materials

Before conducting the main survey, we conducted a pilot study to determine the aspects of the research workflow that are most affected by preregistration. For this pilot study we contacted 176 researchers from our database (described in the following sections) and asked them how their preregistered studies differed from their non-preregistered studies in terms of workflow, data management, and scientific quality. Respondents were asked to list both advantages and disadvantages in a free-text format. In total, we received answers from 49 researchers. The answers were then categorized by three of the authors (A.S., B.A., and M.K.). In total, nine aspects of the research process were identified as being especially impacted by preregistration. These aspects of the research process were then included as items in the main survey.

Table 5.1: Overview of URLs to this Study's Materials Available on the Open Science Framework.

Resource	URL
Project page	`https://osf.io/jcdvb/`
Preregistration of main study	`https://osf.io/qezv5/`
Preregistration of pilot study	`https://osf.io/g3fv7/`
Data and analysis code	`https://osf.io/5ytpk/`
Surveys	`https://osf.io/dzybn/`
Ethics documents	`https://osf.io/atgb7/`

5.3.2 Participants

The researchers in the preregistration group were recruited based on a contact database of published preregistered studies. Initially, we created a collection of 711 research articles in which the authors referred to a preregistered analysis plan. This collection of studies consisted of 404 preregistered and published articles that were part of the bibliographical collection of published preregistered articles from the Center of Open Science (COS), 128 articles mentioned in van den Akker et al. (2021) which originated from a database of articles with open science badges by Kambouris et al. (2020), 22 articles based on a collection from Schäfer and Schwarz (2019), and 157 articles based on a non-systematic collection of the present authors. From this initial collection of articles, we then excluded non-empirical studies (e.g., meta-analyses), Registered Reports, articles that did not include a URL to their preregistration, articles whose preregistration has been published on platforms other than the OSF (e.g., `AsPredicted.org`), and duplicates. This left a final sample of 487 articles from which we extracted the email-addresses of the corresponding authors.

5.3.3 Sampling Plan

No sample size target was specified for the preregistration group; we contacted all authors from our contact database. For the non-preregistration group, we preregistered that data would be collected until we reached a sample size as large as at least 90% of the sample size from the preregistration group. As will be discussed in the section "Sample Characteristics", we were unable to reach that goal.

5.3.4 Materials

The survey was generated using the online survey software Qualtrics (Qualtrics, 2021). The items in the main survey were based on the results of the pilot study and a discussion among the authors. The survey included questions about (1) the nine aspects of the research process that were identified in the pilot study; (2) the respondents' general opinion about preregistration; and (3) the respondents' research background. Respondents from the preregistration group were instructed to relate the questions to their own *experience* (i.e., "Please indicate below how you believe preregistration has affected your work."), whereas researchers from the non-preregistration group were instructed to indicate their *expectations* about preregistration (e.g., "Please indicate below how you believe preregistration would affect your work."). Finally, respondents also had the opportunity to give feedback on the survey and provide us with free-text on the topic of preregistration.

5.3.4.1 Nine Aspects of Research Process

Respondents were asked to indicate whether preregistration has benefited or harmed (preregistration group) or would benefit or harm (non-preregistration group) the following nine aspects of the research process: (1) the formulation of the research hypothesis; (2) the development of the experimental design; (3) the development of the analysis plan; (4) the data management; (5) the project workflow; (6) the collaboration in the team; (7) the preparatory work (e.g., pilot or simulation studies); (8) the total project duration; and (9) the work-related stress. The items concerning the first five aspects were answered using a 7-point rating scale from 1 (*less thought-through*) to 7 (*more thought-through*). The item concerning the sixth aspect (i.e., collaboration in the team) was answered using a 7-point rating scale from 1 (*got worse*) to 7 (*got better*). The item concerning the seventh aspect (i.e., preparatory work) was answered using a 7-point rating scale from 1 (*improved*) to 7 (*did not improve*). The item concerning the eighth aspect (i.e., total project duration) was answered using a 7-point rating scale from 1 (*was longer*) to 7 (*was shorter*). Finally, the item concerning the ninth aspect (i.e., work-related stress) was answered using a 7-point rating scale from 1 (*was reduced*) to 7 (*was increased*). For each question, respondents could also select the options *I do not know* and *Not applicable.*

5.3.4.2 Opinion About Preregistration

Three items asked respondents about their general opinion concerning preregistration. The first item asked about whether respondents thought preregistration has made it easier (preregistration group) or would make it easier (non-preregistration group) to avoid questionable research practices. The item was answered using a 7-point Likert scale from 1 (*Very Strongly Disagree*) to 7 (*Very Strongly Agree*). In addition, respondents could also select the options *I do not know* and *Not applicable.* The second item asked how often respondents would consider preregistration in their future work. The item was answered using a 7-point Likert scale from 1 (*Always*) to 7 (*Never*). The third item asked about whether respondents would recommend preregistration to other researchers in their field. The item was answered using a 7-point Likert scale from 1 (*Very Strongly Disagree*) to 7 (*Very Strongly Agree*).

5.3.4.3 Respondents' Research Background

Two items asked respondents about their research background. The first item asked respondents to categorize their main research approach into either (1) hypothesis testing, (2) estimation, (3) modeling/simulations, (4) qualitative research, or (5) other. The second item asked respondents to write down their specific research background (e.g., developmental psychology) as free text.

5.3.5 Procedure

Responses from the preregistration group were elicited by contacting all authors in our database (including the ones who participated in the pilot survey). Then, for each author in the preregistration group we contacted up to five authors who published a non-preregistered empirical study in the same journal, volume, and issue. When we did not reach the desired sample size for the non-preregistration group, we proceeded to contact authors who had published in previous issues of the journals. This procedure was repeated several times and stopped when we had invited almost 2,000 authors to our study.

In the main survey, respondents were first asked to indicate if they had ever (1) preregistered a study that was not published; (2) preregistered a study that was published; (3) published a study that was neither preregistered nor a Registered Report; (4) created a Registered Report that was not published; or (5) published a Registered Report. Based on their answers, the respondents were assigned to groups. Respondents were assigned to the preregistration group if they had published both preregistered and non-preregistered studies (i.e., they answered "yes" to both option 2 and 3). Respondents were assigned to the non-preregistration group if they had published exclusively non-preregistered studies (i.e., answered "yes" to option 3 and "no" to all other options). In accordance with the preregistration plan, we only analyze and report data from these two groups.

Respondents then answered the remaining survey items and one intermediate attention check item (i.e., $2 + 2 = ?$). The survey items and the attention check were presented in fixed order to the participants. The median amount of time respondents took to fill out the questionnaire was 3 minutes and 18 seconds.

5.3.6 Data Exclusions

As preregistered, we excluded respondents if (1) they were assigned neither to the preregistered group nor to the non-preregistered group ($n = 99$); (2) they did not answer all questions in the survey ($n = 23$); (3) they failed the attention check ($n = 18$); (4) they indicated in the comment section that they could not provide adequate responses or they did not accept the informed consent form ($n = 0$). In total, we received 495 responses to our survey. After exclusion, 355 responses remained for the analysis. Of these, 299 responses came from the preregistration group and 56 responses came from the non-preregistration group.

5.3.7 Analysis

This is an exploratory study and therefore we present our results mainly through descriptive statistics. For the questions relating to nine aspects of the research process, we report both the means and 95% confidence intervals (Figure 5.1; note that the presence of confidence intervals deviates from our preregistration, which stated that no inferential procedure was going to be used). For the questions on the respondents' opinion on preregistration, we visualize the frequency distributions of the survey responses (Figure 5.2). We preregistered the intention to compare, both within the preregistration group and non-preregistration group, the answers of those who choose hypothesis testing as their empirical approach to the answers of those who choose a different approach (i.e., estimation, modeling/simulations, qualitative research, or other). Due to low response rate in the non-preregistration group we could execute the intended comparison only within the preregistration group. We present the results of this comparison in Appendix B. To foreshadow the results, the answers from the hypothesis testing group did not differ notably from those of the other group. For our analyses, we excluded responses that indicated *I do not know* and *Not applicable.*

5.4 Results

5.4.1 Sample Characteristics

We first sent 487 e-mail invitations to our contact database of researchers with experience in preregistration (see the Method section for a description). Out of these 487 e-mails, 30 bounced (i.e., there was an automatic failure to deliver the e-mail, for instance, because an address was no longer active), yielding a total of 457 successfully delivered requests. Removing incomplete surveys and respondents who failed the attention check left a total sample of 299 respondents who had experience with preregistration (i.e., a response rate of 65.43%).

Next we invited a total of 1,999 researchers who had published only non-preregistered studies. Out of these 1,999 e-mails, 146 bounced, yielding a total of 1,853 successfully delivered requests. The response rate for the non-preregistration group was lower than anticipated; receiving 56 responses from 1,999 authors yields a response rate of only 2.80%. Due to this low response rate, we were unable to reach the preregistered target sample size, that is, for the non-preregistration group we only reached 18.7% of the number of responses from the preregistration group instead of the preregistered target of 90%.

Most respondents had a background in psychological science. Specifically, out of the 389 reported research backgrounds (some respondents reported more than one), 112 could be classified as social psychology (28.79%), 104 as experimental and cognitive psychology (26.74%), 36 as developmental and educational psychology (9.25%), 32 as personality psychology (8.23%), 17 as neurophysiology and physiological psychology (4.37%), 15 as applied psychology (3.86%), 12 as clinical psychology (3.08%), and 4 as methodology and statistics (1.03%). The remaining 57 responses (14.7%) could not be categorized into one of the areas above (e.g., anesthesiology).

Out of the combined total of 355 respondents, 291 respondents indicated that hypothesis testing was their primary research approach, 21 indicated estimation, 25 indicated modeling/simulations, 3 indicated qualitative research, and 15 respondents indicated other approaches.

5.4.1.1 Nine Aspects of Research Process

Figure 5.1 illustrates how preregistration was perceived to influence the nine different aspects of the research process. Overall, both groups have a positive opinion on how preregistration influenced or would influence the different aspects of the research process, with the preregistration group generally being more positive than the non-preregistration group. Specifically, respondents were most positive about the benefits of preregistration regarding the analysis plan, the hypotheses, and the study design. For two aspects, however, respondents perceived preregistration to be disadvantageous: specifically, respondents indicated that preregistration would increase both work-related stress and total project duration.

The preregistration group and the non-preregistration group differed mostly in their opinion on how preregistration influences the analysis plan and preparatory work. Although both groups reported that preregistration would benefit these aspects, respondents with preregistration experience were more enthusiastic. That is, the preregistration group reported that preregistration had made the analysis plan more thought-through ($M = 6.01\,[5.88, 6.14]$ versus $M = 4.98\,[4.54, 5.42]$) and that preregistration improved the preparatory work of the project ($M = 5.37\,[5.23, 5.51]$ versus $M = 4.55\,[4.14, 4.96]$).

In four aspects of the research process, that is, research hypothesis, experimental design, work-related stress, and total project duration, the groups showed the smallest differences of opinion. Whereas both groups perceived preregistration to benefit the experimental design ($M = 5.34\,[5.20, 5.48]$ in the preregistration group versus $M = 4.76\,[4.37, 5.15]$ in the non-preregistration group) and the research hypothesis ($M = 5.63\,[5.49, 5.77]$ in the preregistration group versus $M = 5.06\,[4.63, 5.49]$ in the non-preregistration group), preregistration was perceived to be a disadvantage with respect to work-related stress ($M = 3.73\,[3.59, 3.87]$ in the preregistration group versus $M = 3.14\,[2.71, 3.57]$ in the non-preregistration group) and total project duration ($M = 3.07\,[2.93, 3.21]$ in the preregistration group versus $M = 2.96\,[2.60, 3.32]$ in the non-preregistration group).

One aspect in which both groups gave qualitative different answers based on the group means was the influence of preregistration on the collaboration in the team. While respondents in the preregistration group indicated that it had improve the collaboration in the team ($M = 4.57\,[4.45, 4.69]$), respondents in the non-preregistration group indicated that it would be a slight disadvantage ($M = 3.84\,[3.57, 4.11]$).

5.4.1.2 Opinion About Preregistration

Figure 5.2 summarizes the general opinion about preregistration among respondents. The vast majority of respondents in the preregistration group had a positive overall opinion about the practice. Over 80% agreed with the statement that compared to their non-preregistered work, preregistration had helped avoid questionable research practices and would recommend the practice to other researchers in their field. In addition, 83% of the respondents in the preregistration group would consider preregistration in their future work. The results are somewhat more ambiguous in the group of respondents without preregistration experience. Although

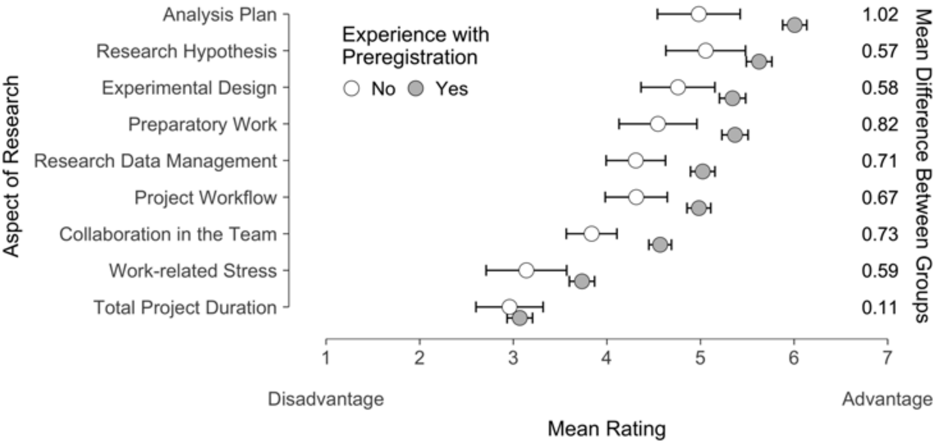


Figure 5.1: Respondents' opinion on how preregistration influenced different aspects of the research process. Grey dots represent the mean ratings from respondents who have experience with preregistration and white dots represent the mean ratings from respondents who have no experience with preregistration. The square skewers represent 95% confidence intervals. Ratings above and below 4 indicate that preregistration helped and harmed a certain research aspect, respectively.

70% agreed with the statement that preregistration would make it easier to avoid questionable research practices, only 45% would recommend the practice to other researchers in their field. Preregistration is also not seen as desirable for future research projects: only 7% in the non-preregistration group would consider this practice in their future work.

5.5 Constraints on Generality

The present study surveyed researchers who have experience with preregistering studies and those who did not. Our sample consisted exclusively of researchers in the field of psychology, presumably from differing career stages. The biggest concern regarding generalizability is that our sample was subject to self-selection. Since participation in the survey was voluntary, researchers who already had a strong opinion about preregistration might have been more likely than others to participate.

Since the proportion of respondents in the preregistration group was relatively high with 65.43%, we assume that our sample therefore reflects the population of these researchers relatively well. Therefore, we expect the results from respondents in the preregistration group to generalize to other researchers within the field of psychology who have experience with preregistration.

The results from the non-preregistration group, on the other hand, might generalize poorly to other researchers in the field since the proportion of respondents in the non-preregistration group was very low (2.80%). In this case the self-selection might have had a stronger effect on the results. However, it should be noted

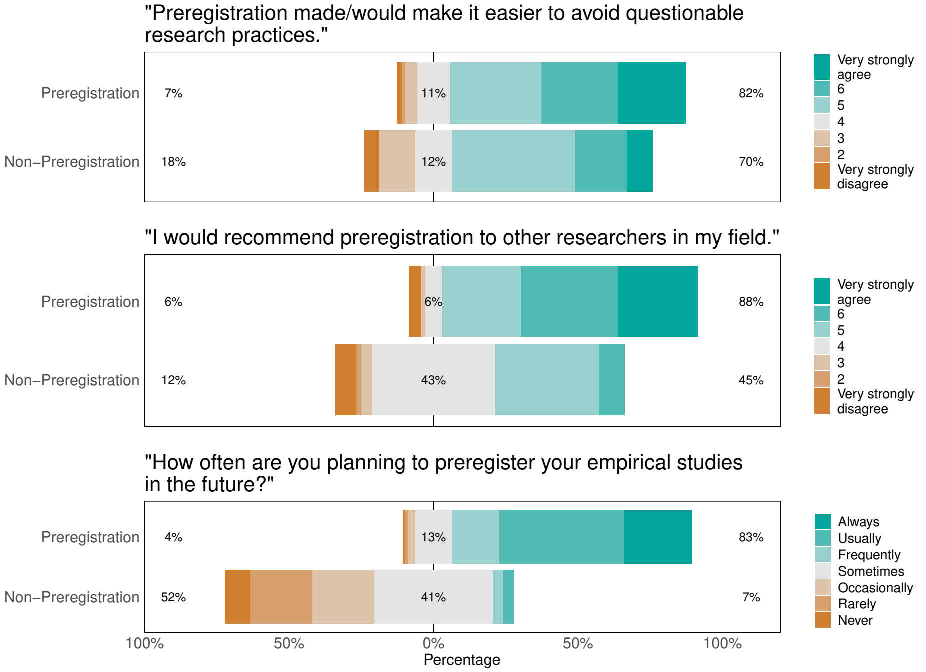


Figure 5.2: Respondents' general opinion about preregistration. The top bar represents answers from respondents who have experience with preregistration, and the bottom bar represents answers from respondents who have no experience with preregistration. For each survey question, the number to the left of the data bar (in brown/orange) indicates the percentage who (slightly or strongly) disagreed or who would recommend preregistration occasionally or less frequently. The number in the center of the data bar (in grey) indicates the percentage who responded with "neither agree or disagree" or "neutral". The number to the right of the data bar (in green/blue) indicates the percentage who (slightly or strongly) agreed or who would recommend preregistration frequently or more.

that despite the low response rates in the non-preregistration group the general response pattern (that is, the ranking of the research aspects) is consistent in both groups. This systematicity might indicate that we were not dealing with a select subgroup.

5.6 Discussion

In the last decade, preregistration has been advocated as a tool to prevent researchers' biases and expectations from contaminating the statistical analyses. It has also been argued that preregistration may have secondary effects on the research process. The current study sought to unveil these expectations and experiences.

Our results suggest that researchers find preregistration to benefit their work in most aspects of the research process. Researchers in our sample reported that preregistration improved the theoretical aspects of the project (e.g., the generation of the research hypothesis, the research design, and the analysis plan) as well as practical aspects of the project (e.g., the design and execution of pilot or simulation studies, and the general project workflow). However, disadvantages of preregistration also became apparent; preregistering a study had increased or was expected to increase the total project duration and the work-related stress.

The increase in time and effort to publish a preregistered study had been acknowledged in the literature (e.g., Nosek & Lindsay, 2018; van 't Veer & Giner-Sorolla, 2016). However, some statements made previously on the influence of preregistration on work-related stress contradict our findings. For instance, Frankenhuis and Nettle (2018, p.441) write: "From hearsay and our own experience, we think that scholars find it relaxing not to have to make [...] critical decisions after having seen the data, accompanied by a lingering sense of guilt, while cognizant of some of their biases and frustratingly unaware of others."

Although researchers with preregistration experience reported that this practice increased the total project duration and work-related stress, the vast majority of this group also indicated that they would recommend the practice to other researchers in their field, and continue to use it for their own research projects. As one respondent mentioned in the free-text comments: "Pre-Reg improves quality, which causes more work, as it should be". For researchers without preregistration experience, the equation does not seem to add up: the majority of this group would not recommend the practice to their peers, or consider this practice for themselves in the future.

How can researchers benefit from the secondary effects of preregistration? Whether or not preregistration improves the secondary aspects of the research process depends largely on the quality of the preregistration document. That is, the thoroughness of the preregistration protocol determines how carefully researchers need to think about the study design and analysis plan. A high-quality preregistration document features detailed information about the experimental conditions, the materials and stimuli used, and a comprehensive analysis plan (preferably featuring a mock data set and analysis code). To ensure that preregistration protocols meet these quality standards without considerable extra effort, researchers can fall back on a range of checklists, guidelines, and preregistration templates. Preregistration templates for the standard experimental framework can be found, for instance, on the websites `aspredicted.org` or on the Open Science Framework (`https://osf.io/zab38/`). The number of preregistration templates and tutorials for other research areas and more complex methods is increasing and includes cognitive modeling (Crüwell & Evans, 2019), secondary data analysis of pre-existing data (Mertens & Krypotos, 2019; Van den Akker et al., 2021), studies using experience sampling methods (Kirtley, Lafit, Achterhof, Hiekkaranta, & Myin-Germeys, 2021), and qualitative research (Haven et al., 2020; Haven & van Grootel, 2019). Finally, the recently developed Transparency Checklist is a quick way to check whether the preregistration and the accompanying paper comply with the current transparency standards (Aczel, Szaszi, et al., 2020).

Some researchers might also prefer alternative methods to preregistration. One of these alternatives that allows for more flexibility while still safeguarding the confirmatory status of the research is analysis blinding (Dutilh, Sarafoglou, & Wagenmakers, 2019; MacCoun, 2020; MacCoun & Perlmutter, 2015; MacCoun & Perlmutter, 2018). With analysis blinding, researchers are in principle not required to write a preregistration document. Instead, they collect their experiment data as usual and develop their analysis plan based on an altered version of the data in which the effect of interest is hidden (e.g., by shuffling the outcome variable). Another alternative would be to minimize bias by trying to map out the uncertainty in the analyses with various statistical practices (Wagenmakers et al., 2021). For instance, researchers could explore the entire universe of outcomes through multiverse analyses (in which all theoretically sensible data-preprocessing steps are explored; Steegen et al., 2016) or multi-analysts approaches (in which multiple analysis teams answer the same research question based on the same dataset; Aczel et al., 2021; Silberzahn & Uhlmann, 2015).

The aim of this study was to obtain an overview of the experienced and expected advantages and disadvantages of the practice of preregistration. Our survey shows that relying on intuition alone when developing open research practices might not be enough. Only if we know how the conceptual advantage of preregistration weighs against the individual experienced benefits and challenges can we find suitable means to improve the methodology so that it finds wider acceptance among researchers or encourage skeptics to try preregistration in their future research endeavors.

5.A Summary of Free-Text Comments

In our survey, respondents both completed the questionnaire and had the opportunity to provide comments on preregistration in an open-ended format. This section summarizes these comments. For this purpose, the authors A.S. and M.K. have divided the comments into different topics and evaluated whether they were positive, negative, or neutral statements. Comments on other topics than preregistration (e.g., comments on the survey) are not here. The full list of comments is available in our online repository at `https://osf.io/5ytpk/`. We would like to emphasize that the results should be interpreted with caution. The comments evaluated below are based on only a fraction of the respondents. Therefore, the overview given here is not necessarily representative of the opinions in our sample.

78 researchers provided us with free-text comments on preregistration. These comments highlighted both the advantages and disadvantages of preregistration: 20 comments were exclusively positive, 22 comments were negative, and 36 comments were mixed. The comments could be categorized roughly into five topics. The topics were (1) the additional workload of preregistration (mentioned by $n = 24$ respondents); (2) the effectiveness of preregistration in solving the crisis of confidence (mentioned by $n = 19$); (3) the impact of preregistration on one's career (mentioned by $n = 16$ respondents); (4) how preregistration might contribute to inequality and stigmatization in different research areas (mentioned by $n = 13$ respondents); (5) and the difficulties in the compliance with the preregistration protocol (mentioned by $n = 11$ respondents).

5.A.1 Additional workload of preregistration: harder, but worthwhile?

Proponents of preregistration argue that despite the additional workload preregistration cases, it is still "worthwhile" (e.g., Nosek & Lindsay, 2018). But do researchers agree with that statement? Not necessarily. From the $n = 24$ respondents who mentioned the additional workload, $n = 11$ respondents believed that preregistration was harder and worthwhile while seven respondents believed that it was harder, but not worthwhile–six respondents mentioned the increased workload without any further judgement. For respondents who thought preregistration was hard, but worthwhile, the added benefit of improved overall quality outweighed the added workload or was perceived as necessary consequence (e.g., "Pre-Reg improves quality, which causes more work, as it should be"). Others recognized the theoretical value of preregistration, but did not see the benefits translating into practice. For instance, one respondent wrote: "I think preregistration is great in theory, but in practice it serves only to increase the red tape and time until publication. In today's hyper-competitive publish-or-perish job market, it amounts to time wasted". The added time it takes to write a preregistration even seems to scare researchers from trying out the practice: "I understand the importance of [preregistration], but the amount of time and effort needed to preregister is probably the biggest reason I have avoided it in the past".

5.A.2 Effectiveness of preregistration in solving the crisis of confidence

19 respondents mentioned that preregistration improved the credibility of their results and the overall quality of their work. Seven respondents, however, questioned whether preregistration was a suitable tool to address the crisis of confidence. Besides the need for theory development and exploratory research, lack of methodological knowledge, and possibilities to cheat the system (by creating multiple preregistration documents) were mentioned. In addition, multiple respondents criticized the incentive structure in science, which is designed to reward research output and thus discourages the adoption of preregistration (e.g.,"[U]nless we rid science from the publication for-profit industry and educate our universities not to use the incentive structure that still very much determines who gets hired and who gets promoted based on where researchers publish rather than what they publish, I am afraid we have left the big elephant in the room untouched."; "[T]he speed at which our institutions expect us to pump through graduate students often means that pre-reg cannot happen for their work [...].").

5.A.3 Influence of preregistration on the career

16 respondents reported how preregistration influenced their career. Two respondents indicated that embracing open science practices helped their career, for instance, by giving them an advantage during the hiring process. With respect to research output, five respondents reported that publishing preregistered studies was easier while six respondents reported that it was harder. The main arguments as to why preregistered articles were easier to publish was that the respondents felt that a preregistration was expected by the journals, or they described that the "in principle acceptance" granted for Registered Reports made the publication process easier. On the other hand, respondents also described how reviewers or editors rejected papers if authors did not adhere to their preregistered plan, or that they pushed them towards rewriting their manuscripts to present polished narratives (e.g., "[R]eviewers sometimes have even criticized that I report non-significant results"; "[I] often encounter editors who still seem to want my team to change a priori aspects of manuscripts to better fit with a *we knew it all along* or in the context of competing hypothesis situations, favor the hypothesis that was ultimately supported by the data").

5.A.4 Inequality and stigmatization

In our survey, 13 respondents addressed disadvantages preregistration can have in research fields outside of psychology and for descriptive and exploratory study designs. As mentioned by some respondents, when working in fields outside of psychology (e.g., animal research) or when the research area has interfaces with industry, preregistration is relatively unknown which makes preregistered studies harder to publish (e.g., "[...] My field (animal research) is substantially behind the curve. To date, of the preregistered studies I have attempted to publish, no reviewer has commented on the preregistration as a positive aspect of the study

[...]. Rather, the reviewers who have mentioned it have used the preregistration to point out deviations (which we take care to explicitly point out in the methods) and thus has led to more challenges with publication rather than fewer. I am of the opinion that if I had submitted identical studies without preregistration, they would have been easier to publish. [...]")

In addition, respondents perceived that preregistration went to the detriment of descriptive and exploratory research. For instance, one respondent argues that confirmatory and preregistered experimental studies are currently perceived as "the gold standard [...] which leaves behind other kinds of exploratory and descriptive studies." Another respondent argues that psychology "needs a clearer distinction between confirmatory and exploratory work, and wider recognition of the value of exploratory, descriptive research that can form the basis for well-specified hypotheses". Lastly, five respondents critiqued that preregistration causes stigmatization for studies that have not been preregistered. In their comments, respondents critiqued that the reviewers often prematurely condemn a non-preregistered study, without considering its individual peculiarities. As suggested by one of the respondents, the scientific community should place more emphasis on positive reinforcement rather than harsh judgement (e.g., "I am still in favor of pre-registration and open science and I plan to pre-register the studies that I lead. At the same time, I wish that the movement was more moderate and based more on positive reinforcement").

5.A.5 Problems with data exploration and compliance with the preregistration protocol

11 respondents commented that preregistration would limit creativity, that it discourages researchers to explore the data and that adherence with the preregistration protocol was problematic, especially for early career researchers "who are still learning as they go", or when working with complex models (e.g., "In my work it's hard or sometimes impossible to know how the data should be analysed before seeing its structure, distribution, etc etc - and there is no way of accounting for every possibility in the prereg.").

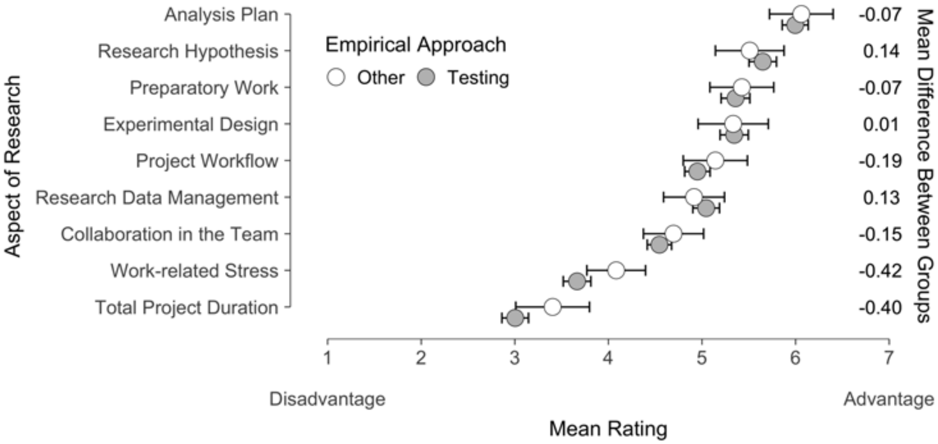


Figure 5.3: Respondents' opinion on how preregistration influenced different aspects of the research process. Grey dots represent the mean ratings from the respondents who indicated that their empirical approach was hypothesis testing and white squares represent the mean ratings from respondents who indicated a different empirical approach. The square skewers represent 95% confidence intervals. Ratings above and below 4 indicate that preregistration helped or harmed a certain research aspect, respectively.

5.B Hypothesis Testing and Exploratory Research

The following section we takes a closer look at the responses within the preregistration group. Specifically, we were interested in whether a researcher's empirical approach influences perceptions of preregistration, for instance, in that researchers who primarily test hypotheses view preregistration as more beneficial than researchers with other empirical approaches.

Within the preregistration group, 250 respondents indicated that hypothesis testing was their main empirical approach while 49 respondents indicated that their main empirical approach was a different one (e.g., estimation, modeling/simulations, qualitative research, other).

Figure 5.3 illustrates how preregistration was perceived to influence the nine different aspects of the research process. Overall, both groups have a positive opinion on how preregistration influenced the different aspects research process. The pattern resembles that of the preregistration group in general, with the analysis plan benefiting the most from preregistration while the total project duration and work-related stress have been negatively affected by the practice. Respondents who do hypothesis-testing seemed to be somewhat more negative than respondents with a different empirical approach. The biggest difference in opinion was regarding work-related stress. Here, the hypothesis-testing group perceived preregistration to be a disadvantage ($M = 3.67\,[3.52, 3.81]$), while respondents with a different empirical approach were neutral ($M = 4.08\,[3.77, 4.40]$).

Figure 5.4 illustrates the general opinion about preregistration among the re-

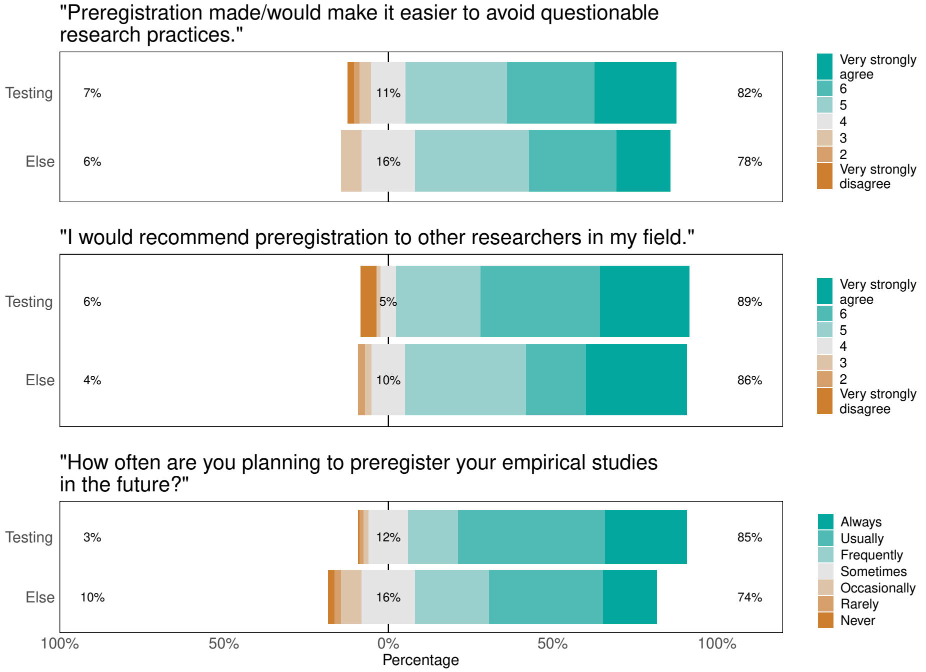
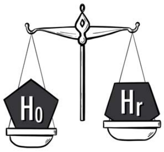


Figure 5.4: Respondents' general opinion about preregistration. The top bar represents answers from respondents whose main empirical approach was hypothesis-testing, the bottom bar represents answers from respondents whose main empirical approach was different. For each survey question, the number to the left of the data bar (in brown/orange) indicates the percentage who (slightly or strongly) disagreed or who would recommend preregistration occasionally or less frequently. The number in the center of the data bar (in grey) indicates the percentage who responded with "neither agree or disagree" or "neutral". The number to the right of the data bar (in green/blue) indicates the percentage who (slightly or strongly) agreed or who would recommend preregistration frequently or more.

spondents. The two groups do not show meaningful differences in opinion. In both groups, more than 75% agreed with the statement that compared to their non-preregistered work preregistration helped them avoid questionable research practices and more than 85% would recommend the practice to other researchers in their field. Finally, over 85% of the respondents who do hypothesis-testing would consider preregistration in their future work and 73% percent of the respondents with a different empirical approach would consider it in their future work.



Chapter 6

Flexible Yet Fair: Blinding Analyses in Experimental Psychology

Abstract

The replicability of findings in experimental psychology can be improved by distinguishing sharply between hypothesis-generating research and hypothesis-testing research. This distinction can be achieved by preregistration, a method that has recently attracted widespread attention. Although preregistration is fair in the sense that it inoculates researchers against hindsight bias and confirmation bias, preregistration does not allow researchers to analyze the data flexibly. To alleviate this concern we discuss how researchers may conduct blinded analyses (MacCoun & Perlmutter, 2015). As with preregistration, blinded analyses break the feedback loop between the analysis plan and analysis outcome, thereby preventing cherry-picking and significance seeking. However, blinded analyses retain the flexibility to account for unexpected peculiarities in the data. We discuss different methods of blinding, offer recommendations for blinding of popular experimental designs, and introduce the design for an online blinding protocol.

This chapter is published as Dutilh, G., Sarafoglou, A., & Wagenmakers, E. J. (2021). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, *198*, 5745–5772.

> When extensive series of observations have to be made, as in astronomical, meteorological, or magnetical observatories, trigonometrical surveys, and extensive chemical or physical researches, it is an advantage that the numerical work should be executed by assistants who are not interested in, and are perhaps unaware of, the expected results. The record is thus rendered perfectly impartial. It may even be desirable that those who perform the purely routine work of measurement and computation should be unacquainted with the principles of the subject.
>
> W. Stanley Jevons, 1874/1913

In recent years, large-scale replication studies revealed what some had foreseen (e.g., Ioannidis, 2005): psychological science appears to suffer from a replication rate that is alarmingly low. For instance, the Open Science Collaboration (2015) showed that out of 100 replication studies, only 39 supported the conclusions that were drawn in the original article (but see Etz & Vandekerckhove, 2016; see also Camerer et al., 2018). Similarly disappointing results were obtained for specific subfields (e.g., R. Klein et al., 2018; Marsman et al., 2017; Nosek & Lakens, 2014), and for particular effects (e.g., de Molière & Harris, 2016; Eerland, Sherrill, Magliano, & Zwaan, 2016; C. R. Harris, Coburn, Rohrer, & Pashler, 2013; Matzke, Nieuwenhuis, et al., 2015; Meyer et al., 2015; Shanks et al., 2013; Unsworth et al., 2015; Wagenmakers et al., 2016, among many others).

One of the contributing factors to the low replication rate is that researchers generally do not have to state their data analysis plans beforehand. Consequently, the reported hypothesis tests run the risk of being misleading and unfair: the tests can be informed by the data, and when this happens the tests lose their predictive interpretation, and, with it, their statistical validity (e.g., Chambers, 2017; De Groot, 1956/2014; Feynman, 1998; Gelman & Loken, 2014; Goldacre, 2009; Munafò et al., 2017; Peirce, 1878a, 1883; Wagenmakers et al., 2012).[1] In other words, researchers may implicitly or explicitly engage in cherry-picking and significance seeking. To break the feedback loop between analysis plan and analysis outcome, and thus prevent hindsight bias from contaminating the conclusions, it has been suggested that researchers should tie their hands and preregister their studies by providing a detailed analysis plan in advance of data collection (e.g., De Groot, 1969). In this article, we argue that in addition to preregistration, *blinding* of analyses can play a crucial role in improving the replicability and productivity of psychological science (e.g., Heinrich, 2003; MacCoun & Perlmutter, 2015; MacCoun & Perlmutter, 2018).

6.0.1 Preregistration

The preregistration of analysis plans ensures that statistical analyses are designed independently of specific data outcomes. Preregistration is an important compo-

[1] In the frequentist paradigm, exploratory analyses introduce a multiple comparisons problem with the number of comparisons unknown; in the Bayesian paradigm, exploratory analyses are vulnerable to a double use of the data, where an informal initial update is used to select a relevant hypothesis, and a second, formal update is used to evaluate the selected hypothesis.

nent of the Transparency and Openness Promotion guidelines (Nosek et al., 2015; see also Munafò et al., 2017) that over 800 journals and societies currently have under consideration. Moreover, the journal *Psychological Science* has adopted a preregistration badge, and in a recent editorial, Steve Lindsay stated that "Personally, I aim never again to submit for publication a report of a study that was not preregistered." (Lindsay, 2015, p. 1827).

Preregistration comes in different forms, ranging from unreviewed preregistration, where researchers upload their plans to an online archive with time stamp, to Chris Chambers' "Registered Report" format (Chambers, 2013, 2015; Lindsay, Simons, & Lilienfeld, 2016; van 't Veer & Giner-Sorolla, 2016) which over 300 journals have now adopted.[2] This especially attractive form of preregistration allows authors to initially submit for publication their introduction and method section together with a detailed data analysis plan. After review and successful revision the authors obtain "In Principle Acceptance", which ensures that the eventual publication of the paper does not depend on the critical outcome (but it does depend on the data being of sufficiently high quality, where "quality" is unambiguously defined up front). This way the Registered Report prevents both data-dependent analyses as well as publication bias while rewarding researchers for ideas and execution rather than outcome.

6.0.2 Lack of Flexibility in Preregistration

Preregistration is a powerful and increasingly popular method to raise the reliability of empirical results. Nevertheless, the appeal of preregistration is lessened by its lack of flexibility: once an analysis plan has been preregistered, that plan needs to be executed mechanically. The advantage of such a mechanical execution is that it prevents significance seeking; the disadvantage is that it also prevents the selection of statistical models that are appropriate in light of the data. Consider the scenario where the data violate the statistical assumptions underlying the planned analyses in one or more unexpected ways. For instance, sequences of response times may show a pronounced and unanticipated fatigue effect. It is evident that modeling the fatigue effect is good statistical practice, but the presence of a fatigue effect was not foreseen in the preregistered analysis plan. This deviation from the analysis plan means that the appropriate statistical analysis has to be downgraded from confirmatory to exploratory.

Two recent high-profile examples of studies where the preregistered analyses had to be discarded due to unexpected peculiarities of the data come from the Reproducibility Project: Cancer Biology, organized by the Open Science Framework. First, Horrigan et al. (2017) investigated if a specific antibody treatment reduces the growth of tumors in mice. The results of their replication attempt could not be interpreted because the authors had to deviate from the preplanned analyses since several mice showed spontaneous tumor regressions and therefore had to be excluded from the analysis. Second, Aird, Kandela, Mantis, and Reproducibility Project: Cancer Biology (2017) attempted to replicate the clinical benefit of

[2] For an overview see `https://osf.io/rr/`; for a discussion of various forms of preregistration see `https://osf.io/crg29/`.

a specific cancer treatment. The authors planned to compare medical parameters from mice that received the treatment with untreated controls by means of a paired analysis. However, due to unexpected early deaths in the control group the authors had to move from the preregistered paired analysis to an unpaired analysis.

The core problem is that preregistration does not discriminate between "significance seeking" (which is bad) and "using appropriate statistical models to account for unanticipated peculiarities in the data" (which is good). Preregistration paints these adjustments with the same brush, considering both to be data-dependent and hence exploratory.

6.0.3 An Alternative to Preregistration: Blinding Analyses

To address the lack of flexibility inherent to preregistration, we follow MacCoun and Perlmutter (2015) and argue for the broader adoption of a technique known as *analysis blinding.* Similar to preregistration, analysis blinding serves to prevent significance seeking and to inoculate researchers against hindsight bias and confirmation bias (e.g., Conley et al., 2006). But in contrast to preregistration, analysis blinding does not prevent the selection of statistical models that are appropriate in light of the observed data. We believe that blinding, in particular when combined with preregistration, makes for an ideal procedure that allows for flexibility in analyses while retaining the virtue of truly confirmatory hypothesis testing. The remainder of this article is organized as follows. We first comment on how biases can enter the research process, and then list the different types of blinding that have been proposed to prevent this from happening. Next we propose specific implementations of blinding for popular experimental designs, and illustrate the use of blinding with a hypothetical example study. Finally, we propose an online registration protocol for blinding procedures.

6.1 How Biases Enter the Research Process

In the classic text "Experimenter Effects in Behavioral Research", Rosenthal (1966) discusses how a researcher's biases can influence the result of a study. On the one hand, such biases can influence the behavior of participants: through study design and task instructions, researchers may transmit and impose their own biases onto their participants. On the other hand, a researcher's biases can also influence the conclusions that are drawn from the observations: researchers project their biases while observing and coding behavior, analyzing and interpreting data. These biases might exert their effects outside of researchers' awareness, which makes them particularly insidious.

Researchers' biases can harm the reliability of results at different stages of a study (L. E. Miller & Stewart, 2011; Schulz & Grimes, 2002). During the *data production stage*, participants can be influenced, intentionally or not, to behave according to expectations (Orne, 1962). During the *measurement stage*, for example when the observed data are submitted to a coding scheme, expectations of the coders can influence the results (Hróbjartsson et al., 2012). For both stages at which biases lurk, blinding procedures have been proposed as a remedy (Barber, 1976): If neither the participant, nor the experimenter knows which experimental condition is administered, biases at the data production and measurement stage can be prevented. Such *double blind* designs are the gold standard in medical trials and are recommended in the widely adopted guidelines provided by the International Council for Harmonisation (ICH).

A third stage at which biases can influence a study's result is the *analysis stage*. Researchers may unwittingly compromise the interpretation of their results by cherry-picking among conditions, variables (Bakker, van Dijk, & Wicherts, 2012; Bohannon, 2015), variable transformations, analyses, statistical thresholds, outlier criteria, and samples sizes (i.e., optional stopping; but see Rouder, 2014a). Such cherry-picking can turn statistical analysis into "a projective technique, such as the Rorschach, because the investigator can project on the data his own expectancies, desires, or biases and can pull out of the data almost any 'finding' he may desire" (Barber, 1976, p. 20).

For the purpose of this article, it is irrelevant whether cherry-picking is performed intentionally, as suggested by the term "p-hacking" (Bakker & Wicherts, 2011; Head, Holman, Lanfear, Kahn, & Jennions, 2015; Simmons et al., 2011), or whether researchers have the best intention but nevertheless get lost in what Gelman and Loken (2014) termed "the garden of forking paths", where the data take control over the analysis and steer it in a direction that appears worth pursuing. Regardless of whether or not bias was introduced intentionally, its end effect is the same: an overly optimistic impression of a study's result.

6.2 Preventing Bias by Analysis Blinding

All methods of analysis blinding aim to hide the analysis outcome from the analyst. Only after the analyst has settled upon a definitive analysis plan is the outcome revealed. A blinding procedure thus requires at least two parties: a *data manager* who blinds the data and an *analyst* who designs the analyses. By interrupting the feedback loop between results and outcomes, blinding eliminates an important source of researcher bias. The unbiased nature of the blinding procedure is symbolized by Lady Justice in Figure 6.1. A detailed cartoon of a blinding procedure is presented in the appendix.

The idea of blinding analyses goes back several decades. Indeed, as early as 1957, Sainz, Bigelow, and Barwise (1957) introduced to medicine the term "triple blind" design, referring to the procedure in which the participants, the experimenter, as well as the analyst are blind to the experimental manipulations. Blinding analyses has not become as widespread as single and double blind designs, but it is commonly advocated as a tool in medical research, for example in the CONSORT reporting guidelines (D. Moher et al., 2010; see also Gøtzsche, 1996). In nuclear and particle physics, blinding of data in the analysis stage is common practice (Heinrich, 2003). In other fields, including psychology, blinding of analyses is exceedingly rare. In experimental psychology, our literature search revealed only a handful of studies reporting blinded analyses (Dutilh et al., 2017a; J. Moher, Lakshmanan, Egeth, & Ewen, 2014; van Dongen-Boomsma, Vollebregt,

Figure 6.1: Lady Justice weighs the evidence in favor of each of two competing hypotheses. The blindfold symbolizes the unbiased nature of the evaluation.

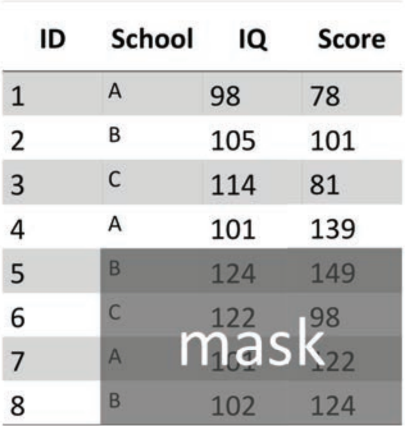
Slaats-Willemse, & Buitelaar, 2013).

6.3 Methods of Analysis Blinding

Analyses may be blinded in various ways and the selection of the appropriate blinding methodology requires careful thought. As Conley et al. (2006, p. 10) write: “[...] the goal is to hide as little information as possible while still acting against experimenter bias.” Thus, a first consideration is how much to distort a variable of interest. The distortion should be strong enough to hide any existing effects of key interest, yet small enough to still allow for sensible selection of an appropriate statistical model. A second consideration is that some relationships in the data may need to be left entirely untouched. For example, in a regression design it is important to know about the extent to which predictors correlate. If this collinearity is not part of the hypothesis, the applied blinding procedure should leave it intact.

Below we outline different blinding procedures and discuss their advantages and limitations. We then propose the blinding methods that are most suitable for popular designs in experimental psychology.

6.3.0.1 Method 1. Masking a subset



ID	School	IQ	Score
1	A	98	78
2	B	105	101
3	C	114	81
4	A	101	139
5	B	124	149
6	C	122	98
7	A	101	122
8	B	102	124

Figure 6.2: Blinding method 1: Masking a subset. The analysis plan is designed using a subset of the data while the remainder of the data have been masked by the data manager.

Each blinding method is illustrated using a fictitious example featuring data of students from different schools who performed a math test. For each of the students, an estimate of their IQ is also available (Figure 6.2).

Blinding can be achieved when the data manager splits the data into a calibration set and a test set (Figure 6.2), similar to what happens in the model-selection technique known as cross-validation (e.g., Browne, 2000; Yarkoni & Westfall, 2017). For the design of an appropriate analysis plan, the analyst is given access only to the calibration set. Once the analyst has committed to a specific analysis plan, the data manager provides the data from the test set, and the proposed analysis plan is then applied mechanically, without any adjustment, on the test set. Importantly, the final conclusions depend exclusively on the analysis outcome for the test set.

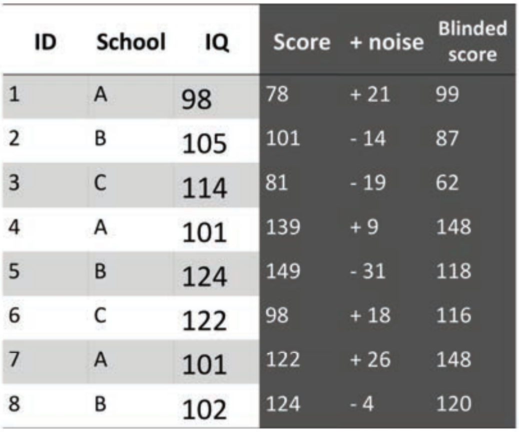
A special case of this procedure is an exact replication study: the complete analysis procedure is defined by the original experiment and is applied unchanged to data from the replication attempt. The main benefit of masking a subset is that it can be executed with relatively little effort, and that it is certain to prevent any feedback from results to analysis.

One drawback of this procedure, which is shared by cross-validation as a model-selection technique, is that it is not clear how to determine the relative size of the calibration and test set. A second drawback is that, once the relative size has been decided upon, the data manager should resist the temptation to examine the two sets and ‘correct’ perceived imbalances in particular characteristics. The third and main drawback of this procedure is that the construction of the calibration set costs data. For example, a study initially thought to have sufficient power (or sufficiently high probability of finding compelling evidence, see Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019), may become underpowered in light of a split-half cross-validation technique in which 50% of the data have to be sacrificed to construct the calibration set.

One way to alleviate this problem is to simulate data based on a small calibration set and design the analyses on the basis of these simulated data (Heinrich, 2003). This simulation procedure, however, involves a range of non-trivial choices that might lower the representativeness of the simulated calibration set.

In conclusion, the advantage of masking a subset is particularly pronounced for the analysis of large data sets with ample opportunity for significance seeking. In such cases, the costs of subset-masking are low (as the test set will still be sufficiently large so as to draw confident conclusions) and the benefits are substantial.

6.3.0.2 Method 2. Adding Noise



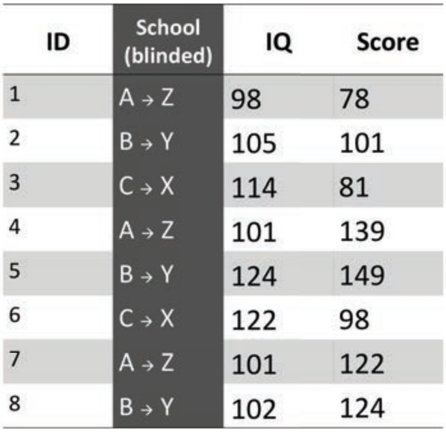
ID	School	IQ	Score	+ noise	Blinded score
1	A	98	78	+ 21	99
2	B	105	101	- 14	87
3	C	114	81	- 19	62
4	A	101	139	+ 9	148
5	B	124	149	- 31	118
6	C	122	98	+ 18	116
7	A	101	122	+ 26	148
8	B	102	124	- 4	120

Figure 6.3: Blinding method 2: Adding noise. The analysis plan is designed using a ‘Blinded score’, a version of the dependent variable to which the data manager has added noise.

A straightforward method of blinding is where the data manager adds random noise to all values of the dependent variable. For example, in Figure 6.3, noise is added to the dependent variable ‘Score’. The proper amount of noise will mask any real effects in the data, such that, when executed on the contaminated data, the blinded analysis is unlikely to show the predicted effect. After the analyst has settled on a data analysis plan, the data manager supplies the original data and the analysis plan is mechanically executed on the original data.

For this blinding method, the precise amount of noise is of critical importance. For example, consider a researcher who compares test scores of children on different schools (dummy data presented in Figure 6.3). When the added noise is drawn from a uniform distribution between, say, $-.1$ and $.1$, this would not hide existing effects. On the other extreme, when the added noise is drawn from a uniform distribution between, say, -1000 and 1000, not only would all effects be hidden, but the noise would also dramatically alter the distributional properties of the score variable. As a result, the analyst is no longer able to define a sensible outlier removal protocol or choose an appropriate transformation for the test score variable.

6.3.0.3 Method 3. Masking Labels



ID	School (blinded)	IQ	Score
1	A → Z	98	78
2	B → Y	105	101
3	C → X	114	81
4	A → Z	101	139
5	B → Y	124	149
6	C → X	122	98
7	A → Z	101	122
8	B → Y	102	124

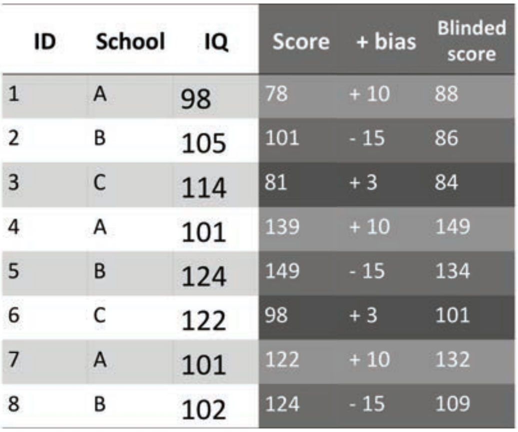
Figure 6.4: Blinding method 3: Masking labels. The analysis plan is designed on a data set for which the data manager has masked the labels of the factor levels. In this example, school labels A, B, and C are replaced by Z, Y, and X, respectively.

Another straightforward method of blinding is achieved by masking or shuffling the level labels of an experimental factor. By masking the condition labels, the data stay entirely intact but the analyst does not know whether the effects she finds are in the expected direction. Figure 6.4 illustrates how the labels of the different schools have been masked by the data manager before the analyst is allowed to access the data.

One drawback of this method is that the analyst is still able to see whether or not there are significant effects between the cells. If the researcher prefers to find any effect over no effect, this method will not stop this bias from influencing the results. Practically, this drawback disappears when factors with many levels are studied. For example, an anthropologist who studies in which countries people are the most generous might use this blinding method and mask the country indicators.

6.3.0.4 Method 4. Adding Cell Bias and Equalizing Means

A relatively subtle method of blinding is to add the same random number to all observations within the same cell of an experimental design in order to “shift the answer” (Heinrich, 2003). Consider the researcher who studies the difference between test scores from students on three different schools (Figure 6.5). To blind the data analysis, the data manager could change the means in each condition by adding a random number to each of the observations, for instance +10 to all observations from condition A, −15 to all observations from condition B, and +3 to all observations from condition C. This implementation of cell bias leaves the



ID	School	IQ	Score	+ bias	Blinded score
1	A	98	78	+ 10	88
2	B	105	101	- 15	86
3	C	114	81	+ 3	84
4	A	101	139	+ 10	149
5	B	124	149	- 15	134
6	C	122	98	+ 3	101
7	A	101	122	+ 10	132
8	B	102	124	- 15	109

Figure 6.5: Blinding method 4: Adding cell bias. The analysis plan is designed using a 'Blinded score', a version of the dependent variable to which the data manager has added cell-specific bias.

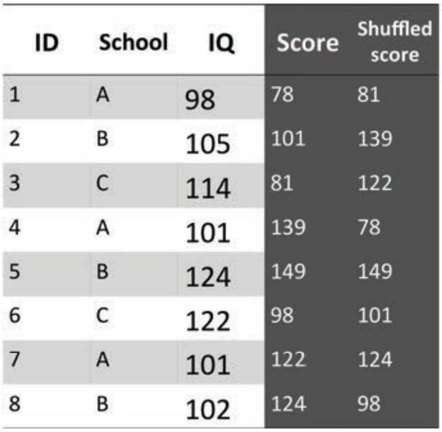
distribution of test scores for the three schools intact. At the same time, the right amount of bias will obscure the differences between the groups. The analyst is unable to search for significance, because the available group differences may just reflect cell bias that was injected by the data manager. As for the method of adding noise, the distribution from which these cell biases are chosen is crucial. Specifically, too little noise will not blind the analyses. In addition, adding cell bias obscures only the location of the dependent variable. If the mean of the dependent variable is correlated with its spread (e.g., for a response time distribution, the standard deviation generally increases with the mean; Wagenmakers & Brown, 2007), then the data analyst can use the spread to discover the hidden information about the mean.

A special way to shift the answer is by removing all effects that are present in the data set such that the mean is equal for all cells. This *equalizing of means* results in a blinded data set in which the null-hypothesis is true by construction. One advantage of the equalizing of means method is that a biased analyst cannot p-hack data that were blinded in this way. On the contrary, the imposed truth of the null hypothesis can serve as a sanity check: any analysis that supports an effect of the experimental manipulation here must be reconsidered.

Note that when the means of the criterion variable have been equalized, an absolute outlier exclusion rule can no longer be used for this variable. For instance, consider a situation in which response time is the criterion variable. The blinding prevents the analyst from knowing the absolute response times. This lack of knowledge makes it impossible to argue from the data that, say, 1200 milliseconds is a good cutoff for outlier removal. Instead, the analyst needs to either formulate a relative outlier criterion, for instance removing the 1% slowest responses, or to formulate an absolute criterion based on theoretical grounds.

Still, no form of adding cell bias provides a bullet-proof solution, as the true ordering of means may sometimes be reconstructed from aspects of the data that the means are correlated with (e.g., the ordering of the standard deviations).

6.3.0.5 Method 5. Shuffling of Variables



ID	School	IQ	Score	Shuffled score
1	A	98	78	81
2	B	105	101	139
3	C	114	81	122
4	A	101	139	78
5	B	124	149	149
6	C	122	98	101
7	A	101	122	124
8	B	102	124	98

Figure 6.6: Blinding method 5: Shuffle variables. The analysis plan is designed using 'Shuffled Rows', a version of the dependent variable that was shuffled by the data manager.

A versatile and effective method of blinding is to shuffle the key variables, while leaving the remaining variables untouched. This procedure can be applied to both correlational and factorial designs.

Shuffling variables in a correlational design For correlation or regression analyses, both predictor or criterion variables can be shuffled. Any correlation with a shuffled variable is based on chance, which breaks the results–analysis feedback loop. When the design is bivariate, it does not matter which variable is shuffled. However, in the case of a multiple regression, it is preferable to shuffle the criterion variable, so that eventual collinearity of the predictors stays intact and can be accounted for. An example of such blinding is performed in Dutilh et al. (2017a), who studied whether people's scores on a working memory task can be predicted by response time and accuracy on a simple two-choice task. Only the criterion variable (i.e., score on a working memory test) was shuffled, whereas the collinearity of the predictor variables (i.e., response time and accuracy) could be explicitly accounted for by a cognitive model. Relative to blinding methods that add noise to observations, shuffling of a variable has the advantage that the distributional properties of the variable stay intact.

Shuffling variables in a factorial design For a factorial design, one can shuffle the predictor variable(s) across observations, that is, randomly permute the condition labels for all participants. The result is a blinded data set in which all differences between cells of the factorial design are based on chance. Shuffling factorial predictors, however, might lead to a misrepresentation of the distributional properties of the original data set. Consider the analyst who plans to perform an ANOVA comparing the test scores on three different schools. Assume that the test scores from the three schools are very different, but within each school, there is a highly skewed distribution of test scores, violating the normality assumptions of an ANOVA. When the analyst would model the data without blinding, he or she would rightfully decide to transform the test scores before executing the ANOVA. With the condition indicator shuffled, however, the true skew of the test scores for the individual schools might be warped by mixing the three differently skewed distributions. As a result, the analyst may recommend an inappropriate transformation.

6.3.0.6 Method 6. Cloned Data Analyses

This overarching method addresses a potential problem with many of the blinding methods described above, namely that particular aspects of the blinded data are defined by chance. Depending on chance, the blinding procedure might eliminate existing effects, induce effects where none exist, or change the direction of effects. Consequently, some of the blinded data sets may not be representative for the original data and hence provoke the stipulation of an analysis plan that is inappropriate. For example, the blinding method might have distorted the distributional shape of the data.

To address this issue, MacCoun and Perlmutter (2018) proposed a procedure that we call *cloned data analysis*. Here the analyst works with multiple data sets (e.g., six), one of which is the original data set. A suitable blinding method makes it impossible to reliably identify the original data set from among the blinded clones. By working with multiple blinded data sets as well as the original data set, the analysis that is planned is certain to be appropriate for the original data set.[3]

6.4 Application to Standard Designs in Experimental Psychology

The various blinding procedures come with advantages and disadvantages, the relative importance of which depends on the experimental design. Below we recommend specific blinding methods for the three standard inferential situations in psychology and the social sciences more generally: regressions, contingency tables, and ANOVA designs. Our recommendations are meant as a starting point, as the specifics of the research design sometimes require tailor–made methods. Only when more studies apply blinded analysis techniques will we learn what methods (or what combination of methods) are most appropriate.

[3] This is why we believe it is important that the original data set is always among the set of clones; instead, MacCoun and Perlmutter (2018) propose to let chance determine whether or not the original data set is included.

6.4.1 Regression Designs

When one or more continuous variables are assumed to predict a criterion variable we argue that the best method of blinding is to shuffle the criterion variable Y and leave the predictors intact (i.e., method 5 'shuffling of variables' above).

Consider again the study by Dutilh et al. (2017a), who set out to test whether elementary processing speed (as measured from performance on a perceptual two-choice task) predicts working memory capacity. Performance on the perceptual task was measured by response time and accuracy, whereas working memory capacity was measured by a single composite score obtained from a battery of memory tests. The data manager blinded the analysis by shuffling the working memory capacity variable and then sent the shuffled data set to the analyst. The analyst was free to explore different ways to model the relation between response time and accuracy. The analyst was also free to account for peculiarities in the distribution of working memory capacity (e.g., eliminate outliers). Once the analyst was satisfied with the statistical procedure, he shared the intended analysis plan with the co-authors by publishing it online. The blind was then lifted and the planned procedure was applied without any changes to the original version of the working memory variable.

A similar situation often occurs in neuroscience, when researchers seek to study the correlation between behavior and particular measures of brain activation. Much like in the example above, there is a need for flexibility of analysis on one side of the regression equation. For example, when functional MRI signals are to be correlated with behavioral measures, the rich fMRI data first need to be preprocessed and compressed, and this can be done in many different plausible ways (Carp, 2012; Poldrack et al., 2017). When behavior is measured with one variable such as a test score, the easiest way to blind the analyses is to shuffle this test score variable. The analyst is then free to explore different ways to process the fMRI data without unduly and unwittingly influencing the results concerning the correlation with the criterion variable.

The situation becomes only slightly more complicated when there is a need for flexibility at both sides of the regression equation. Consider for instance a study that aimed to relate activity in certain brain regions to behavior as expressed by the parameters of a cognitive model (e.g., Forstmann et al., 2008). Here, both the neural analyses and the cognitive modeling require flexibility of analysis. In this case, the solution is to keep the variables of interest intact for each participant at each side of the equation. Blinding is achieved by shuffling the case identifier, thereby destroying the connection between the brain activity and behavior.

6.4.2 Contingency Tables

For fully categorical data, we again recommend to shuffle the dependent variable. For example, when studying whether class attendance ('always', 'sometimes', 'never') predicts whether students 'pass' or 'fail' a course, it is convenient to shuffle the pass/fail variable. Table 6.1 shows fictitious original data. Table 6.2 shows how this table could look after shuffling the pass/fail variable.

Table 6.1: Fictitious data for students who pass or fail a course depending on their class attendance.

	Outcome		
Class Attendance	Pass	Fail	Total
'Always'	30	3	33
'Sometimes'	10	20	30
'Never'	3	6	9
Total	43	29	**72**

Table 6.2: Fictitious data for students who pass or fail a course depending on their class attendance, with shuffled outcome variable.

	Outcome		
Class Attendance	Pass	Fail	Total
'Always'	17	16	33
'Sometimes'	19	11	30
'Never'	7	2	9
Total	43	29	**72**

Note that this shuffling assures that the margin counts are kept intact. This way the analyst is given access to the total number of students who pass and fail, and the total number of students who reported to attend class never, sometimes, and always. The analyst is then free to decide on sensible outlier removal criteria and variable transformations without having to fear that unconscious biases unduly influence the analysis outcome. For example, the analyst might want to merge two very similar categories since one of these has very few counts.

6.4.3 ANOVA Designs

The ANOVA design is ubiquitous in experimental psychology. In the simplest scenario, the ANOVA concerns a comparison between the means of two groups. For example, a researcher may seek to study whether people who hold a pen between their teeth perceive cartoons to be funnier than do people who hold the pen between their lips (Wagenmakers et al., 2016). In this design, it is essential to use a blinding technique that distorts the mean perceived funniness in each group. Distortion of the cell means could be achieved by shuffling the condition indicators (holding a pen or not).

However, because this shuffling of condition indicators will also distort the form of the within-cell distributions of the dependent variable, we propose for ANOVA designs to *equalize the cell means*. This way, the effects of interest are masked while the distribution of the data is left intact. The easiest way to equalize cell means is by setting them all to zero, i.e., subtracting the cell mean from each observation. Importantly, the coordinator does not adjust variables that are not the focus of the hypothesis, thereby allowing the analyst the freedom to use these extra variables sensibly. For example, the answers to an exit interview can be used to exclude participants.

Still, even after equalizing the means a particularly determined analyst may still try and learn the identity of the conditions (e.g., by considering the spread of the distributions), after which the resulting analyses are again susceptible to bias. Therefore, in addition to equalizing means, we recommend to shuffle the labels of the factor levels. For instance, in the pen study the analyst would not know whether a particular participant was in the 'teeth' condition or in the 'lips' condition.

6.5 Blinding as Integral Part of Preregistration

Below we analyze fictitious data to illustrate the strength of blinding when combined with preregistration. The example shows how blinding can prevent a real and substantial effect from being downgraded from a confirmatory to an exploratory finding.[4]

6.5.1 Blinding and Preregistration in a Hypothetical Research Project

Consider the following hypothetical research project, preregistered but without a blinding procedure: An experimental psychologist aims to test the hypothesis that priming participants with the concept of god makes them more willing to help others (Shariff & Norenzayan, 2007). More interestingly, the psychologist hypothesizes that this positive effect is attenuated by paying participants for their participation, a speculation motivated by Deci, Koestner, and Ryan (1999) who suggested that monetary incentives decrease participants' intrinsic motivation. To test this hypothesis, the psychologist measures helpfulness as the amount of time that a participant voluntarily commits to perform extra tasks for the experimenter. The design features two factors with two levels each: god prime vs. no god prime and monetary reward vs. no reward. The hypothesis is defined as an *interaction*, such that the size of the boost in helpfulness due to the god prime depends on whether or not participants get payed for their participation.

In order to protect herself from publication bias in case the results turn out in favor of the null-hypothesis, the psychologist submits her proposal as a Registered Report (`https://osf.io/rr/`). The preregistration protocol includes a sampling plan (i.e., testing 50 participants in each cell of the design for a total of 200 participants) and an analysis plan (i.e., a two-factor ANOVA, where the dependent variable is the time that participants voluntarily commit). The protocol states that the hypothesis is said to be supported when the ANOVA interaction of god prime and payment condition shows a p value lower than .05. Similar studies in the past

[4] The .jasp file for this example is accessible via: `https://osf.io/p7nkx/`.

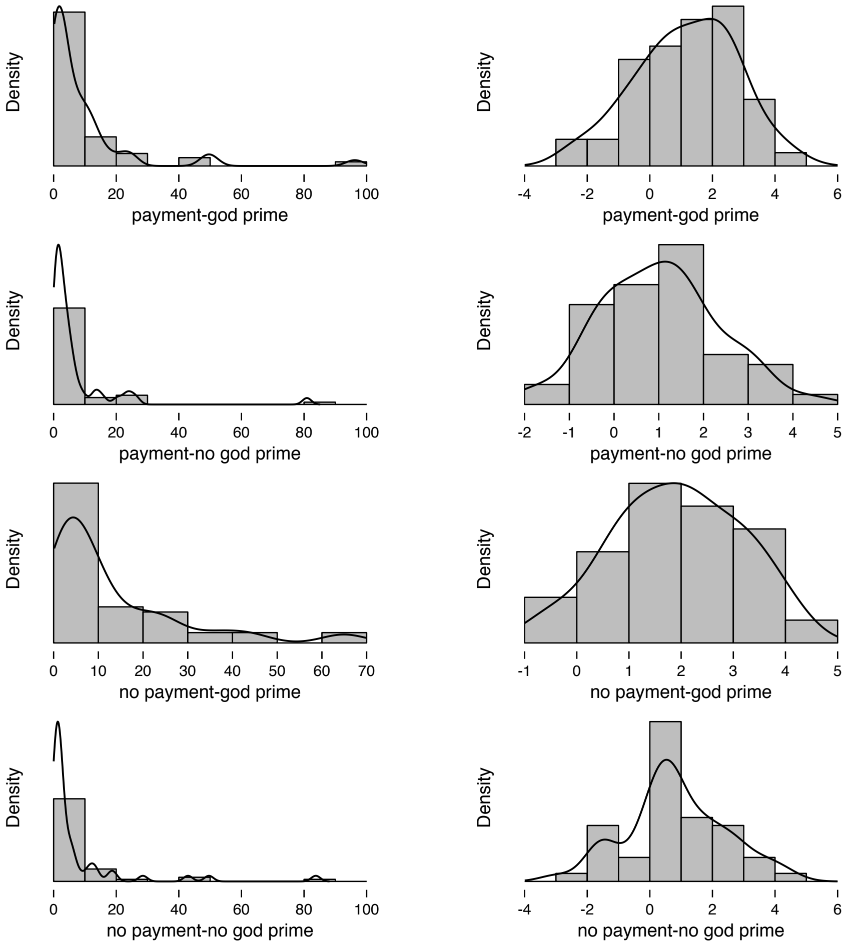


Figure 6.7: Raw data (left hand panels) and log–transformed data (right hand panels). For each payment × god–prime condition, the raw data are heavily skewed. After log–transformation, the distributions look approximately normal.

had performed the exact same analysis. The preregistered proposal is reviewed and eventually accepted by the editor, who rewards the psychologist with "In Principle Acceptance" conditional on a data quality check involving a significant ($p < .05$) main effect of the god-prime manipulation. [5]

The collected data are depicted in the four left-most histograms of Figure 6.7 (simulated data). The researcher notices that the data are heavily skewed but is required to execute the preregistered analysis plan, which produces the ANOVA output shown in Table 6.3.

The psychologist is disappointed to find that none of the effects is significant at

[5] Both the psychologist and the editor are unaware that the ANOVA harbors a hidden multiplicity problem, as explained in Cramer et al. (2016).

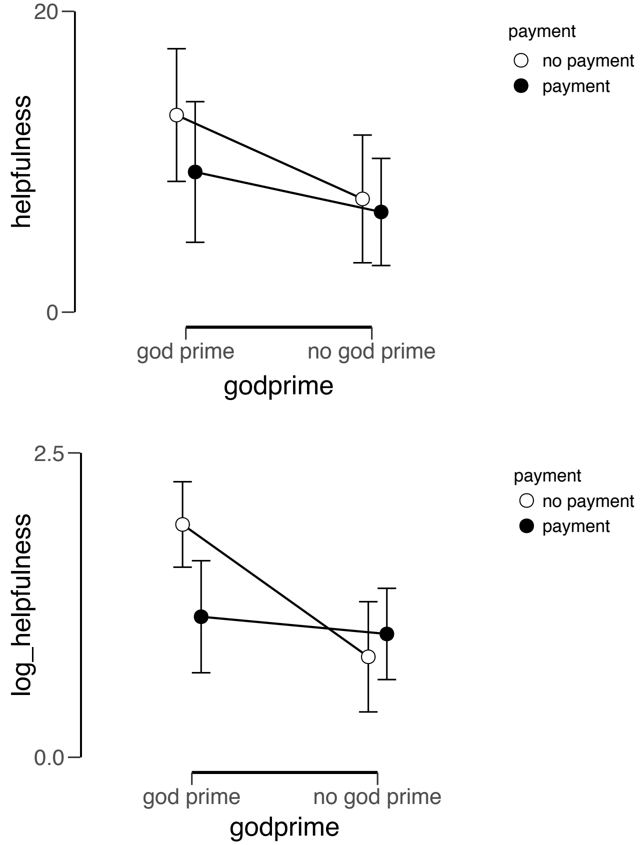


Figure 6.8: Means and 95% confidence intervals based on the raw data (upper panel) and log–transformed data (lower panel).

the .05 level. She decides to deviate from the preregistered analysis plan and perform a log-transformation on the dependent variable (i.e., the number of minutes volunteered) to account for the obvious skew. The right four histograms of Figure 6.7 show that the transformation indeed removed the skew of the dependent variable. The means and their 95% confidence intervals for both the raw and the log-transformed data are shown in Figure 6.8. The ANOVA on the log-transformed data leads to the result in the Table 6.4.

The ANOVA on the transformed data shows the effect precisely as predicted: the data quality check is met (i.e., there is a significant main effect of the god-prime manipulation with $p = .0031$) and, more importantly, there is a significant interaction between the payment and god-prime manipulations (i.e., $p = .0222$) in the expected direction, supporting the hypothesis that the payment reduces the effect of the god prime on helpfulness.[6]

[6] The psychologist is inconvenienced by recent arguments that p-values higher than .005 pro-

Table 6.3: ANOVA results from a hypothetical Registered Report investigating the interaction between god-priming and monetary reward on willingness to help.

Cases	Sum of Squares	df	Mean Square	F	p
payment	270.657	1	270.657	1.214	0.272
godprime	847.498	1	847.498	3.801	0.053
payment × godprime	106.296	1	106.296	0.477	0.491
Residual	43700.868	196	222.964		

Note. ANOVA table based on the raw data from Figure 6.7.

Table 6.4: ANOVA results based on the log-transformed data.

Cases	Sum of Squares	df	Mean Square	F	p
payment	4.065	1	4.065	1.931	0.166
godprime	18.865	1	18.865	8.959	0.003
payment × godprime	11.183	1	11.183	5.311	0.022
Residual	412.709	196	2.106		

Note. ANOVA table based on data from Figure 6.7 after a log-transformation.

The psychologist now has a serious problem: the analysis with the log-transformation cannot be reported as a preregistered confirmatory analysis. Instead, in the results section of the preregistered study, the researcher has to conclude that the confirmatory test did not show support for the hypothesis. In the section "exploratory results", the ANOVA on the log-transformation is reported with the encouragement to carry out a confirmatory test of this hypothesis in future work.

This unsatisfactory course of events could have been prevented by including a blinding protocol in the preregistered analyses plan. The blinding protocol could outline that in the blinded data set the cell means will be equalized and the condition labels will be shuffled, as we recommended above. This blinding procedure would have enabled the analyst to observe the extreme right skew of the data, while offering the flexibility to explore several transformations before settling on a definite analysis plan. Because these transformations were all applied on a properly blinded version of the data, the results could still have been presented as truly confirmatory.

6.6 Incentivizing Blinding

The example above illustrates that, in addition to guarding against bias, blinding can also prevent confirmatory findings from being demoted to exploratory findings. Thus, for preregistered studies there is a strong incentive to include a blinding

vide only suggestive evidence (Benjamin et al., 2018) and therefore decides that these arguments are not compelling and can best be ignored.

protocol. In studies that are not preregistered, however, the incentive to apply blinding may not be readily apparent, and the blinding procedure itself may seem relatively involved. We now present the structure of an online blinding protocol that facilitates blinded analyses.

6.6.1 Online Blinding Protocol

We propose an online blinding protocol that allows researchers to declare the blinding procedure they follow and receive a certificate. The protocol kills two birds with one stone: The protocol serves science by making explicit the difference between exploration and confirmation, and it serves the scientist by convincing editors, reviewers, and colleagues that the reported results are untainted by bias.

The protocol is easy to follow. The only requirement is that there are two actors involved: (1) the data manager, who has the access to the original data set and (2) the analyst, who designs the analyses and has no direct access to the original data set. The protocol consists of three steps, preferably preceded by a preregistration step (see Figure 6.9).

Phase 0 In case of a Registered Report, the study's method and analysis plan are preregistered, including the blinding protocol. All of this is peer-reviewed and revised if required.

Phase 1 The data manager creates a project on the online blinding portal. She uploads the raw data that is kept private until Phase 3. Then, she uploads the blinded data, which become publicly available. By signing an online contract, she declares that she did not share the raw data or sensitive information about the blinding procedure with the analyst.

Phase 2 The analyst downloads the blinded data and designs the analyses based on that data. When she is satisfied with the analysis plan, she uploads it to the online blinding portal. By signing an online contract, she declares that she did not have access to the raw data or sensitive information about the blinding procedure when designing the analyses.

Phase 3 The raw data is revealed to the analyst, so that she can apply her analysis plan. If both data manager and analyst agree, at this point the data is also made publicly available. A blinding certificate is issued in the form of an online document that describes the blinding procedure followed. The data manager and author can include a link to this certificate in the manuscript that reports the results.

Thanks to the blinding certificate, the reader of the eventual report can trust that specific analyses were performed without knowledge of the outcome. This is important, since only when an article precisely reports how analyses are performed (as advocated in D. Moher et al., 2010), can the results be interpreted appropriately. Sadly, the reader of psychological articles is currently most often blind to which analyses, data transformations, and exclusion criteria have been tried before

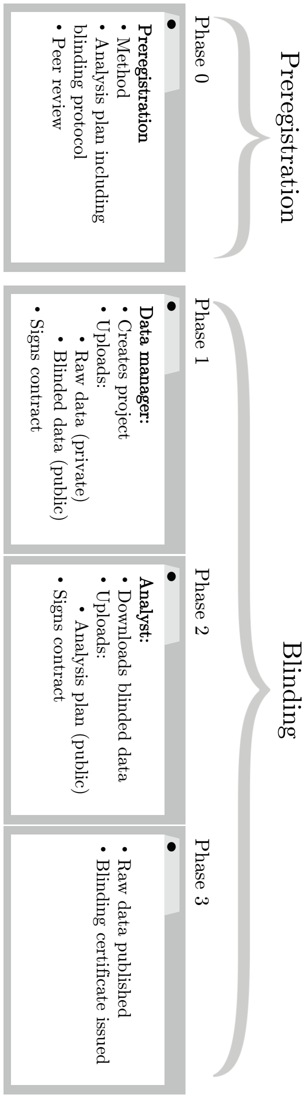


Figure 6.9: An online protocol for data blinding. In three steps, the blinding procedure is standardized and registered. The full transparency is awarded with a blinding certificate.

the authors settled upon the reported analyses. Thus, the online blinding protocol presents a promising opportunity to unblind the reader.

Signing the contract does of course not prevent cheating. In principle, researchers could sign the contract, but still perform all the hazardous post-hoc practices described earlier. Thanks to the signing of the contract, however, these questionable research practices are now clearly classified as fraud and cannot be engaged in unwittingly.

6.7 Discussion and Conclusion

This article advocates analysis blinding and add it to the toolbox of improved practices and standards that are currently revolutionizing psychology. We believe that the blinding of analyses is an intuitive procedure that can be achieved in almost any study. The method is of particular interest for preregistered studies, because blinding keeps intact the virtue of true confirmatory hypothesis testing, while offering the flexibility that an analyst needs to account for peculiarities of a data set. Nonetheless, for studies that are not preregistered, blinding analyses can substantially improve the rigor and reliability of experimental findings. Blinding allows a sharp distinction between exploratory and confirmatory analyses, while allowing the analyst almost complete flexibility in selecting appropriate statistical models.

We are aware that, as any other method to improve the reliability of science, blinding is not a silver-bullet solution. In addition, blinding has to be performed honestly and accurately.

6.7.1 Honesty

First of all, blinding brings the acclaimed virtues only when performed honestly. Claiming that you have performed analyses blinded although you peeked at the data is of course highly questionable. It is, however, not always easy to abstain from doing so. Data is most often collected in the same lab as where it is analyzed. A discussion during lunch between the data manager and the analyst might supply the analyst with information he or she did not want to know. As a partial solution to this problem, the proposed online registration of a blinding protocol increases the awareness of sticking to the rules.

6.7.2 Errors in Analysis Discovered After Unblinding

Another potential problem that can occur when data analyses are designed blindly, is that they simply turn out not to work when the blind is lifted. For example, in spite of a careful choice of the blinding method, the analyses turn out not to be able to account for a crucial property of the data, e.g., bimodality of a variable's distribution. Also, it is possible that simple coding errors are only discovered after the blind is lifted. Such mistakes are frustrating: the analyses cannot be interpreted as purely confirmatory anymore.

When analyses turn out not to work on the unblinded data, there are two possible solutions. First, one can simply describe what went wrong and include a description of both the planned and the corrected analyses in the manuscript. Another solution is to go one step further and get a second analyst involved and repeat the blinding procedure.

We want to stress that without blinding, the chances of ending up with exploratory analyses is much larger. Researchers often try a number of analyses on the real data before settling on the analysis to be described in the eventual article (John et al., 2012). The analyses they eventually present should often be labeled exploratory.

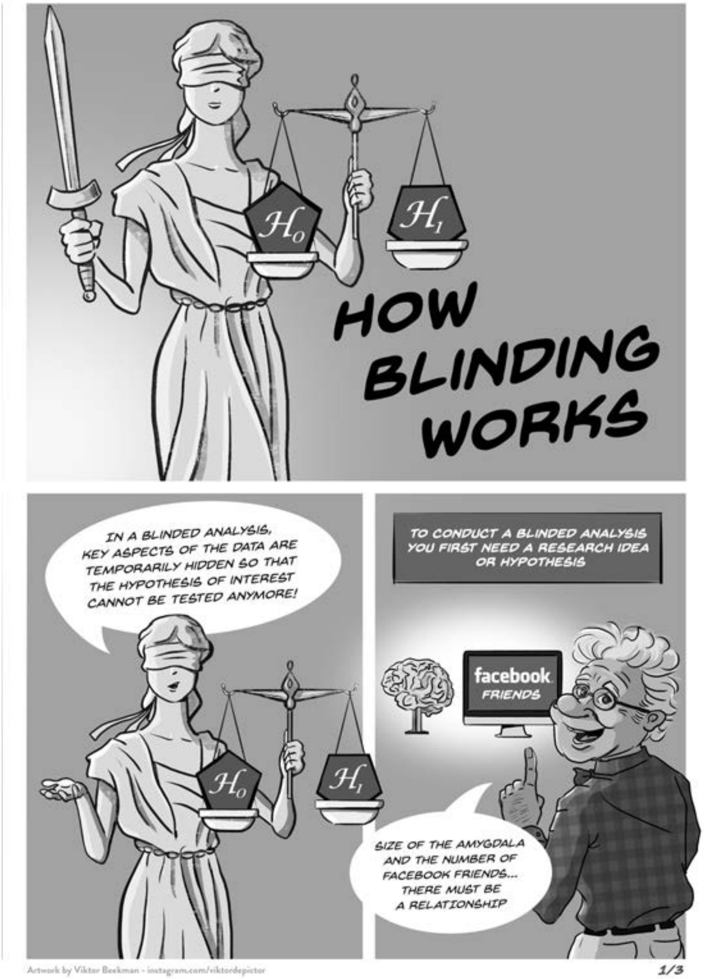
6.7.3 Can Blinding Really Improve Reproducibility?

As we noted above, analysis blinding is already being employed in other fields such as medicine and physics. Meta studies have revealed that experiments in which a blinding technique is applied show on average fewer positive results (Hróbjartsson et al., 2014) and report smaller effect sizes (Bello et al., 2014; Holman, Head, Lanfear, & Jennions, 2015) than studies without blinding procedure. These results should be viewed in relation to the findings from the reproducibility project by Open Science Collaboration (2015), who reported that only 39% of the effects found in the original articles were qualitatively replicated, and that the average effect size of the replication studies was about half as large as the average effect sizes reported in the original studies. These and other results suggest that blinding (and other procedures to tighten the methodological screws) can increase the reproducibility and reliability of results reported in experimental psychology.

6.7.4 Get Excited Again

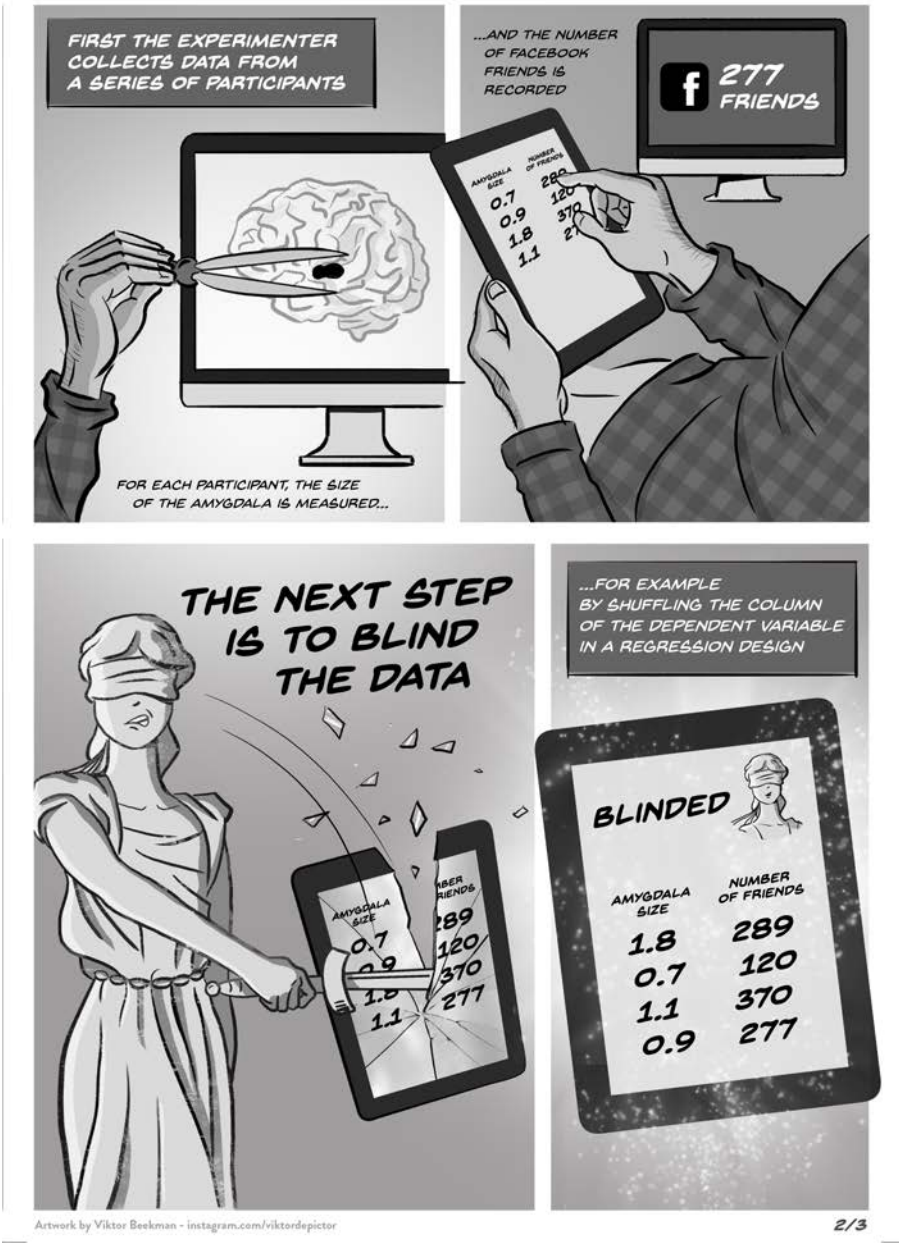
We want to finish our plea for analysis blinding on a personal note. We know many students and colleagues who analyze their data inside and out. So much time is spent on the analysis, iterating between analysis and outcome, that the eventual results can hardly be called exciting. We ourselves have had this experience too. Now, having used a blinding protocol in our own work, we have experienced how blinding can bring back the excitement in research. Once you have settled on a particular set of analyses, lifting the blind is an exciting event – it can reveal the extent to which the data support or undermine your hypotheses, without having to worry about whether the analysis was either biased or inappropriate.

6.A A Cartoon to Explain How Blinding Works


$\mathcal{H}_0$
$\mathcal{H}_1$
HOW BLINDING WORKS
IN A BLINDED ANALYSIS, KEY ASPECTS OF THE DATA ARE TEMPORARILY HIDDEN SO THAT THE HYPOTHESIS OF INTEREST CANNOT BE TESTED ANYMORE!
$\mathcal{H}_0$
$\mathcal{H}_1$
TO CONDUCT A BLINDED ANALYSIS YOU FIRST NEED A RESEARCH IDEA OR HYPOTHESIS
facebook FRIENDS
SIZE OF THE AMYGDALA AND THE NUMBER OF FACEBOOK FRIENDS... THERE MUST BE A RELATIONSHIP

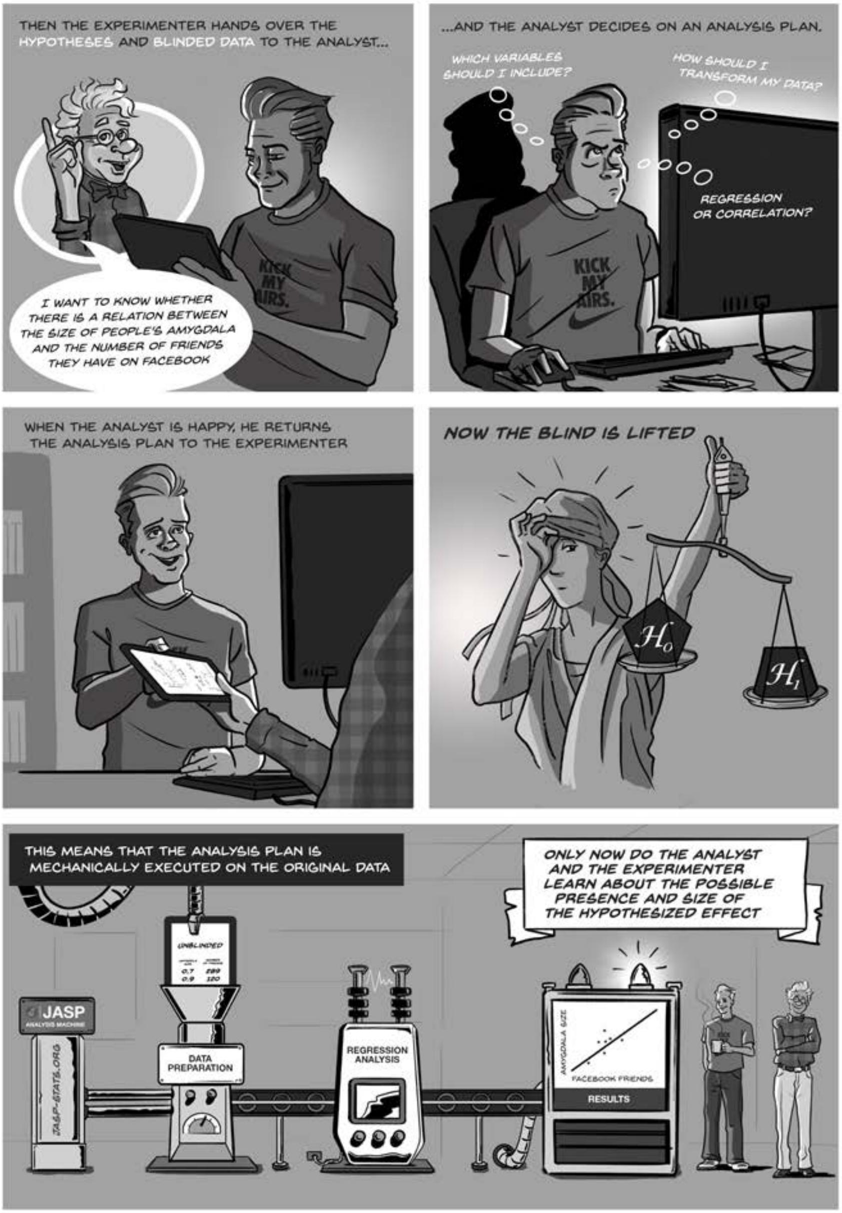
Artwork by Viktor Beekman - instagram.com/viktordepictor

1/3


FIRST THE EXPERIMENTER COLLECTS DATA FROM A SERIES OF PARTICIPANTS
FOR EACH PARTICIPANT, THE SIZE OF THE AMYGDALA IS MEASURED...
...AND THE NUMBER OF FACEBOOK FRIENDS IS RECORDED
f 277 FRIENDS
AMYGDALA SIZE
NUMBER OF FRIENDS
0.7 289
0.9 120
1.8 370
1.1 27
THE NEXT STEP IS TO BLIND THE DATA
...FOR EXAMPLE BY SHUFFLING THE COLUMN OF THE DEPENDENT VARIABLE IN A REGRESSION DESIGN
BLINDED
2/3

AMYGDALA SIZE	NUMBER OF FRIENDS
1.8	289
0.7	120
1.1	370
0.9	277

Artwork by Viktor Beekman - instagram.com/viktordepictor


THEN THE EXPERIMENTER HANDS OVER THE HYPOTHESES AND BLINDED DATA TO THE ANALYST...
I WANT TO KNOW WHETHER THERE IS A RELATION BETWEEN THE SIZE OF PEOPLE'S AMYGDALA AND THE NUMBER OF FRIENDS THEY HAVE ON FACEBOOK
KICK MY AIRS.
...AND THE ANALYST DECIDES ON AN ANALYSIS PLAN.
WHICH VARIABLES SHOULD I INCLUDE?
HOW SHOULD I TRANSFORM MY DATA?
REGRESSION OR CORRELATION?
WHEN THE ANALYST IS HAPPY, HE RETURNS THE ANALYSIS PLAN TO THE EXPERIMENTER
NOW THE BLIND IS LIFTED
$\mathcal{H}_0$
$\mathcal{H}_1$
THIS MEANS THAT THE ANALYSIS PLAN IS MECHANICALLY EXECUTED ON THE ORIGINAL DATA
ONLY NOW DO THE ANALYST AND THE EXPERIMENTER LEARN ABOUT THE POSSIBLE PRESENCE AND SIZE OF THE HYPOTHESIZED EFFECT
JASP
ANALYSIS MACHINE
JASP-STATS.ORG
UNBLINDED
DATA PREPARATION
REGRESSION ANALYSIS
AMYGDALA SIZE
FACEBOOK FRIENDS
RESULTS

Artwork by Viktor Beekman - instagram.com/viktordepictor

3/3

Chapter 7

Comparing Analysis Blinding with Preregistration in the Many-Analysts Religion Project

Abstract

In psychology, preregistration is the most widely used method to ensure the confirmatory status of analyses. However, the method has disadvantages: not only is it perceived as effortful and time consuming, but reasonable deviations from the analysis plan demote the status of the study to exploratory. An alternative to preregistration is analysis blinding, where researchers develop their analysis on an altered version of the data. In this study, we compare the reported efficiency and convenience of the two methods in the context of the Many-Analysts Religion Project. In this project, 120 teams answered the same research questions on the same dataset, either preregistering their analysis (n = 61) or using analysis blinding (n = 59). Our results provide strong evidence (BF = 11.40) for the hypothesis that analysis blinding leads to fewer deviations from the analysis plan and if teams deviated they did so on fewer aspects. Contrary to our hypothesis, we found strong evidence (BF = 13.19) that both methods involved approximately the same amount of work. Finally, we found no and moderate evidence on whether analysis blinding was perceived as less effortful and frustrating, respectively. We conclude that analysis blinding does not mean less work, but researchers can still benefit from the method since they can plan more appropriate analyses from which they deviate less frequently.

This chapter has been submitted for publication as Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E. J. (2022). Comparing Analysis Blinding With Preregistration In The Many-Analysts Religion Project. Available as *PsyArXiv preprint*: `https://psyarxiv.com/6dn8f/`

7.1 Introduction

The “crisis of confidence” in psychological science (Pashler & Wagenmakers, 2012) inspired a variety of methodological reforms that aim to increase the quality and credibility of confirmatory empirical research. Among these reforms, preregistration is arguably the most vigorous and widespread. Preregistration protects the confirmatory status of the study by restricting the researchers’ degrees of freedom in conducting a study and analyzing the data (e.g., Chambers, 2017; Munafò et al., 2017; Wagenmakers et al., 2012). When preregistering studies, researchers specify in detail the study design, sampling plan, measures, and analysis plan before data collection. By specifying these aspects beforehand, researchers protect themselves against their (subconscious) tendencies to select favorable –that is, statistically significant– results.

Preregistration is fair in the sense that it restricts the researchers’ degrees of freedom. However, this implies that researchers must anticipate all possible peculiarities of the data and define analysis paths for each scenario, which can be perceived as effortful and time-consuming (Nosek & Lindsay, 2018; Sarafoglou, Kovacs, Bakos, Wagenmakers, & Aczel, 2021). Indeed, it is rare for researchers to adhere fully to their preregistration plan. When comparing preregistrations to published manuscripts, two recent studies found that only a small minority did not contain any deviations from the preregistration: two out of 27 in Claesen, Gomes, Tuerlinckx, and Vanpaemel (2021) and seven out of 20 in Heirene et al. (2021). More serious still is the dilemma that preregistration does not distinguish between significance seeking and selecting appropriate methods to analyze the data. Researchers face a harsh penalty for reasonable deviations from their preregistered analysis plan, for instance, by removing outliers, transforming skewed data, or account for measurement invariance. By adjusting the analysis plan to properties of the data, the analysis will be demoted from “confirmatory” to “exploratory” even when the adjustments were entirely appropriate and independent from any significance test that was entertained. This makes preregistration a challenge for research that includes any sort of non-trivial statistical modeling (e.g., Dutilh et al., 2017b).

An alternative to preregistration is analysis blinding (Dutilh, Sarafoglou, & Wagenmakers, 2019; MacCoun, 2020; MacCoun & Perlmutter, 2015; MacCoun & Perlmutter, 2018). Just like preregistration, analysis blinding safeguards the confirmatory status of the analysis. However, the analyst does not specify their analysis before data collection. Instead, the analyst develops their analysis plan based on a blinded version of the data, that is, a dataset in which a collaborator or an independent researcher has removed any potentially biasing information.

One can create a blinded version of the data, for instance, by providing the analyst with a subset of the data (i.e., data that only feature a subset of participants, or data in which the key outcome measure is removed), by shuffling the key outcome measures in regression designs, or by equalizing the group means across experimental conditions in factorial designs (see Dutilh, Sarafoglou, & Wagenmakers, 2019 for an overview on different blinding techniques for common study designs in experimental psychology). Then, the analyst creates an analysis script that preprocesses the blinded data (e.g., explores the factor structure of relevant measure, identifies outliers, handles missing cases) and executes the appropriate statistical analysis. After the analyst is satisfied with their analysis plan they receive access to the real data and execute their script without any changes. To make this process transparent, the analyst may choose to publish their analytic script to a public repository such as the Open Science Framework (OSF; Center for Open Science, 2021) before accessing the data.

The benefit of analysis blinding is that it offers the flexibility to explore the data and fit statistical models to its idiosyncrasies, yet preventing an analysis that is tailored to the outcomes. In addition, it could save researchers time and effort since the additional step of creating a preregistration document is omitted.

7.1.1 Current Study

The current study assesses the potential benefits of analysis blinding over the preregistration of analysis plans in terms of efficiency and convenience. As part of the Many-Analysts Religion Project (MARP; The MARP Team, 2022), we invited teams to answer two research questions on the relationship between religiosity and well-being. Specifically, the teams investigated (1) whether religious people self-report higher well-being, and (2) whether the relation between religiosity and self-reported well-being depends on perceived cultural norms of religion. Relevant to this study is that we assigned the teams to two conditions, that is, they either preregistered their analysis plan or used analysis blinding.

To complete the project, the teams had to go through two distinct stages. In stage 1 the teams had to conceptualize, write, and submit their analysis plan. They did so either by submitting a completed preregistration template, or by submitting an executable analysis script based on the blinded version of the data. In stage 2, the teams were granted access to the real dataset to execute their planned analysis. After the sign-up and after each stage of the project, the teams completed brief surveys on their experiences with planning and executing the analysis and on how their change of beliefs on the two MARP research questions.

7.1.2 Research Question and Hypotheses

Our overarching research question was: *Does analysis blinding have benefits over preregistration in terms of workload and convenience?* We predicted four benefits of analysis blinding, which led to the following hypotheses:

1. The total workload spent on planning and executing the analysis is higher for teams in the preregistration condition than for teams in the analysis blinding condition

2. The perceived effort for planning and executing the analysis is higher for teams in the preregistration condition than for teams in the analysis blinding condition

3. The perceived frustration when planning and executing the analysis is higher for teams in the preregistration condition than for teams in the analysis blinding condition.

Table 7.1: Overview of URLs to this Study's Materials Available on the Open Science Framework.

Resource	URL
Project page	`https://osf.io/vy8z7/`
Preregistration	`https://osf.io/2cdht/`
Data and analysis code	`https://osf.io/gkxqy/`
Stage 1 materials for preregistration teams	`https://osf.io/a5ent/`
Stage 1 materials for analysis blinding teams	`https://osf.io/ktvqw/`
Surveys and ethics documents	`https://osf.io/kgqze/`
MARP data	`https://osf.io/6njsy/`

4. Teams in the preregistration condition deviate more often from their planned analysis than teams in the analysis blinding condition and when they deviate from their analysis plan, teams in the preregistration condition deviate on more items than teams in the analysis blinding condition.

7.2 Disclosures

7.2.1 Preregistration and Analysis Blinding

Prior to collecting data, we preregistered the intended analyses on the Open Science Framework. These analyses were then verified and adjusted –if necessary– based on the blinded version of the data. The author SH acted as data manager (i.e., blinded the dataset) and author AS verified and adjusted the data analysis. The final analysis pipeline was uploaded to the OSF project page, before the analysis on the real data was carried out. Any deviations from the preregistration are mentioned in this chapter.

7.2.2 Data and Materials

Table 7.1 shows an overview of important resources of the study. Readers can access the preregistration, the materials for the study, the blinded and real data (including relevant documentation), and the R code to conduct all analyses (including all figures), in our OSF folder at: `https://osf.io/vy8z7/`.

7.2.3 Reporting

We report how we determined our sample size, all data exclusions, and all manipulations in the study. However, since this project was part of the MARP we will not describe all measures in this study. Here, we only describe measures relevant to the research question. The description of the remaining measures can be found in The MARP Team (2022).

7.2.4 Ethical approval

The study was approved by the local ethics board of the University of Amsterdam (registration number: 2019-PML-12707). All participants were treated in accordance with the Declaration of Helsinki.

7.3 Methods

7.3.1 Participants and Recruitment

The analysis teams were recruited through advertisements in various newsletters and email lists (e.g., the International Association for the Psychology of Religion (IAPR), Cognitive Science of Religion (CSR), Society for Personality and Social Psychology (SPSP), and the Society for the Psychology of Religion and Spirituality (Div. 36 of the APA)), on social media platforms (i.e., blogposts and Twitter), and through the authors' personal network. We invited researchers from all career stages (i.e., from doctoral student to full professor). Teams were allowed to include graduate and undergraduate students in their teams as long as each team also included a PhD candidate or a more senior researcher. Initially, $N = 173$ teams signed up to participate in the MARP. From those teams, $N = 127$ submitted an analysis plan and $N = 120$ completed the whole project. Out of the final sample of $N = 120$ teams, 61 had been assigned to the preregistration condition, and 59 had been assigned to the analysis blinding condition. As compensation, the members from each analysis team were included as co-authors on the MARP manuscript. No teams were excluded from the study.

7.3.2 Sampling Plan

The preregistered sample size target was set to a minimum of 20 participating teams, which was based on the number of recruited teams in the many analysts project from Silberzahn and Uhlmann (2015). However, we did not set a maximum number of participating teams. The recruitment of teams was ended on December 22, 2020.

7.3.3 Study Design

The current design was a between-subjects design (at the team level). Our dependent variables were (1) total workload in hours, (2) perceived effort, (3) perceived frustration, and (4) deviation from the analysis plan. Our independent variable was the assigned analytic strategy which had two levels (preregistration, analysis blinding).

7.3.4 Randomization

The assignment of teams to conditions was done with block randomization. After sign-up, each analysis team was randomly assigned to one of the two conditions in blocks of four so that the groups were approximately equally sized at all times. In four cases, members from different teams requested to collaborate. When those teams were assigned to different conditions and they had not yet submitted an analysis plan, they were instructed not to fill out the preregistration template but to follow the instructions of the analysis blinding condition instead.

7.3.5 Materials

In stage 1 each team received the research questions, a project description and a brief summary of the theoretical background on the relationship between religiosity and well-being, the original materials, the documentation for the MARP data, and instructions specific to their assigned condition. In stage 2, teams were granted access to the MARP data. After sign-up, and after completing stage 1 and 2, the teams were instructed to fill out surveys, further referred to as pre-survey, mid-survey, and post-survey. The pre-survey included questions about the background of the teams. The mid-survey and the post-survey included questions about the workload and about their perceived level of frustration and effort during the process. The post-survey also inquired whether and how the teams deviated from their submitted analysis plan. Only one survey per analysis team was required and the teams were instructed to either sum up the responses from each team member (for workload items) or give joint answers depending on the consensus within the team. The pre-survey, mid-survey, and post-survey were generated using Google Forms.

7.3.5.1 Project Description and Theoretical Background

Teams received a 5 page document with an overview of the MARP, the research questions, two paragraphs on the theoretical background on the relationship between religiosity and well-being, and a description of the measures and some features in the MARP data (i.e., number of participants, number of countries).

7.3.5.2 Original Materials

The teams received the cross-cultural survey used to collect the MARP data. This survey was provided in English and contained all items and answer options.

7.3.5.3 MARP Data and Data Documentation

The MARP data featured information of 10,535 participants from 24 countries collected in 2019. The data were collected as part of the cross-cultural religious replication project (see also Hoogeveen et al., 2021; Hoogeveen & van Elk, 2018). The MARP data contained measures of religiosity, well-being, perceived cultural norms of religion, as well as some demographics.

To achieve analysis blinding, we shuffled the key outcome variable, that is the well-being scores. In the blinded data, we ensured that the scores on a country level remained intact to facilitate hierarchical modeling and outlier detection. That is, we shuffled well-being within countries so that the average well-being score for each country was the same in the real and blinded data. In addition, we ensured that the well-being scores within each individual remained intact, that is, well-being scores associated with one individual were shuffled together.

The data documentation featured a detailed description for each of the 46 columns in the data. It disclosed the scaling of the items and whether and how many missing values there were in each variable.

7.3.5.4 Independent Variable: Assigned Analytic Strategy

Teams were randomly assigned to the preregistration condition or to the analysis blinding condition. These conditions differed with respect to the instructions and materials they received in stage 1. Teams in the preregistration condition received a document which briefly explained preregistration and a preregistration template (see appendix). The template was a shortened version of the "OSF Preregistration" template from the Center of Open Science. It entailed only the aspects of preregistration related to the analysis plan that is the (1) operationalization of the variables, (2) the analytic approach, (3) outlier removal and handling of missing cases, and (4) inference criteria.

Teams in the analysis blinding condition received a document which briefly explained analysis blinding and a blinded version of the MARP data. Participants received the following information about the blinded data:
In this blinded dataset, we made sure that

- *The relationship between well-being and all other independent variables is destroyed.*
- *Data on the country level are intact. This means that, for instance, the mean religiosity we measured in Germany is identical in the blinded version of the data as well as in the real data.*
- *All well-being scores are intact within a person.*
- *All religiosity scores are intact within a person.*

7.3.5.5 Dependent Variables: Experienced Workload, Experienced Effort, Experienced Frustration, and Deviations From the Planned Analysis

In the mid-survey and in the post-survey we asked participants to indicate their experienced, effort, and frustration to accomplish the tasks from stage 1 (i.e., writing and submitting the analysis plan) and stage 2 (i.e., executing the analysis), respectively.

One item asked to indicate how many hours it took the team to accomplish the tasks at the respective stage of the project. The teams could respond by giving numerical values and were instructed to add up the work hours for each team member.

One item asked to indicate how hard the team had to work to accomplish the task during the respective stage. This item was answered using a 7-point Likert-type scale from 1 (*Effort was very low*) to 7 (*Effort was very high*). Lastly, one item asked to indicate how frustrated the team was during the respective stage (i.e., whether they felt insecure, discouraged, irritated, stressed, or annoyed). This item was answered using a 7-point Likert-type scale from 1 (*Frustration was very low*) to 7 (*Frustration was very high*). The items concerning the perceived effort and frustration were inspired by Hart (2021). The measures “Workload”, “Perceived effort”, and “Perceived frustration” were computed by summing up the indicated values for stage 1 and stage 2 for each team.

In the post-survey, we asked teams whether they deviated from their analysis plan after they received the real data. If they answered “Yes” to that question, they indicated out of a catalogue of eight aspects which aspects they deviated on. These aspects were: (1) hypothesis, (2) included variables, (3) operationalization of dependent variables, (4) operationalization of independent variables, (5) exclusion criteria, (6) statistical test, (7) statistical model, and (8) direction of the effect.

The items concerning the deviations from the analysis plan were based on a subset of the catalogue presented in Claesen et al. (2021). In addition, the teams could describe in a text field which peculiarities caused them to deviate from their analysis plan.[1]

7.3.5.6 Anticipated Workload

As an additional exploratory variable we measured whether the indicated work hours were more time than the team had anticipated. This item was answered using a 5-point Likert-type scale from 1 (*No, much less*) to 5 (*Yes, much more*). We computed the measure “Anticipated Workload” by summing up the indicated values for stage 1 and stage 2 for each team.

7.3.5.7 Respondents’ Research Background

In the pre-survey, five items asked respondents about their research background. The first item asked how many people the analysis team consists of. In the final dataset, this number was updated for teams that requested to collaborate, meaning that in these cases the number of team members were summed. The second item asked to describe the represented subfield(s) of research in the team. The third item asked about what positions were represented in the team. The answer options were (1) doctoral student, (2) post-doc, (3) assistant professor, (4) associate professor, and (5) full professor. The fourth item asked the teams to rate their theoretical knowledge on the topic of religion and well-being. The fifth item asked the teams to rate their knowledge on methodology and statistics. The fourth and fifth item were answered using a 5-point Likert-type scale from 1 (*No knowledge*) to 5 (*Expert*). The teams were instructed that if they participated as a team that they should indicate their collective knowledge.

7.3.6 Procedure

We started advertising MARP on September 11, 2020. After teams had signed-up to the project we asked them to complete the pre-survey. The teams then received their analysis team number, access to their OSF project folder, and all materials

[1] Four teams indicated that they deviated from their analysis plan, but selected “no” to all the options. These teams were coded to have one deviation.

and instructions needed to complete stage 1 of the project. To complete stage 1, the teams had to upload their analysis plans to their OSF project page and complete the mid-survey. We then "checked-out" the submitted analysis plans (i.e., created a file in their OSF project folder that cannot be edited or deleted). The deadline to complete stage 1 was December 22, 2020. In stage 2, the teams then were granted access to the real data. To finalize stage 2 of the project, the teams had to complete the post-survey. We also encouraged the teams to upload all relevant files, together with a brief "ReadMe" document and a summary of their results to their project folder. We discouraged the open communication of analysis strategies or results (e.g., through Twitter) until after the official deadline of stage 2 of the project, which was February 28, 2021.

7.3.7 Statistical Model

We used Bayesian inference for all statistical analyses. As preregistered, we aimed to collect at least strong evidence (i.e., a Bayes factor of at least 10) in favor for our hypotheses. Each hypothesis was tested against the null hypothesis that the respective outcomes are the same under both conditions. To test hypothesis 1 and 2, we conducted one-sided Bayesian independent samples *t*-tests. To test hypothesis 3, we conducted a one-sided Bayesian Mann-Whitney U test. For hypothesis 1 and 2, we additionally conducted a robustness analysis to check how different prior specifications influence the results and a sequential analysis to check how the evidence changes as the data accumulates. For all three analyses, we assigned a one-sided Cauchy prior distribution with scale 0.707 to the effect size (i.e., $\delta \sim \text{Cauchy}^{-}(0, 0.707)$). These analyses were conducted in JASP (JASP Team, 2021).

To test hypothesis 4, we fitted two zero-inflated Poisson regression models as defined by Lambert (1992) and implemented in McElreath (2016). This model assumes that with probability θ a team will report zero deviations and with probability $1-\theta$ the number of reported deviations (i.e., zero or higher) are estimated using a Poisson(λ) distribution. The first model included "analysis method" as predictor, the second model did not. McElreath (2016) expressed the logit-transformed parameter θ' as the additive term of an intercept and a predictor variable. Following their recommendations, we assigned a standard normal distribution as prior to both the intercept parameter and the predictor variable. Similarly, McElreath (2016) expressed the log-transformed parameter λ' as the additive term of an intercept and a predictor variable, to which we assigned a Normal$(0, 10)$ distribution and a standard normal distribution as prior, respectively.

We then estimated the log marginal likelihoods of these models using bridge sampling and computed the Bayes factor for these two models (Gronau et al., 2017; Gronau, Singmann, & Wagenmakers, 2020). This Bayes factor compared the null hypothesis to the encompassing hypothesis which lets all parameters free to vary. Afterwards, we applied the unconditional encompassing method on the first model to estimate the proportion of prior and posterior samples in agreement with our hypothesis and again computed a Bayes factor (Gelfand, Smith, & Lee, 1992; Hoijtink, 2011; Klugkist, 2008; Klugkist et al., 2005; Sedransk, Monahan, & Chiu, 1985). This Bayes factor compared hypothesis 4 to the encompassing hypothesis which lets all parameters free to vary. Finally, we received the Bayes factor comparing hypothesis 4 to the null hypothesis by multiplying the two Bayes factors. The analysis was conducted in R (R Core Team, 2021).

Deviations from the Preregistration In our preregistration, we mentioned that the catalogue listing on which aspects the teams deviated on would span six items. However, when preparing the study materials we decided to split the aspects "operationalization of variables" into " operationalization of dependent variables" and "operationalization of independent variables" and to add the aspect "statistical test".

We preregistered that we would exclude no teams from the analyses. However, some teams did not complete all surveys and thus we were unable to calculate all relevant outcome measures. These teams were excluded from the analysis of those hypotheses for which no outcome measures could be calculated.

Concerning hypothesis 1, we preregistered to conduct a one-sided Bayesian independent samples *t*-test with "total workload" as dependent variable and "analysis method" as independent variable. We preregistered that we did not plan to transform any variables. However, after inspecting the blinded data, we decided to log transform the variable "total workload" since this variable was heavily right-skewed.

Concerning hypothesis 2, we preregistered to conduct a one-sided Bayesian Mann-Whitney test with "perceived effort" as dependent variable and "analysis method" as independent variable. After inspecting the blinded data, we decided that a Bayesian independent samples t-test would be more appropriate since we treated the variable "perceived effort" as continuous.

Concerning hypothesis 3, we preregistered that we test this hypothesis using a one-sided Bayesian Mann-Whitney test with "perceived frustration" as dependent variable and "analysis method" as independent variable. We did not change the preregistered analysis plan. Even though we treat the variable "perceived frustration" as continuous, a Mann-Whitney test seemed most appropriate since the variable did not meet the normality assumption even after we applied transformations.

7.4 Results

7.4.1 Sample Characteristics

The career stages and research backgrounds featured in each team are shown in Table 7.2. As apparent from Figure 7.1, for both conditions the teams reported less knowledge on the topic of religion and well-being (left panel; 25% and 31% of teams reported to have (some) expertise on this topic in the preregistration and analysis blinding condition, respectively) than on their knowledge on methodology and statistics (right panel; 75% and 89% of teams reported to have (some) expertise on this topic in the preregistration and analysis blinding condition, respectively).

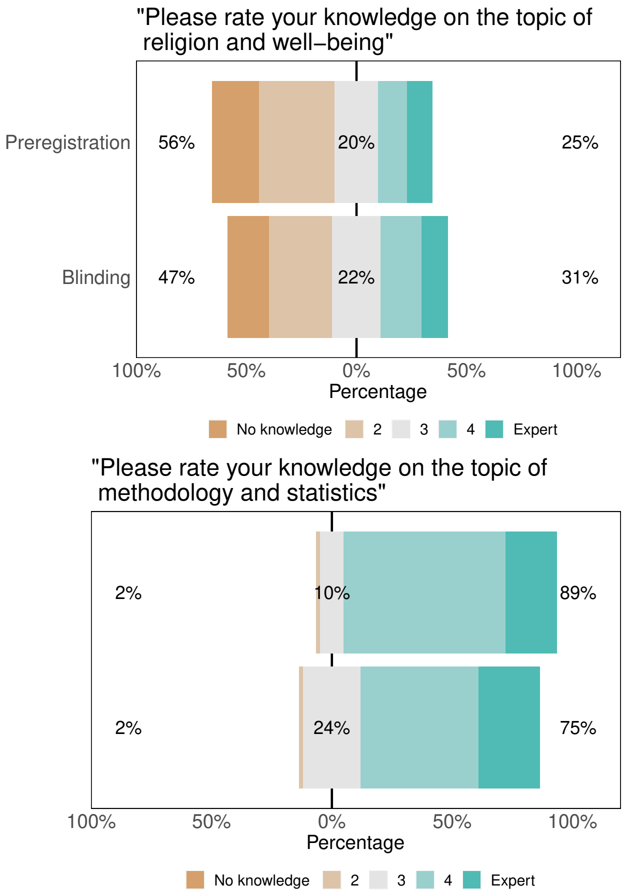


Figure 7.1: Responses to the survey questions on the teams' reported knowledge regarding religion and well-being (top panel) and knowledge regarding methodology and statistics (bottom panel). In each panel, the top bar represents responses from teams who preregistered and the bottom bar represents responses from teams who did analysis blinding. For each item, the number to the left of the data bar (in brown/orange) indicates the percentage of teams that reported little to no knowledge. The number in the center of the data bar (in grey) indicates the percentage of teams that were neutral. The number to the right of the data bar (in green/blue) indicates the percentage of teams that reported (some) expertise.

Table 7.2: Positions and domains featured in the analysis teams per condition.

	Preregistration	Analysis Blinding
Positions		
Doctoral Student	24/61 (39.34 %)	30/59 (50.85 %)
Post-doc	19/61 (31.15 %)	26/59 (44.07 %)
Assistant Professor	18/61 (29.51 %)	14/59 (23.73 %)
Associate Professor	16/61 (26.23 %)	13/59 (22.03 %)
Full Professor	7/61 (11.48 %)	10/59 (16.95 %)
Domains		
Social Psychology	24/61 (39.34 %)	19/59 (32.2 %)
Cognition	14/61 (22.95 %)	14/59 (23.73 %)
Religion and Culture	14/61 (22.95 %)	14/59 (23.73 %)
Methodology and Statistics	11/61 (18.03 %)	11/59 (18.64 %)
Health	9/61 (14.75 %)	10/59 (16.95 %)
Psychology (Other)	9/61 (14.75 %)	8/59 (13.56 %)

Note. Teams may include multiple members of the same position and in the same domain.

7.4.2 Exclusions

One team in the analysis blinding condition and one team in the preregistration condition did not fill in the stage 1 survey therefore could not be included in the analysis. In addition, one team in the preregistration condition did not report their perceived effort in the survey from stage 1 and was therefore excluded from the analysis regarding hypothesis 2. Note that one team did not report deviations because they did not submit a final analysis.

7.4.3 Confirmatory Analyses

Workload Hypothesis 1 stated that the total workload of planning and executing the analysis is *lower* for teams in the analysis blinding condition than for teams in the preregistration condition. We collected strong evidence for the null hypothesis, that is, that both teams take the same amount of time, with a Bayes factor of $\text{BF}_{0-} = 13.19$. Figure 7.2 illustrates the responses of the reported workload. Based on the descriptives, the effect seems to go in the direction opposite to our predictions, that is, the total hours spent on executing the task was in fact lower for teams in the preregistration condition ($M = 23.94$, $SD = 24.90$; log-transformed $M = 2.79$, $SD = 0.88$) than for teams in the analysis blinding condition ($M = 33.12$, $SD = 35.34$; log-transformed $M = 3.08$, $SD = 0.89$). The results are robust against different prior settings. A sequential analysis showed that as the data accumulate, the evidence in favor for the null hypothesis gradually increases.

Figure 7.3 illustrates the responses of the reported workload separately for stage 1 and stage 2. The difference in total workload spend was the largest in

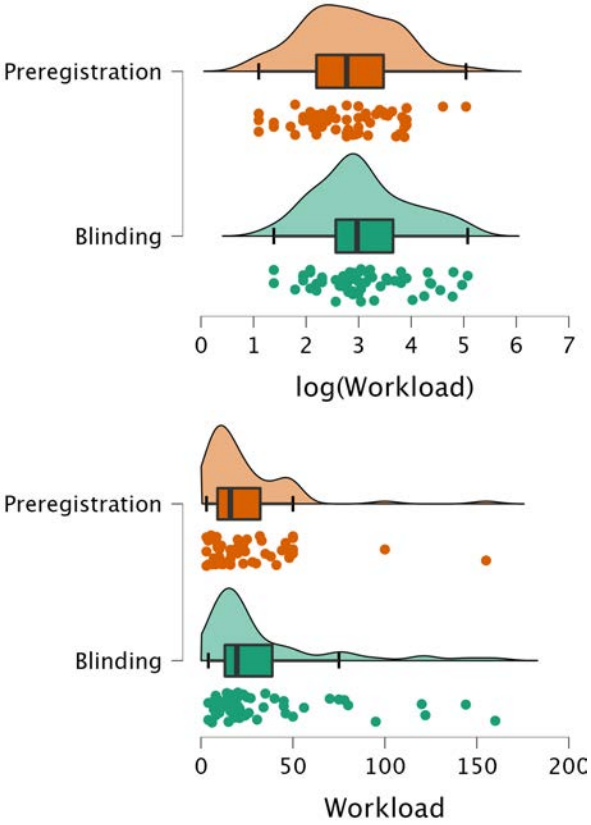


Figure 7.2: Reported total workload of stage 1 and stage 2 for each analysis team. The top panel depicts the workload on the log scale, the bottom panel on the original scale. The upper panel shows (in orange) responses of teams in the preregistration condition. The lower panel shows (in green) responses of teams in the analysis blinding condition. The data suggests strong evidence in favor of the null hypothesis that both teams take an equal amount of time planning and executing the analysis. Points are jittered to enhance visibility.

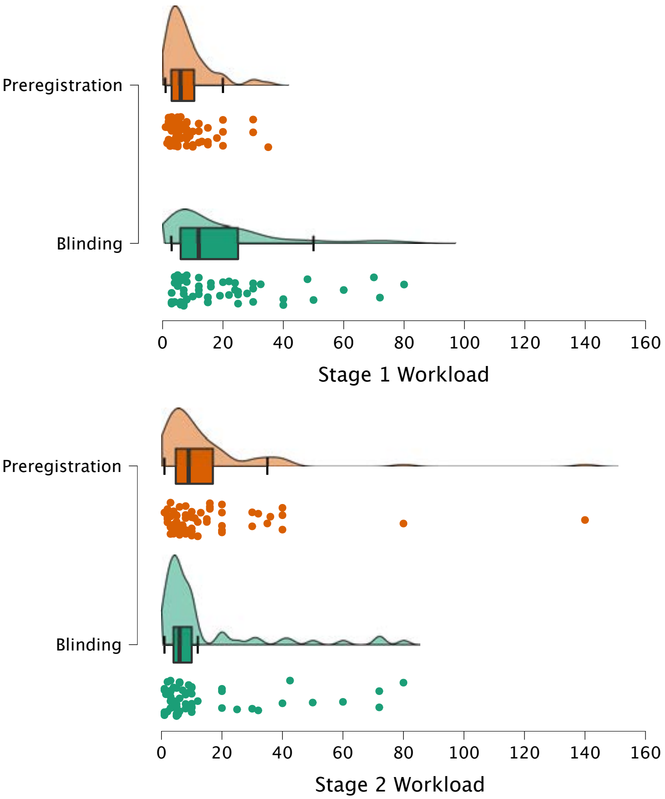


Figure 7.3: Reported total workload of stage 1 (top) and stage 2 (bottom) for each analysis team. The upper panel shows (in orange) responses of teams in the preregistration condition. The lower panel shows (in green) responses of teams in the analysis blinding condition. In stage 1, teams required more time on creating an executable script based on the blinded data than teams who created a preregistration. In stage 2, teams in both conditions required approximately the same amount of time for executing their analysis. Points are jittered to enhance visibility.

stage 1 of the project, that is, when preregistering the analysis or analyzing the blinded data. Here, teams in the analysis blinding condition took about twice as much time ($M = 19.25$) than teams in the preregistration condition ($M = 8.90$).

For stage 1, 25.0% of teams who preregistered reported that completing the task was more work than anticipated, compared to 48.3% of teams who did analysis blinding. When executing the analysis (i.e., stage 2 of the project), teams in both conditions approximately needed 15 hours to complete the task. For stage 2, 29.5% of teams who preregistered reported that this was more work than anticipated, compared to 35.6% of teams who did analysis blinding.

Perceived Effort and Frustration Hypothesis 2 stated that the perceived effort of planning and executing the analysis is *lower* for teams in the analysis blinding condition than for teams in the preregistration condition. The data were inconclusive. We found no evidence either in favor or against our hypothesis, with a Bayes factor of $\text{BF}_{-0} = 0.41$. These results are not robust against different prior settings. Depending on the prior choices, the evidence in favor of the null hypothesis fluctuates between being completely uninformative (i.e., $\text{BF}_{0-} = 0.92$) to being moderately high (i.e., $\text{BF}_{0-} = 4.52$). As the data accumulates, the evidence in favor for $\mathcal{H}_0$ fluctuates, suggesting that more data is needed to draw an informative conclusion. The left panel in Figure 7.4 illustrates the responses of teams concerning the perceived effort. Both groups reported perceived effort to be moderate to somewhat high, with an average of $M = 8.78$, $SD = 2.17$ for teams in the preregistration condition and $M = 8.44$, $SD = 2.46$ for teams in the analysis blinding condition.

Hypothesis 3 stated that the perceived frustration when planning and executing the analysis is *lower* for teams in the analysis blinding condition than for teams in the preregistration condition. We collected moderate evidence for the null hypothesis, with a Bayes factor of $\text{BF}_{0-} = 5.00$. The right panel in Figure 7.4 illustrates the responses of teams concerning the perceived frustration. Both groups reported perceived frustration to be somewhat low, with an average of $M = 5.97, SD = 2.22$ for teams in the preregistration condition and $M = 5.98, SD = 2.66$ for teams in the analysis blinding condition.

Deviation from Analysis Plan Hypothesis 4 stated that teams in the preregistration condition deviate more often from their planned analysis than teams in the analysis blinding condition and when they deviate from their analysis plan, teams in the preregistration condition deviate on more aspects than teams in the analysis blinding condition. An overview of the reported deviations are given in Table 7.3. We collected strong evidence in favor for our hypothesis, that is, $\text{BF}_{r0} = 11.40$. The estimated probability that a team would deviate from their analysis plan was almost twice as high for for teams who preregistered (i.e., 38%) compared to team who did analysis blinding (i.e., 20%).

The aspect most teams deviated from was their exclusion criteria (11 teams), the included variables in the model (9 teams), the operationalization of the independent variables (8 teams) and the statistical model (8 teams). A difference between teams who did analysis blinding and preregistration was most apparent in the exclusion criteria; from eleven teams, 10 were in the preregistration condition. Also in the operationalization of the independent variable, almost all deviations were reported by teams who preregistered (8 out of 9).

7.4.4 Exploratory Analysis

We conducted an exploratory analysis to test whether the effect of workload goes in the direction opposite to our predictions, that is, whether the total workload to plan and execute the task is *higher* for teams in the analysis blinding condition than for teams in the preregistration condition. The data suggests inconclusive evidence for this hypothesis, $\text{BF}_{+0} = 1.511$.

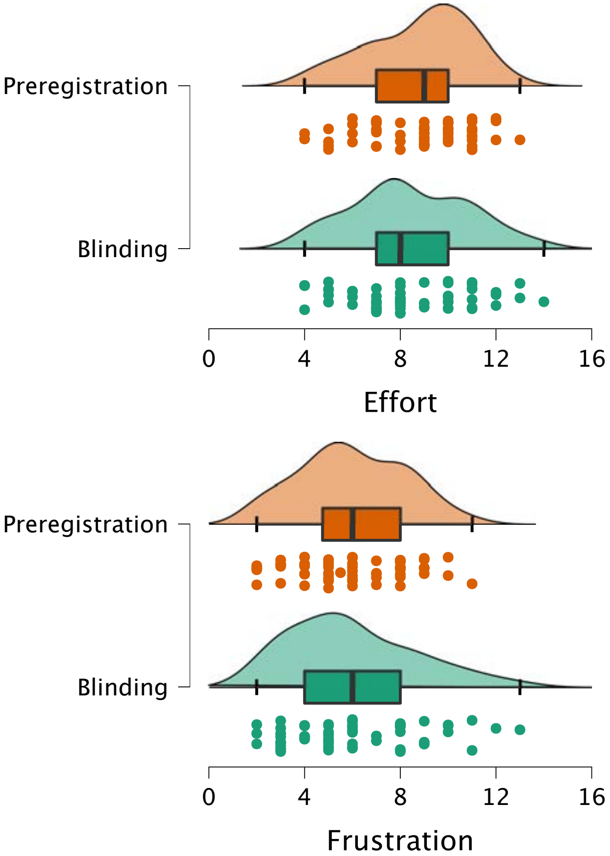


Figure 7.4: Responses to the survey questions about the perceived effort (top panel) and frustration (bottom panel) of planning and executing the analysis. The top panel shows responses of teams in the preregistration condition. The bottom panel shows responses of teams in the analysis blinding condition. The data suggests no and moderate evidence on whether analysis blinding was perceived as less effortful and frustrating, respectively. Points are jittered to enhance visibility.

Table 7.3: Reported deviations form planned analysis per condition.

	Preregistration	Analysis Blinding
Nr. of Teams Reporting Deviations	24/61 (39.34 %)	10/59 (16.95 %)
Aspects		
Exclusion Criteria	10/61 (16.39 %)	1/59 (1.69 %)
Included Variables	5/61 (8.20 %)	4/59 (6.78 %)
Operationalization of IV	8/61 (13.11 %)	1/59 (1.69 %)
Statistical Model	4/61 (6.56 %)	4/59 (6.78 %)
Statistical Test	5/61 (8.20 %)	1/59 (1.69 %)
Operationalization of DV	2/61 (3.28 %)	1/59 (1.69 %)
Hypothesis	0/61 (0 %)	0/59 (0 %)
Direction of Effect	0/61 (0 %)	0/59 (0 %)

Note. Teams may report multiple deviations.

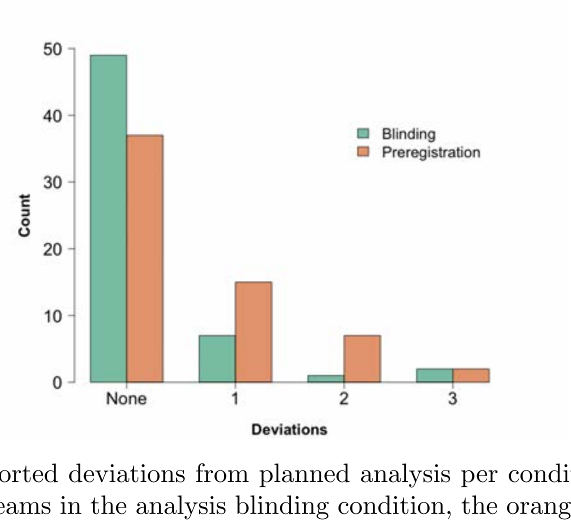


Figure 7.5: Reported deviations from planned analysis per condition. The green bars represent teams in the analysis blinding condition, the orange bars represent teams in the preregistration condition. More teams in the analysis blinding condition reported no deviations from their planned analysis and if they had deviated, they did so on less aspects than teams in the preregistration condition.

7.5 Constraints on Generality

The outcomes of this study might be dependent on the complexity of the data and hypotheses researchers are investigating. Specifically, we expect data with a simpler structure than the MARP data (i.e., non-nested structure, no composite measures) to lead to fewer deviations from the analysis plans, whereas data with a more complex structure (e.g., requiring an extensive amount of preprocessing, such as in fMRI analyses) to magnify the present results.

7.6 Discussion

The current study investigated whether analysis blinding has benefits over the preregistration of the analysis plan in terms of efficiency and convenience. We analyzed data from 120 teams participating in the Many-Analysts Religion Project who either preregistered their analysis or created a reproducible script based on blinded data. We hypothesized that analysis blinding would save researchers time, and reduce their perceived effort and frustration to complete the project. Additionally, we hypothesized that analysis blinding would lead to fewer deviations from the analysis plan.

One of the four hypotheses was supported. Compared to teams who preregistered, teams who did analysis blinding deviated less often from the analysis plan and if they did, they did so for fewer aspects. Teams in the analysis blinding condition better anticipated their final analysis strategies, particularly with respect to exclusion criteria and operationalization of the independent variable. We regard the finding that analysis blinding has a protective effect against deviations as good news for the field of meta-science, since (fear of) deviation is a well-known problem of preregistration (Claesen et al., 2021; Heirene et al., 2021; Nosek et al., 2019).

Contrary to our prediction, we found strong evidence against our hypothesis that analysis blinding would reduce workload. Teams who did analysis blinding and teams who preregistered spent approximately the same amount of time planning and executing the analysis. We assumed that teams who preregistered had a higher workload since they were required to create a preregistration document in stage 1 and write and execute this plan in stage 2. Teams who did analysis blinding wrote their analysis scripts already in stage 1 and only had to execute it in stage 2. This workload benefit for analysis blinding was expected especially since some of the proposed analyses were quite complex (including factor analyses, structural equation models, and hierarchical regression models).

Lastly, we cannot draw conclusions about the hypotheses on perceived effort and frustration since the data did not provide strong evidence either in favor of or against our hypotheses. Our data suggested moderate evidence for the hypothesis that teams in both conditions experienced equal amounts of frustration and no evidence either in favor or against the hypothesis that analysis blinding would be experienced as less effortful.

Why was workload approximately equal under preregistration versus analysis blinding? Descriptives on stage 1 showed that teams who preregistered were in fact quicker than teams who did analysis blinding. In itself, this result is not surprising: one would expect preregistration to be somewhat faster in stage 1 and that the expected benefit of analysis blinding would mostly occur in stage 2. What was surprising, however, was how much faster the teams who preregistered were in stage 1: they took only about half as much time than teams who did analysis blinding.

One explanation is be that in the current study the preregistration of the analysis was particularly simple. The literature is recommending structured workflows and templates to assist researchers with their preregistrations (Nosek et al., 2019; van 't Veer & Giner-Sorolla, 2016). That applied to the MARP in that the researchers adhered to a highly structured workflow. That is, the research questions were fixed, the teams were provided with a preregistration template, and they had access to the theoretical background of the research question and a comprehensive data documentation. In addition, since the teams analyzed preexisting data, they preregistered only their analysis plan instead of all aspects of the study (i.e., study design, sampling plan, materials).

Descriptives on stage 2 showed that teams who preregistered and teams who did analysis blinding took about the same amount of time to execute the analysis. We speculate that this result may be due to an improper communication to the teams. To complete stage 2, the teams were instructed to execute their planned analyses on the real data and fill out the post-survey to indicate their conclusions and summarize their results. We also provided teams with the type of information required to fill in the post-survey and recommendations about how to organize their OSF folder. These recommendations included to add a “ReadMe” file that documents the uploaded files and a brief summary of the main conclusions. The time associated with creating these files might have distorted our workload measure. It may be that in stage 2 most of the time was spent not on conducting the analyses but on writing the report, so that differences in workload related to the execution of the analysis may have gone undetected. If true, this would imply that differences between the two methods may not be as relevant in real-world research, where again most of the time may be spent on writing up the results rather than executing the analyses. To gain more insight into the time it takes teams to execute the analysis, future research should provide teams with instructions on how to document their files and results (or more generally speaking how to complete the project) only after workload is measured.

Lastly, future research could assess whether the quality of preregistrations is sufficiently high, or whether the quality of analyses plans are equal in both conditions. We consider an analysis plan to be of high quality if it is “specific, precise, and exhaustive” (J. M. Wicherts et al., 2016, p. 2). The quality of the submitted preregistrations could be rated with the coding protocol used by C. Veldkamp et al. (2017). However, to our knowledge there exists no comparable coding protocol for submitted analysis code, checking, for instance, its clarity and reproducibility. Such a protocol would still have to be developed and validated so that the assessments of preregistrations and analysis scripts are comparable. Along the same lines, future research could assess the quality of the final analysis, for instance, by letting participating teams rate the work of their peers. However, such a quality check should be done with caution: assessing the quality of an analysis imposes significant additional work on participating teams, is highly sensitive to subjective analytic preferences, and ignores theoretical considerations.

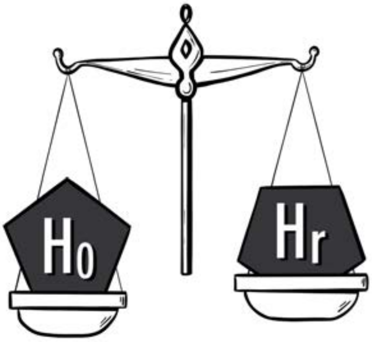
The current study mainly focused on planning and executing a confirmatory analysis. However, preregistration and analysis blinding involve other aspects as well. Specifically, we cannot draw conclusions about the perceived workload and convenience when researchers are required to preregister the whole study, including the study design, sampling plan, and materials, or when researchers need to blind a dataset first themselves, before they are handed to the analysts. Additionally, we are unable to determine how analysis blinding and preregistration compare to standard research. We deliberately decided not to include such a baseline condition since the teams answered a theoretically relevant research question and thus we saw the necessity to safeguarded the confirmatory status of all analyses.

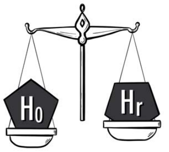
We would like to emphasize that researchers do not have to choose between preregistration and analysis blinding but they can use them in combination. In a survey by Sarafoglou, Kovacs, et al. (2021) researchers reported that preregistration benefited multiple aspects of the research process, including the research hypothesis, study design, and preparatory work. We therefore regard it as most beneficial if researchers preregister the study but finalize the statistical analysis on a blinded version of the data–in fact this was the procedure we used in the present report.

To our knowledge, this is the first study that sought to investigate analysis blinding empirically. Analysis blinding ties in with current methodological reforms for more transparency since it safeguards the confirmatory status of the analyses while simultaneously allowing researchers to explore peculiarities of the data and account for them in their analysis plan. Our results showed that analysis blinding and preregistration imply approximately the same amount of work but that in addition, analysis blinding reduced deviations from analysis plans. As such, analysis blinding constitutes an important addition to the toolbox of effective methodological reforms to combat the crisis of confidence.

Part II

Multinomial Order Restrictions





Chapter 8

Evaluating Multinomial Order Restrictions with Bridge Sampling

Abstract

Hypotheses concerning the distribution of multinomial proportions typically entail exact equality constraints that can be evaluated using standard tests. Whenever researchers formulate inequality constrained hypotheses, however, they must rely on sampling-based methods that are relatively inefficient and computationally expensive. To address this problem we developed a bridge sampling routine that allows an efficient evaluation of multinomial inequality constraints. An empirical application showcases that bridge sampling outperforms current Bayesian methods, especially when relatively little posterior mass falls in the restricted parameter space. The method is extended to mixtures between equality and inequality constrained hypotheses.

This chapter is published as Sarafoglou, A., Haaf, J. M., Ly, A., Gronau, Q. F., Wagenmakers, E. J., & Marsman, M. (2021). Evaluating multinomial order restrictions with bridge sampling. *Psychological Methods.*

8.1 Introduction

In many scientific fields the analysis of categorical variables is of major importance. Applications range from the analysis of declared numeric values in forensic accounting, auditing, and fraud detection (M. Nigrini, 2012; Rauch, Göttsche, Brähler, & Engel, 2011), the analysis of descriptive measures in survey studies (e.g., Haberman, 1978; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016; Sedransk et al., 1985; C. L. Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, 2014), the analysis of gut microbiome composition (Song, Zhao, & Wang, 2020), to the validation of model assumptions and axioms in the field of psychometrics (see e.g., Cavagnaro & Davis-Stober, 2014; Davis-Stober, 2009; Guo & Regenwetter, 2014; J. Myung, Karabatsos, & Iverson, 2005; Regenwetter et al., 2018; Regenwetter, Dana, & Davis-Stober, 2011; Tijmstra, Hoijtink, & Sijtsma, 2015). The breadth and depth of these examples underscore the importance of having efficient tools for their analysis readily available.

In each of the examples above, researchers are interested in quantifying evidence for hypotheses that impose certain restrictions on the underlying category proportions. These hypotheses often predict that all category proportions are exactly equal (e.g., the prevalence for a statistical reporting error is equal across different psychological journals; C. L. Veldkamp et al., 2014), or that they are fixed and follow a specific pattern (e.g., the digit proportions in non-fraudulent auditing data conform to Benford's law; Benford, 1938; M. Nigrini, 2012). However, research hypotheses also often stipulate ordinal expectations among category proportions (e.g., students with higher abilities have a higher chance to solve any particular item correctly; Grayson, 1988), or a mix of equality and inequality parameter constraints (e.g., according to the recognition heuristic, when laypeople predict which sports team will win a tournament they assign a higher probability of winning to more familiar teams and equal but lower probabilities to unknown teams; Goldstein & Gigerenzer, 2002).

Ordinal expectations about underlying category proportions are a regular occurrence in scientific theories. However, the evaluation of hypotheses that go beyond exact equality constraints is not very popular, particularly among researchers who use frequentist statistics (Iverson, 2006). As motivating example, consider the study conducted by Uhlenhuth, Lipman, Balter, and Stern (1974), who surveyed 735 adults to investigate the association between symptoms of mental disorders and experienced life stress. To measure participants' life stress, the authors asked them to indicate, out of a list of negative life events, life stresses, and illnesses, which event they had experienced during the last 18 months prior to the interview. A subset of these data was reanalyzed by Haberman (1978, p. 3). Haberman noted that retrospective surveys tend to fall prey to the fallibility of human memory, causing participants to report primarily those negative events that happened most recently. He therefore investigated the 147 participants who reported only one negative life event over this time span and tested whether the frequency of the reported events was equally distributed over the 18 month period. However, Haberman did not directly test the ordinal pattern implied by his assumption of forgetting, namely that the number of reported negative life events decreases as a function of the time passed. Figure 8.1 shows the frequency of reported negative life events in Haberman's sample.

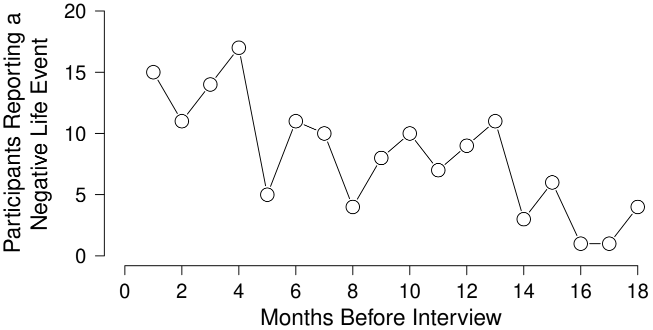


Figure 8.1: Frequency of reported negative life events over the course of the 18 months prior to the interview for Haberman's (1978) sample of the data collected by Uhlenhuth et al. (1974).

To evaluate ordinal multinomial patterns such as the one hypothesized by Haberman (1978) we focus on Bayesian methods. In the Bayesian statistical framework, researchers may quantify the evidence for or against a specific restriction on the model parameters using the Bayes factor (Jeffreys, 1935; Kass & Raftery, 1995). The Bayes factor is defined as the relative predictive performance of the models with and without the restriction, which is reflected in the ratio of their normalizing constants. For the usual scenario of equal or fixed underlying category proportions, the Bayes factor is available analytically. This is not the case, unfortunately, when inequality constraints are in play. In these cases, the Bayes factor can be approximated using the encompassing prior approach which defines the Bayes factor as the ratio of the prior and posterior probabilities that the inequality constraints hold (Klugkist et al., 2005). In the following, we highlight two particularly popular implementations of the encompassing prior approach. The first implementation is the *unconditional encompassing method* which uses straightforward sampling from the unrestricted (i.e., encompassing) distributions to compute the Bayes factor (Haaf, Merkle, & Rouder, 2020; Haaf & Rouder, 2017, 2021; Hoijtink, 2011; Hoijtink, Klugkist, & Boelen, 2008; Klugkist et al., 2005; Schnuerch, Nadarevic, & Rouder, 2020; van der Lans, Cremers, Klugkist, & Zwart, 2020). The second implementation is the *conditional encompassing method* which decomposes the Bayes factor into a product of conditional probabilities (Gu, Mulder, Deković, & Hoijtink, 2014; Laudy, 2006; Mulder, 2014, 2016; Mulder et al., 2009) and is implemented in the `R` packages `multinomineq` (Heck & Davis-Stober, 2019), `bain` (Gu, Hoijtink, Mulder, & Rosseel, 2019), `BFpack` (Mulder et al., 2021), and the software program `BIEMS` (Mulder, Hoijtink, & de Leeuw, 2012).[1]

[1] For more examples in which the conditional encompassing method has been used, see the website of the R software package `bain` (Gu et al., 2019) at `https://informative-hypotheses.sites.uu.nl/software/bain/`.

The main disadvantage of both methods is computational: the approximation of the Bayes factor becomes harder (i.e., more time-consuming and less accurate) as researchers are interested in a smaller part of the parameter space. For the unconditional encompassing method the problem is that the probability of samples falling within the parameter space of the restricted distribution is very low, making it practically impossible to obtain accurate estimates of the Bayes factor by sampling from the unrestricted distribution.

For instance, in the Haberman example the ordinal restrictions on the 18 categories are associated with a minuscule prior mass of $1/18!$ or 1 over $6,402,373,705,728,000$. Consequently, a single posterior draw that obeys the restriction will catapult the Bayes factor to extreme values unless the number of draws is impractically large. For instance, with 5 million draws a single posterior draw that obeys the restriction will yield an estimated Bayes factor of 1.28×10^9, massively favoring the inequality-constrained hypothesis. On the other hand, if none of the 5 million draws obeys the restriction, the estimated Bayes factor shows infinite support *against* the inequality-constrained hypothesis. This illustrates that in order to obtain a precise estimate of the Bayes factor, researchers need to draw millions of samples from the posterior. These ratios become increasingly problematic as the models become more complex (Mulder et al., 2009; Sedransk et al., 1985).

By decomposing the Bayes factor into a product of conditional probabilities, the conditional encompassing method is more stable than the unconditional encompassing method. However, the increased stability of the conditional encompassing method is accompanied by a steep increase in runtime. This increase has three reasons. The first reason follows directly from the sequential evaluation of the individual constraints. To refer again to our motivating: Since the associated model features seventeen constraints, seventeen sets of prior and posterior samples must be drawn for the evaluation. The resulting runtime is thus seventeen times higher than that of the unconditional encompassing method. The second reason is that even though the method is more stable, there is still the risk that the relative size of the restricted area for each individual restriction is too small to effectively sample from it (but see Gu et al., 2019 who estimate these conditional probabilities more efficiently). The third reason is the implementation of the conditional encompassing method. When evaluating the individual constraints it is not enough to simply draw samples from the unrestricted distribution; this is only possible for the first constraint. For each additional constraint, samples are drawn from distributions that are conditional on previous constraints, with a new constraint added at each step. Thus, we need to draw samples from restricted distributions using Markov chain Monte Carlo (MCMC), and this is slower than the standard Monte Carlo methods used in the unconditional encompassing method.

To overcome the above limitation we present a bridge sampling routine (e.g., Gronau et al., 2017; Meng & Wong, 1996) to estimate the Bayes factor for multinomial inequality constraints. The advantage of the bridge sampling routine is that its efficiency does not suffer when the size of the restricted parameter space decreases. The resulting Bayes factor estimates are relatively unbiased and precise. In addition, the bridge sampling approach has a fixed cost in terms of runtime, which makes it appealing for the implementation in standard statistical software packages. The bridge sampling method outlined in this chapter can be used to evaluate hypotheses that postulate (a) a monotonic increase or decrease for (a subset of) multinomial parameters (e.g., $\theta_1 < \theta_2 < \theta_3$ or $\theta_1 > \theta_2 > \theta_3$); (b) mixtures of inequality constraints and equality constraints (e.g., $\theta_1 < \theta_2 = \theta_3$); (c) mixtures of inequality constraints and free parameters (e.g., $\theta_1 < \theta_2, \theta_3$); and (d) mixtures of the first three cases (e.g., $\theta_1 < (\theta_2 = \theta_3), \theta_4$).

The outline of this chapter is as follows. First, we introduce the basic theoretical concepts of Bayesian parameter estimation and the computation of Bayes factors for the multinomial model featuring equality constrained hypotheses. We then extend these concepts to inequality constrained hypotheses to a mixture of equality and inequality constrained hypotheses. Third, we show how the bridge sampling approach compares to the established methods such as the unconditional encompassing method and the conditional encompassing method in terms of precision and efficiency by applying the methods to our motivating example. The last section contains a short discussion and the appendix compares the accuracy of the bridge sampling approach to the established methods.

8.2 Bayesian Analysis of Multinomial Variables

This section introduces the theoretical concepts of Bayesian inference for the multinomial model, that is, Bayesian parameter estimation using posterior distributions and Bayesian hypothesis testing using Bayes factors. We denote the number of observations in a category k with x_k, and the total number of observations with $N = \sum_{k=1}^{K} x_k$. The multinomial distribution is a generalization of the binomial distribution to variables that can take values in $K \geq 2$ categories, and it assigns the following probabilities to the different ways that N observations distribute across the K categories,

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = p(x_1,\, x_2,\, \ldots,\, x_K \mid \theta_1,\, \theta_2,\, \ldots,\, \theta_K) = \binom{N}{x_1,\, x_2,\, \ldots,\, x_K} \prod_{k=1}^{K} \theta_k^{x_k},$$

where the first factor in the likelihood denotes an extension of the binomial coefficient known as the multinomial coefficient. The parameters of the multinomial model, θ_k, reflect the probability of observing a value in a particular category, and need to sum to one. Note that due to the sum-to-one constraint, the K-th parameter is sometimes expressed as $\theta_K = 1 - \sum_{k=1}^{K-1} \theta_k$.

8.2.1 Prior and Posterior Distribution Without Inequality Constraints

Bayesian parameter estimation concerns the expression of a posterior distribution for model parameters capturing *a priori* information and information from the data (i.e., the likelihood). For the vector of probability parameters, $\boldsymbol{\theta} = (\theta_1,\, \theta_2,\, \ldots,\, \theta_K)$, we choose a Dirichlet distribution with concentration parameters ($\alpha_1, \alpha_2, \ldots, \alpha_K$), where each element in $\boldsymbol{\alpha}$ is larger than zero:

$$p(\boldsymbol{\theta}) = p(\theta_1, \theta_2, \ldots, \theta_K) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}.$$

The concentration parameters α_k of the Dirichlet distribution have an intuitive interpretation: they may be interpreted as *a priori* category counts, and their exact values determine both the relative values of category probabilities and their variability. For the problem at hand, the posterior is also a Dirichlet distribution of the form

$$p(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{\Gamma\left(N + \sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(x_k + \alpha_k)} \prod_{k=1}^{K} \theta_k^{x_k + \alpha_k - 1},$$

with the updated concentration parameters $\alpha'_k = x_k + \alpha_k$ (O'Hagan & Forster, 2004). The concentration parameters of the posterior Dirichlet distribution can be interpreted as *a posteriori* category counts, the sum of the prior and observed category counts.

8.2.2 Bayes Factor Hypothesis Testing Without Inequality Constraints

When stipulating exact equality constraints on the parameters of interest, researchers formulate a point null hypothesis $\mathcal{H}_0$ that assigns expected values $\mathbf{c}$ to the underlying category proportions $\boldsymbol{\theta}$, that is $\mathcal{H}_0 : \boldsymbol{\theta} \ = \mathbf{c}$. We first consider the Bayes factor

$$\text{BF}_{0e} = \frac{p(\mathbf{x} \mid \mathcal{H}_0)}{p(\mathbf{x} \mid \mathcal{H}_e)},$$

which is defined as the ratio of normalizing constants of the null hypothesis and the encompassing hypothesis. Here, the hypothesis $\mathcal{H}_0$ stipulates exact values for all of the model parameters, i.e., $\mathcal{H}_0 : \boldsymbol{\theta} = \mathbf{c}$. In the standard multinomial test the null hypothesis states that all model parameters are exactly equal. Since the parameters are sum-to-one constrained, it follows that all elements in $\mathbf{c}$ are set equal to $1/K$. We test the null hypothesis against the encompassing hypotheses which states that all category proportions are free to vary without any ordinal restrictions. We call this hypothesis the encompassing hypothesis $\mathcal{H}_e$, since it encompasses all possible orders of the parameters. The parameter space of the encompassing hypothesis is denoted as $\mathcal{R}_e$. When stipulating exact equality constraints, it is assumed that there is no prior uncertainty about the model parameters, and the marginal likelihood of the null hypothesis is simply a multinomial distribution. Due to the conjugacy of the Dirichlet distribution to the parameters of the multinomial model, the marginal likelihood for the encompassing hypothesis has a simple, closed-form solution. Thus, if all model parameters of the null hypothesis are *a priori* specified, the Bayes factor BF_{0e} is equal to

$$\text{BF}_{0e} = \prod_{k=1}^{K} c_k^{x_k} \times \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)} \times \frac{\Gamma\left(N + \sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k + x_k)},$$

as derived already by Good (1967). There is another way to express the Bayes factor, which relates the Bayes factor to Bayesian parameter estimation. By rearranging Bayes' rule the marginal likelihood of the encompassing hypothesis can be expressed as:

$$\underbrace{p(\mathbf{x} \mid \mathcal{H}_e)}_{\substack{\text{marginal}\\\text{likelihood}\\\text{of } \mathcal{H}_e}} = \frac{\overbrace{p(\mathbf{x} \mid \boldsymbol{\theta}, \mathcal{H}_e)}^{\text{likelihood}}\ \overbrace{p(\boldsymbol{\theta} \mid \mathcal{H}_e)}^{\substack{\text{prior}\\\text{density}}}}{\underbrace{p(\boldsymbol{\theta} \mid \mathbf{x}, \mathcal{H}_e)}_{\substack{\text{posterior}\\\text{density}}}},$$

which is known as Chib's identity (Chib, 1995). Chib's identity allows us to arrive at an alternative characterization of the Bayes factor that only requires the prior and posterior distribution under the alternative hypothesis at $\mathbf{c}$:

$$\text{BF}_{0e} = \frac{p(\mathbf{x} \mid \mathcal{H}_0)}{p(\mathbf{x} \mid \mathcal{H}_e)} = \frac{p(\mathbf{x} \mid \boldsymbol{\theta} = \mathbf{c}, \mathcal{H}_e)}{\frac{p(\mathbf{x} \mid \boldsymbol{\theta}=\mathbf{c}, \mathcal{H}_e)\, p(\boldsymbol{\theta}=\mathbf{c} \mid \mathcal{H}_e)}{p(\boldsymbol{\theta}=\mathbf{c} \mid \mathbf{x}, \mathcal{H}_e)}} = \frac{\overbrace{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathbf{x}, \mathcal{H}_e)}^{\substack{\text{Height of posterior density of } \mathcal{H}_e\\ \text{at } \theta = c}}}{\underbrace{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathcal{H}_e)}_{\substack{\text{Height of prior density of } \mathcal{H}_e\\ \text{at } \theta = c}}}.$$

This expression is known as the Savage-Dickey density ratio (Dickey, 1971; Dickey & Lientz, 1970; O'Hagan & Forster, 2004; Verdinelli & Wasserman, 1995). The underlying principle of the Savage-Dickey density ratio is to compute the Bayes factor by dividing the height of the posterior density under $\mathcal{H}_e$ at the point of interest (i.e., $\mathbf{c}$) by the height of the prior density under $\mathcal{H}_e$ at the same point.

For concreteness, we will demonstrate the Bayesian multinomial test for exact equality constraints by reanalyzing the research question of Habermann (1978). The null hypothesis entails that the probability of reporting a negative life event is equally distributed over the 18 months prior to the interview. In particular, the expected category proportions under $\mathcal{H}_0$ are

$$\mathbf{c} : \theta_1, \theta_2, \cdots, \theta_K = {}^1\!/\!{}_K.$$

Assuming that every parameter value is equally likely before we see any data, we assign a uniform prior distribution across the parameter vector $\boldsymbol{\theta}$, such that, $p(\boldsymbol{\theta} \mid \mathcal{H}_e) \sim \text{Dirichlet}(\boldsymbol{\alpha})$ with all concentration parameters set to 1. Using the observed frequencies from Haberman (1978), that is,

$$\mathbf{x} = (15, 11, 14, 17, 5, 11, 10, 4, 8, 10, 7, 9, 11, 3, 6, 1, 1, 4)',$$

the Bayes factor comparing the null and encompassing hypotheses is:

$$\text{BF}_{0e} = \frac{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathbf{x}, \mathcal{H}_e)}{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathcal{H}_e)} = \frac{\frac{\Gamma\left(\sum_{k=1}^K \alpha_k + x_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k + x_k)} \prod_{k=1}^K \theta_k^{x_k + \alpha_k - 1}}{\frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}} = \frac{1}{27.1}.$$

This result indicated that the data are about 27 times more likely under $\mathcal{H}_e$ (in which the parameters are free to vary) than under $\mathcal{H}_0$ (in which the parameters are constrained to be equal).

Sometimes it is desirable to compare the null hypothesis $\mathcal{H}_0$ not against the encompassing hypothesis $\mathcal{H}_e$, but against a more informed hypothesis $\mathcal{H}_1$ that makes specific theoretically-motivated predictions. With Bayes factors against the encompassing hypotheses in hand, the desired comparison between the null and the informed alternative can be obtained through transitivity:

$$\mathrm{BF}_{01} = \frac{\mathrm{BF}_{0e}}{\mathrm{BF}_{1e}} = \frac{\dfrac{p(\mathbf{x} \mid \mathcal{H}_0)}{p(\mathbf{x} \mid \mathcal{H}_e)}}{\dfrac{p(\mathbf{x} \mid \mathcal{H}_1)}{p(\mathbf{x} \mid \mathcal{H}_e)}} = \frac{p(\mathbf{x} \mid \mathcal{H}_0)}{p(\mathbf{x} \mid \mathcal{H}_1)}.$$

This transitivity property is especially relevant when comparisons are made with restricted hypotheses, since it is more challenging to compute BF_{01} directly. Furthermore, we can use posterior model probabilities to compare the relative plausibility of any number of hypotheses (Berger & Molina, 2005). Assuming that all hypotheses are equally likely *a priori*, the posterior model probability for a particular hypothesis $\mathcal{H}_1$ is defined as:

$$p(\mathcal{H}_1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathcal{H}_1)}{\sum_{i=1}^{I} p(\mathbf{x} \mid \mathcal{H}_i)} = \frac{\mathrm{BF}_{1e}}{\sum_{i=1}^{I} \mathrm{BF}_{ie}}.$$

Here we have outlined how to express the prior and posterior distribution for the multinomial model using a Dirichlet prior. In addition, we expressed the Bayes factor in terms of the change of belief about the parameter value and outlined how to compare multiple hypotheses by utilizing the transitivity property of the Bayes factor and posterior model probabilities. A related expression for the Bayes factor can be derived in the case of inequality constraints, to which we turn next.

8.2.3 Prior and Posterior Distribution With Inequality Constraints

When stipulating inequality-constrained hypotheses we can predict, for instance, an increasing trend of the first two categories, $\mathcal{H}_r : \theta_1 < \theta_2$. We refer to such inequality-constrained hypotheses as $\mathcal{H}_r$. Here, the parameter space, $\mathcal{R}_r$ is a subset of $\mathcal{R}_e$ by restrictions imposed on $\boldsymbol{\theta}$, that is, $\mathcal{R}_r = \{\boldsymbol{\theta} \in \mathcal{R}_e \,;\, \mathcal{H}_r\}$. The prior and posterior distributions of the parameters subject to an inequality-constrained hypothesis $\mathcal{H}_r$ thus take the following form:

$$p(\boldsymbol{\theta} \mid \mathcal{H}_r) = \frac{p(\boldsymbol{\theta} \mid \mathcal{H}_e)\, \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)} \tag{8.1}$$

$$p(\boldsymbol{\theta} \mid \mathbf{x}, \mathcal{H}_r) = \frac{p(\boldsymbol{\theta} \mid \mathbf{x}, \mathcal{H}_e)\, \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)}, \tag{8.2}$$

where $\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)$ is an indicator function that is one for parameter values $\boldsymbol{\theta}$ in the restricted space $\mathcal{R}_r$ and zero otherwise. As apparent from the equations above, the prior and posterior distributions under an inequality-constrained hypothesis are proportional to their unconstrained counterparts. In principle, whenever the concentration parameters in the Dirichlet distribution are natural numbers for every k, thus, $\alpha_k \in \mathbb{N}$, we are able to achieve an exact result for the normalizing constants for the restricted prior and posterior distribution (see our online appendix for a description of the exact procedure).[2] However, the exact procedure is far more inefficient than sampling-based methods, especially as the number of categories in the model and the number of observations for a fixed K increases. Here, we were only able to obtain exact results for simple cases, involving models with no more than $K = 6$ categories and no more than $N = 63$ observations. For this reason, in the following we limit our descriptions of Bayesian parameter estimation and the computation of Bayes factors to sampling-based procedures.

In general researchers rely on Monte Carlo sampling methods to compute the normalizing constants of the restricted prior and posterior distribution. In the simplest case we can use rejection sampling to simulate values from the unconstrained prior and posterior distributions and only keep those values that conform to the restrictions. The proportion of the retained samples to the total number of samples is then an approximation for the normalizing constant of the restricted distribution. Unfortunately, when many inequality constraints are proposed, the approach outlined above, although intuitive, can be terribly inefficient. For instance, in the Haberman example, when drawing from a uniform prior only 1 in over $18! = 6.4 \times 10^{15}$ samples will obey the restriction. As an alternative, we can use a MCMC approach, that allows us through random variable transformation to simulate the values directly from the restricted distribution. Devroye (1986, p. 594), for instance, shows that one can simulate values from a Dirichlet distribution by first simulating K independent random variables γ_k with a Gamma(α_k, 1) density, for $k = 1, \ldots, K$, and then setting

$$\theta_k = \frac{\gamma_k}{\sum_{k=1}^{K} \gamma_k}.$$

The variables θ_k that are generated in this way follow the desired Dirichlet($\boldsymbol{\alpha}$) distribution (see Klugkist, Laudy, & Hoijtink, 2010 for an application in the context of contingency tables). Note that with the transformation from $\boldsymbol{\theta}$ to $\boldsymbol{\gamma}$ the sum-to-one constraint is conveniently removed. Additionally, this MCMC method is suitable for drawing values from the restricted distribution because the transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ is order-preserving. Thus, an inequality-constrained hypothesis $\mathcal{H}_r : \theta_1 < \theta_2$ on the category probabilities translates into the inequality-constrained hypothesis $\mathcal{H}_r : \gamma_1 < \gamma_2$ on the gamma variables. If we simulate the gamma variables consistent with the order restrictions imposed by $\mathcal{H}_r$, that is, $p(\boldsymbol{\gamma} \mid \mathcal{H}_r)$, the transformed gamma variables then generate Dirichlet variables that are consistent with $\mathcal{H}_r$, that is, $p(\boldsymbol{\theta} \mid \mathcal{H}_r)$.

To draw gamma variables that obey the order imposed by the inequality-constrained hypotheses we use the Gibbs sampling algorithm proposed by Damien

[2] Whenever the concentration parameters are natural numbers and we express the problem in terms of stick-breaking parameters the computations involve integrating polynomials, which makes the result exact. For general $\boldsymbol{\alpha}$, however, we do not have polynomials and thus an exact result is not expected.

and Walker (2001). Their Gibbs sampling algorithm assumes fixed upper and lower bounds for each parameter. However, the algorithm can easily be generalized to cases where we wish to draw from gamma variables whose upper and lower bounds are not known, but are itself random variables (as it is the case for inequality-constrained hypotheses).

Instead of simulating values directly from the multivariate distribution of gamma variables that are subject to inequality constraints —$p(\boldsymbol{\gamma} \mid \mathcal{H}_r)$—, the Gibbs sampler operates by iteratively simulating values from the full-conditional posterior distributions, that is, the distribution of one gamma variable given the remaining gamma variables and inequality constraints —$p(\gamma_k \mid \boldsymbol{\gamma}^{(k)}, \mathcal{H}_r)$, where $\boldsymbol{\gamma}^{(k)}$ refers to the vector of gamma variables with the kth parameter removed. If there is no constraint on a gamma variable γ_k then the full conditional is simply the regular Gamma$(\alpha_k, 1)$ density. However, if γ_k is subject to a constraint, for instance, $\gamma_j < \gamma_k < \gamma_q$, then the gamma variable γ_k has the bounded support $[\gamma_j, \gamma_q]$ instead of $[0, \infty)$. This implies that the full conditional distribution of γ_k subject to an inequality constraint is a truncated gamma distribution:

$$p(\gamma_k \mid \boldsymbol{\gamma}^{(k)}, \mathcal{H}_r) = p(\gamma_k \mid \gamma_j < \gamma_k < \gamma_q) = \frac{\frac{1}{\Gamma(\gamma_k)}\, \gamma_k^{\alpha_k - 1}\, e^{-\gamma_k}\, \mathbb{I}(\gamma_k \in [\gamma_j, \gamma_q])}{p(\gamma_k \in [\gamma_j, \gamma_q])}.$$

For gamma variables with bounded support $[\gamma_j, \gamma_q]$, the bounds at iteration t are calculated using the current values of the parameters. After the gamma variables have been simulated in this manner, they can be transformed back into category probabilities to yield samples from the Dirichlet distribution. Sampling from the prior and posterior is useful when we wish to estimate the parameters or when draws from restricted distributions are required to compute the Bayes factor, as is the case with the conditional encompassing method and the bridge sampling method.

8.2.4 Bayes Factor Hypothesis Testing for Inequality Constraints

We consider the Bayes factor

$$\text{BF}_{re} = \frac{p(\mathbf{x} \mid \mathcal{H}_r)}{p(\mathbf{x} \mid \mathcal{H}_e)},$$

where the hypothesis $\mathcal{H}_r$ stipulates inequality constraints on the model parameters, for instance,

$$\mathcal{H}_r : \theta_1 < \cdots < \theta_K.$$

In order to obtain the marginal likelihood of the inequality-constrained hypothesis we need to integrate over the restricted parameter space $\mathcal{R}_r$, which makes the Bayes factor BF_{re} difficult to compute:

$$p(\mathbf{x} \mid \mathcal{H}_r) = \int_{\mathcal{R}_e} p(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{H}_r)\, \mathrm{d}\boldsymbol{\theta}. = \frac{\int_{\mathcal{R}_r} p(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{H}_e)\, \mathrm{d}\boldsymbol{\theta}}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)}$$

It is nevertheless possible to arrive at an intuitive expression of the Bayes factor. This expression is a generalization of the Savage-Dickey density ratio mentioned above and follows from an alternative characterization of $p(\mathbf{x} \mid \mathcal{H}_r)$:

$$p(\mathbf{x} \mid \mathcal{H}_r) = \frac{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x},\, \mathcal{H}_e)\, p(\mathbf{x} \mid \mathcal{H}_e)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)},$$

which was derived in Klugkist et al. (2005). With this characterization the Bayes factor amounts to

$$\mathrm{BF}_{re} = \frac{\frac{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x},\, \mathcal{H}_e)\, p(\mathbf{x} \mid \mathcal{H}_e)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)}}{p(\mathbf{x} \mid \mathcal{H}_e)} = \frac{\overbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x},\, \mathcal{H}_e)}^{\substack{\text{Proportion of posterior parameter} \\ \text{space consistent with the restriction}}}}{\underbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)}_{\substack{\text{Proportion of prior parameter} \\ \text{space consistent with the restriction}}}}. \tag{8.3}$$

Like the Savage-Dickey density ratio, this presents the Bayes factor as the change of belief that the parameters lie in the restricted parameter space $\mathcal{R}_r$ (see also Wetzels, Grasman, & Wagenmakers, 2010 and Mulder et al., 2009). We discuss two established procedures to approximate the Bayes factor BF_{re} in the next section.

8.2.5 Established Procedures to Estimate the Bayes Factor For Inequality-Constraints

One popular method to estimate the Bayes factor for inequality-constrained hypotheses is the unconditional encompassing method which relies on simple Monte Carlo estimates (Gelfand et al., 1992; Klugkist et al., 2005; Sedransk et al., 1985). This method estimates the Bayes factor in Equation 8.3 by considering the proportion of the prior and posterior distributions of the unrestricted distribution that are in agreement with the constraints. That is, the numerator can be estimated by sampling from the encompassing posterior density and then calculating the proportion of draws in accordance with the restrictions imposed by the inequality-constrained hypothesis. Likewise, the denominator can be estimated by sampling from the encompassing prior density and then calculating the proportion of draws in accordance with the restrictions:

$$\begin{aligned}\mathrm{BF}_{re} &= \frac{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x},\, \mathcal{H}_e)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)} \\ &\approx \frac{\frac{1}{S}\sum_{s=1}^{S} \mathbb{I}(\boldsymbol{\theta}'_s \in \mathcal{R}_r)}{\frac{1}{S}\sum_{s=1}^{S} \mathbb{I}(\boldsymbol{\theta}^*_s \in \mathcal{R}_r)},\end{aligned}$$

where $\boldsymbol{\theta}^*_s$ and $\boldsymbol{\theta}'_s$ denote the s-th sample from the encompassing prior and posterior distribution, respectively, for samples $s = 1, \ldots, S$. The simplicity of the method has stimulated application to numerous statistical scenarios, ranging from contingency tables and the analysis of variance and covariance models (Hoijtink et al., 2008), to item response theory (Haaf et al., 2020), to meta-analysis models (Haaf & Rouder, 2021), to linear mixed models (Haaf & Rouder, 2017), and to circular mixed effects models (van der Lans et al., 2020). However, it is also widely recognized that this method is not particularly efficient for models with an increasing number of independent constraints (J. Myung, Karabatsos, & Iverson, 2008; Sedransk et al., 1985). The same holds true for models with a small number of constraints that are extremely restrictive or models for which the data do not align with the inequality-constrained hypothesis. This is the case because the efficiency of the method relies on the relative size of the restricted area: if prior and posterior samples almost never fall inside the area of interest, a large number of samples is required to estimate the proportions accurately (Gelfand et al., 1992; Hoijtink, 2011).

A method that is more stable for larger models is the conditional encompassing method (Mulder et al., 2009; for an application to multinomial models, see Heck & Davis-Stober, 2019). The conditional encompassing method also utilizes the identity in Equation 8.3. But instead of estimating the normalizing constants of the constrained distribution based on a single set of samples from the encompassing distribution, Mulder et al. (2009) proposed a stepwise approach. For instance, when evaluating a hypotheses concerning $K = 4$ ordered parameters $\mathcal{H}_r : \theta_1 < \theta_2 < \theta_3 < \theta_4$, the proportion of prior parameter space consistent with the restriction can be factored as follows:

$$p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e) = p(\theta_1 < \theta_2 \mid \mathcal{H}_e) \times p(\theta_2 < \theta_3 \mid \theta_1 < \theta_2, \mathcal{H}_e) \times p(\theta_3 < \theta_4 \mid \theta_1 < \theta_2 < \theta_3, \mathcal{H}_e).$$

The proportion of posterior samples consistent with the restriction are estimated in a similar fashion, which yields the Bayes factor:

$$\begin{aligned}\text{BF}_{re} &= \frac{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)} \\ &= \frac{p(\theta_1 < \theta_2 \mid \mathbf{x}, \mathcal{H}_e) \times \cdots \times p(\theta_3 < \theta_4 \mid \theta_1 < \theta_2 < \theta_3, \mathbf{x}, \mathcal{H}_e)}{p(\theta_1 < \theta_2 \mid \mathcal{H}_e) \times \cdots \times p(\theta_3 < \theta_4 \mid \theta_1 < \theta_2 < \theta_3, \mathcal{H}_e)},\end{aligned}$$

where each Bayes factor is estimated independently. By evaluating the constraints sequentially, the conditional encompassing method yields better results for models featuring larger numbers of constraints (Mulder et al., 2009). It is noteworthy that the conditional encompassing method was first used to evaluate almost-equality constraints: using the transitivity property of the Bayes factor, Laudy (2006, p. 115) and Klugkist (2008) proposed to approximate the Bayes factor for almost-equality constraints by evaluating a series of hypotheses of increasing narrowness, such that for each pair of parameters $\theta_1 \approx \theta_2$ the distance between the them approaches zero (i.e., $\mid \theta_1 - \theta_2 \mid \to 0$).[3] However, care must be taken not to set the values for the distance $\mid \theta_1 - \theta_2 \mid$ too small, or otherwise the method becomes inefficient (Klugkist, 2008).

[3] Wetzels et al. (2010) showed that the proposed almost-equality constrained method approximates the Savage-Dickey density ratio.

8.2.6 A Bridge Sampling Routine to Estimate the Bayes Factor

The main limitations of the unconditional encompassing method and the conditional encompassing method–lack of precision, lack of scalability, and long runtimes–come from the effort to estimate the proportion of the encompassing parameter space in accordance with the constraint. In contrast, bridge sampling (C. H. Bennett, 1976; Meng & Wong, 1996) estimates the Bayes factor using a different approach. The basic principle of bridge sampling is that the ratio between two normalizing constants operating on the same parameter space can be estimated by the following identity:

$$\mathrm{BF}_{12} = \frac{p(\mathbf{x} \mid \mathcal{H}_1)}{p(\mathbf{x} \mid \mathcal{H}_2)} = \frac{\mathbb{E}_{\mathcal{H}_2}\left(p(\mathbf{x} \mid \boldsymbol{\theta}, \mathcal{H}_1) p(\boldsymbol{\theta} \mid \mathcal{H}_1) h(\boldsymbol{\theta})\right)}{\mathbb{E}_{\mathcal{H}_1}\left(p(\mathbf{x} \mid \boldsymbol{\theta}, \mathcal{H}_2) p(\boldsymbol{\theta} \mid \mathcal{H}_2) h(\boldsymbol{\theta})\right)}, \tag{8.4}$$

where the term $h(\boldsymbol{\theta})$ refers to a bridge function that ensures that the denominator is non-zero. In this case we choose the optimal bridge function as proposed by Meng and Wong (1996). Instead of estimating the Bayes factor directly, we use a modified form of the bridge identity proposed by Overstall and Forster (2010) which estimates only a single normalizing constant instead of the Bayes factor to further increase the precision of the estimates (for a tutorial on the bridge sampling method, see Gronau et al., 2017). The modified form of the bridge identity requires that the second distribution is chosen such that it has overlapping support with the target distribution and has a known normalizing constant. In the following, we will refer to this distribution as proposal distribution $g(\boldsymbol{\theta})$. The modified identity then becomes:

$$p(\mathbf{x} \mid \mathcal{H}_1) = \frac{\mathbb{E}_{g(\boldsymbol{\theta})}\left(p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{H}_1) h(\boldsymbol{\theta})\right)}{\mathbb{E}_{\mathcal{H}_1}\left(g(\boldsymbol{\theta}) h(\boldsymbol{\theta})\right)}, \tag{8.5}$$

where $p(\mathbf{x} \mid \mathcal{H}_1)$ indicates a normalizing constant we wish to estimate, that is, the normalizing constant of the constrained prior distribution, or the normalizing constant of the constrained posterior distribution, that is, $p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)$ or $p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)$, respectively. Since these normalizing constants are of the form

$$\int_{\mathcal{R}_r} p(\boldsymbol{\theta} \mid \mathcal{H}_e)\, \mathrm{d}\boldsymbol{\theta} \quad \text{and} \quad \int_{\mathcal{R}_r} p(\boldsymbol{\theta} \mid \mathbf{x}, \mathcal{H}_e)\, \mathrm{d}\boldsymbol{\theta}$$

the bridge sampler can be used to estimate them, if the support of the proposal distribution $g(\boldsymbol{\theta})$ is $\mathcal{R}_r$. That is, the restricted distribution and the proposal distribution need to operate on the same parameter space. As we will discuss in the next section, we will facilitate this overlapping support by applying a series of transformation on the parameters of the restricted distribution.

To arrive at the expression for the bridge sampling identity for the normalizing constant of the constrained prior distribution we now simply replace the terms related to $\mathcal{H}_1$. Specifically, since

$$p(\boldsymbol{\theta} \mid \mathcal{H}_r) = \frac{p(\boldsymbol{\theta} \mid \mathcal{H}_e)\, \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)},$$

we can replace the term for the unnormalized density under $\mathcal{H}_1$ in the numerator of Equation 8.5 (i.e., $p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{H}_1)$) by the corresponding term for the constrained prior distribution, that is, $p(\boldsymbol{\theta} \mid \mathcal{H}_e)\,\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)$. Thus, the resulting bridge sampling identity can be described as follows:

$$p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e) = \frac{\mathbb{E}_{g(\boldsymbol{\theta})}\left(p(\boldsymbol{\theta} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)h(\boldsymbol{\theta})\right)}{\mathbb{E}_{\text{prior}}\left(g(\boldsymbol{\theta})h(\boldsymbol{\theta})\right)}. \tag{8.6}$$

The normalizing constant for the constrained posterior distribution can be described similarly. Based on this identity, we can now define the corresponding estimator. We substitute the expectations by sample averages, using N_1 samples from the constrained prior distribution, that is, $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta} \mid \mathcal{H}_r)$ and N_2 samples from a suitable proposal distribution, that is $\tilde{\boldsymbol{\theta}} \sim g(\boldsymbol{\theta})$. Then, we can estimate $p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)$ by:

$$\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e) \approx \frac{\frac{1}{N_2}\sum_{m=1}^{N_2} p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \mid \mathcal{H}_e)\mathbb{I}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \in \mathcal{R}_r)h(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\frac{1}{N_1}\sum_{n=1}^{N_1} g(\boldsymbol{\theta}^*_{\boldsymbol{n}})h(\boldsymbol{\theta}^*_{\boldsymbol{n}})}. \tag{8.7}$$

There are many possible choices for $h(\boldsymbol{\theta})$. Meng and Wong (1996) suggested the use of a bridge function that has been shown to minimize the relative mean square error of the estimate. However, when following this recommendation, the specific choice for $h(\boldsymbol{\theta})$ depends on the unknown normalization constant:

$$h(\boldsymbol{\theta}) = c \times \frac{1}{s_1 p(\boldsymbol{\theta} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r) + s_2 p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)g(\boldsymbol{\theta})},$$

where $s_1 = \frac{N_1}{N_2+N_1}$, $s_2 = \frac{N_2}{N_2+N_1}$ and c is a constant that has no influence on the results. To be able to estimate the normalizing constant of the constrained prior distribution we use the iterative scheme proposed by Meng and Wong (1996). Thus, we yield the following formula for the bridge sampling estimator at iteration $t+1$:

$$\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)} \approx \frac{\frac{1}{N_2}\sum_{m=1}^{N_2} \frac{\ell_{2,m}}{s_1\ell_{2,m} + s_2 p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)}}}{\frac{1}{N_1}\sum_{n=1}^{N_1} \frac{1}{s_1\ell_{1,n} + s_2 p(\boldsymbol{\theta}^*_{\boldsymbol{n}} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)}}}, \tag{8.8}$$

where $\ell_{1,n} = \dfrac{p(\boldsymbol{\theta}^*_{\boldsymbol{n}} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta}^*_{\boldsymbol{n}} \in \mathcal{R}_r)}{g(\boldsymbol{\theta}^*_{\boldsymbol{n}})}$ and $\ell_{2,m} = \dfrac{p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \mid \mathcal{H}_e)\mathbb{I}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \in \mathcal{R}_r)}{g(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}$.[4] We then run the iterative scheme until a predefined tolerance criterion is reached. We follow

[4] Equation 8.8 illustrates another advantage of bridge sampling: its robustness to the tail

the suggestion by Gronau et al. (2017) to use a tolerance criterion of

$$\frac{\mid \hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)} - \hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)} \mid}{\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)}} \leq 10^{-10},$$

while setting $\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(1)} = 0$ as initial guess. To summarize, we can use bridge sampling to separately estimate the normalizing constants for the restricted prior distribution and the restricted posterior distribution. Then, we can use these two estimates to compute the Bayes factor:

$$\text{BF}_{re} \approx \frac{\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)}{\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)}.$$

8.2.7 Transformations To Facilitate Bridge Sampling

Since the bridge function is defined on the common support of the proposal and target distribution, both distributions have to operate on the same parameter space. In addition, the normalizing constant of the proposal distribution must be known, which means that we cannot choose another constrained Dirichlet distribution. To resolve this problem we move the prior and posterior draws from the probability space to the real line using a probit transformation. This transfomation aims to eliminate the constraints inherent to the restricted Dirichlet distribution, namely the sum-to-one constraint and the inequality constraints. Furthermore, the transformation enables us to choose a convenient proposal distribution that is easy to sample from and easy to evaluate, for instance, the multivariate normal distribution (Overstall & Forster, 2010).

The general idea is as follows: $\boldsymbol{\theta}$ is a probability vector, therefore, its elements must sum to one. As a result, the vector is completely determined by its first $K-1$ elements. For the transformation we therefore only consider the first $K-1$ elements and transform them to $K-1$ elements of a new vector $\boldsymbol{\xi}$ with $\boldsymbol{\xi} \in \mathbb{R}^{K-1}$. To retain the inequality constraints imposed on the parameters, we need to account for the lower bound l_k and the upper bound u_k of each θ_k. These bounds can be determined by adapting a stick-breaking approach (Frigyik, Kapila, & Gupta, 2010; Stan Development Team, 2021). The stick-breaking approach represents $\boldsymbol{\theta}$ as a stick of length one which we subsequently break into K elements. Assuming $\theta_{k-1} < \theta_k$, for $k \in \{1 \cdots, K\}$, the lower bound for any element in $\boldsymbol{\theta}$ is defined as

$$l_k = \begin{cases} 0 & \text{if } k = 1 \\ \theta_{k-1} & \text{if } 1 < k < K. \end{cases} \tag{8.9}$$

The upper bound is defined as

behavior of the proposal distribution. Since both the numerator and denominator are bounded, samples from the tail region of the distributions cannot dominate the bridge sampling estimate. In this sense the bridge sampler improves on other estimation methods —such as the importance sampling estimator or the generalized harmonic mean estimator— whose variance depend on ratios that are potentially unbounded for poorly chosen proposal distributions (Frühwirth-Schnatter, 2004; Gronau et al., 2017).

$$u_k = \begin{cases} \dfrac{1}{K} & \text{if } k = 1 \\ \dfrac{1 - \sum_{j<k} \theta_j}{K + 1 - k} & \text{if } 1 < k < K, \end{cases} \tag{8.10}$$

where $1 - \sum_{j<k} \theta_j$ represents the length of the remaining stick and $K + 1 - k$ is the number of elements in the remaining stick. Let ϕ denote the density of a normal variable with a mean of zero and a variance of one, Φ its cumulative density function, and Φ^{-1} its inverse cumulative density function. Then, the transformation of $\boldsymbol{\theta}$ is given by:

$$\begin{aligned} \xi_k &= \Phi^{-1}\left(\frac{\theta_k - l_k}{u_k - l_k}\right). \\ &= \begin{cases} \Phi^{-1}\left(\dfrac{\theta_k}{1/K}\right) & \text{if } k = 1 \\ \Phi^{-1}\left(\dfrac{\theta_k - \theta_{k-1}}{\dfrac{1 - \sum_{j<k} \theta_j}{K + 1 - k} - \theta_{k-1}}\right) & \text{if } 1 < k < K - 1. \end{cases} \end{aligned}$$

The inverse transformation is given by:

$$\begin{aligned} \theta_k &= (u_k - l_k)\Phi(\xi_k) + l_k \\ &= \begin{cases} \dfrac{1}{K}\Phi(\xi_k) & \text{if } k = 1 \\ \left(\dfrac{1 - \sum_{j<k} \theta_j}{K + 1 - k} - \theta_{k-1}\right)\Phi(\xi_k) + \theta_{k-1} & \text{if } 1 < k < K. \end{cases} \end{aligned}$$

In the inverse transformation θ_k depends only on the first k elements of $\boldsymbol{\xi}$. Therefore, we know that the Jacobian matrix will be lower triangular, and the determinant of the Jacobian matrix will be the product of the diagonal entries given by:

$$\frac{\partial \theta_k}{\partial \xi_k} = \begin{cases} \dfrac{1}{K}\phi(\xi_k) & \text{if } k = 1 \\ (u_k - l_k)\,\phi(\xi_k) & \text{if } 1 < k < K. \end{cases}$$

Therefore, the Jacobian can be computed using the upper and lower bounds for all samples determined in the transformation step:

$$|J| = \frac{1}{K}\phi(\xi_1) \prod_{k=2}^{K-1} \left((u_k - l_k)\,\phi(\xi_k)\right).$$

Taking this transformation into account the bridge sampling estimator computes $\ell_{1,n}$ and $\ell_{2,m}$ as follows:

$$\ell_{1,n} = \frac{p(\boldsymbol{\theta_n^*} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta_n^*} \in \mathcal{R}_r)}{g(\boldsymbol{\xi_n}^*)},$$
$$\ell_{2,m} = \frac{p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \mid \mathcal{H}_e)\mathbb{I}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \in \mathcal{R}_r)}{g(\tilde{\boldsymbol{\xi}}_{\boldsymbol{m}})},$$

where $\boldsymbol{\xi_n}^* = \Phi^{-1}\left(\frac{\boldsymbol{\theta_n^*} - \mathbf{l}}{\mathbf{u} - \mathbf{l}}\right)$, and $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} = ((\mathbf{u} - \mathbf{l})\Phi(\tilde{\boldsymbol{\xi}}_{\boldsymbol{m}}) + \mathbf{l})\,|J|)$.

Taken together, to apply the proposed bridge sampling routine the following three conditions must be met. First, we need to be able to sample directly from the constrained prior and posterior densities, which can be achieved by using the adapted version of the Gibbs sampling method by Damien and Walker (2001) described above. Second, we need to select a suitable proposal distribution for the bridge sampling algorithm; here we choose a multivariate normal distribution that achieves sufficient overlap with our target distribution by moving the samples from the restricted Dirichlet distribution to the real line. Third, we need to choose a bridge function; here, we have chosen the bridge function proposed in Meng and Wong (1996) which has the favorable property that it minimizes the estimated relative mean-squared error.

Given that bridge sampling only requires draws of the restricted distribution and the proposal distribution, this method is more efficient than the unconditional encompassing method (because fewer samples are typically needed) and the conditional encompassing method (because fewer instances of the Gibbs sampler are needed). In addition, the precision of the bridge sampling estimator depends not on the relative size of the restricted parameter space, but on the overlap between the target and proposal distribution; when the proposal distribution resembles the target distribution more closely, the resulting estimates are more accurate (Meng & Wong, 1996).

8.3 Bayes Factors for Mixed Constraints

In addition to pure equality-constrained and pure inequality-constrained hypotheses, researchers may want to specify hypotheses with some parameters that are exactly equal to each other while others can vary freely and again others are ordered (see e.g., Pericchi Guerra, Liu, & Torres, 2008). However, it is not intuitively clear how to compute Bayes factors when parametric constraints are mixed. Without loss of generality, we first consider a mixed hypothesis $\mathcal{H}_m$ where the first j category parameters are constrained to be exactly equal and where the remaining $K - j$ parameters are increasing:

$$\mathcal{H}_m : (\theta_1 = \theta_2 = \cdots = \theta_j) < \theta_{j+1} < \cdots < \theta_K.$$

As shown in Equation (8.3), the Bayes factor of restricted hypotheses against the encompassing hypothesis can be formulated as

$$\mathrm{BF}_{me} = \frac{p(\boldsymbol{\theta} \in \mathcal{R}_m \mid \mathbf{x}, \mathcal{H}_e)}{p(\boldsymbol{\theta} \in \mathcal{R}_m \mid \mathcal{H}_e)}.$$

The mixed hypothesis stipulates the following set of constraints

$$\mathcal{R}_m : (\theta_1 = \cdots = \theta_j) \cap (\theta_j < \cdots < \theta_K) = \mathcal{R}_0 \cap \mathcal{R}_r.$$

The first set of constraints, which we denote with $\mathcal{R}_0$, are the equality constraints, and the second set of constraints, which we denote with $\mathcal{R}_r$, are the inequality constraints. Using this notation, the Bayes factor can be reformulated as

$$\mathrm{BF}_{me} = \underbrace{\frac{p(\boldsymbol{\theta}_r \in \mathcal{R}_r \mid \boldsymbol{\theta}_0 \in \mathcal{R}_0, \mathbf{x}, \mathcal{H}_e)}{p(\boldsymbol{\theta}_r \in \mathcal{R}_r \mid \boldsymbol{\theta}_0 \in \mathcal{R}_0, \mathcal{H}_e)}}_{\mathrm{BF}_{re}} \times \underbrace{\frac{p(\boldsymbol{\theta}_0 \in \mathcal{R}_0 \mid \mathbf{x}, \mathcal{H}_e)}{p(\boldsymbol{\theta}_0 \in \mathcal{R}_0 \mid \mathcal{H}_e)}}_{\mathrm{BF}_{0e}},$$

that is, a conditional Bayes factor for the inequality constraints given the equality constraints and a Bayes factor for the equality constraints. The latter is similar to the Savage-Dickey ratio that we discussed before, but involves a correction for marginalization.

The probabilities above crucially depend on the marginal probabilities $p(\boldsymbol{\theta}_0 \in \mathcal{R}_0 \mid \mathcal{H}_e)$ and $p(\boldsymbol{\theta}_0 \in \mathcal{R}_0 \mid \mathbf{x}, \mathcal{H}_e)$, which are derived from the prior and posterior Dirichlet distributions, respectively. Since the derivations and results are the same for the prior and posterior probabilities, we derive it here for the prior distribution. The prior probability is of the form

$$p(\boldsymbol{\theta}_0 \in \mathcal{R}_0 \mid \mathcal{H}_e) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \int_{\mathcal{R}_e \setminus \mathcal{R}_0} \theta_j^{\sum_{k=1}^{j} \alpha_k - j} \prod_{k=j+1}^{K-1} \theta_k^{\alpha_k - 1} \left(1 - j\theta_j - \sum_{k=j+1}^{K-1} \theta_k\right)^{\alpha_K - 1} \mathrm{d}\boldsymbol{\theta}_r,$$

and involves a Dirichlet integral, except that the first j probabilities are now collapsed. Here, we have used $\mathcal{R}_e \setminus \mathcal{R}_0$ to denote the unconstrained parameter space for the parameters $\boldsymbol{\theta}_r = (\theta_j, \ldots, \theta_{K-1})^{\mathrm{T}}$. We introduce a change of variable $\lambda_j = j\theta_j$, and $\lambda_k = \theta_k$, for $k = j+1, \ldots, K-1$, with $|J| = {}^1/_j$, such that

$$\begin{aligned} p(\boldsymbol{\theta}_0 \in \mathcal{R}_0 \mid \mathcal{H}_e) &= \frac{1}{j\mathrm{B}(\boldsymbol{\alpha})} \int_{\mathcal{R}_e \setminus \mathcal{R}_0} \left(\frac{\lambda_j}{j}\right)^{\sum_{k=1}^{j} \alpha_k - j} \prod_{k=j+1}^{K-1} \theta_k^{\alpha_k - 1} \left(1 - \lambda_j - \sum_{k=j+1}^{K-1} \theta_k\right)^{\alpha_K - 1} \mathrm{d}\boldsymbol{\lambda}_r \\ &= \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \left(\frac{1}{j}\right)^{\sum_{k=1}^{j} \alpha_k - j + 1} \mathrm{B}\left(\sum_{k=1}^{j} \alpha_k - j + 1, \alpha_{j+1}, \ldots, \alpha_K\right), \end{aligned}$$

which allows us to express the (marginal) Bayes factor for the equality constraints as

$$\mathrm{BF}_{e0} = \frac{\mathrm{B}(\boldsymbol{\alpha})}{\mathrm{B}(\boldsymbol{\alpha} + \mathbf{x})} \left(\frac{1}{j}\right)^{\sum_{k=1}^{j} x_k} \frac{\mathrm{B}\left(\sum_{k=1}^{j} (\alpha_k + x_k) - j + 1, \alpha_{j+1} + x_{j+1}, \ldots, \alpha_K + x_K\right)}{\mathrm{B}\left(\sum_{k=1}^{j} \alpha_k - j + 1, \alpha_{j+1}, \ldots, \alpha_K\right)},$$

where the latter factor introduces a correction for marginalizing which originates from the marginalization of the remaining free parameters, including the collapsed category parameter. If it is the case that no free parameters are involved, that is, $\mathcal{H}_0$ assigns expected category proportions to the entire parameter vector $\boldsymbol{\theta}$ (such as in the multinomial test), then the Bayes factor for the equality constraints corresponds to the Savage-Dickey density ratio.[5] It readily follows that the conditional Bayes factor of inequality constraints given the equality constraints now involves expectations over the conditional Dirichlet distributions

$$p(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_0 \in \mathcal{R}_0,\, \mathcal{H}_e) = \text{Dirichlet}\left(\sum_{k=1}^{j} \alpha_k - j + 1,\, \alpha_{j+1} \ldots,\, \alpha_K\right)$$

and

$$p(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_0 \in \mathcal{R}_0,\, \mathbf{x},\, \mathcal{H}_e) = \text{Dirichlet}\left(\sum_{k=1}^{j} (\alpha_k + x_k) - j + 1,\, \alpha_{j+1} + x_{j+1} \ldots,\, \alpha_K + x_K\right),$$

which can be computed, as before, using bridge sampling. To generalize the above derivations for any set of mixed constraints, we note that the conditional Dirichlet distribution adds the parameters for the collapsed categories and corrects for the change in degrees of freedom by subtracting the degrees of freedom it lost; $j - 1$ degrees of freedom are lost if j categories are collapsed. Thus, for mixed hypotheses of the form

$$\mathcal{H}_m : \theta_1 < \theta_2 = \theta_3 < \theta_4 = \theta_5 = \theta_6,$$

we find the following conditional Dirichlet distribution $p(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_0 \in \mathcal{R}_0,\, \mathcal{H}_e) = \text{Dirichlet}\,(\alpha_1,\, \alpha_2 + \alpha_3 - 1,\, \alpha_4 + \alpha_5 + \alpha_6 - 2)$, which has two sets of collapsed categories, and we lose one degree of freedom for the first, and lose two degrees for the second collapsed category.

The marginal probability has two corrections. First, a uniform probability is stipulated for the collapsed categories, i.e., $1/j$ if j categories are collapsed. Its concentration parameter is equal to the sum of the collapsed categories minus the change in degrees of freedom. Second, a multivariate beta function is introduced that incorporates the corrected concentration parameters. For the mixed hypothesis

$$\mathcal{H}_m : \theta_1 < \theta_2 = \theta_3 < \theta_4 = \theta_5 = \theta_6,$$

we readily find the following marginal probability

$$\frac{\mathrm{B}\,(\alpha_1,\, \alpha_2 + \alpha_3 - 1,\, \alpha_4 + \alpha_5 + \alpha_6 - 2)}{\mathrm{B}(\boldsymbol{\alpha})} \left(\frac{1}{2}\right)^{\alpha_2+\alpha_3-1} \left(\frac{1}{3}\right)^{\alpha_4+\alpha_5+\alpha_6-2},$$

[5] When stipulating exact equality constraints on all parameters, it is assumed that there is no prior uncertainty about the model parameters, and the likelihood of the constrained hypothesis marginalized over the parameter space is simply a multinomial distribution. This expression follows from the fact that the prior distribution under $\mathcal{H}_0$ is

$$p(\boldsymbol{\theta} \mid \mathcal{H}_0) = \frac{p(\boldsymbol{\theta} \mid \mathcal{H}_e)\,\mathbb{I}(\boldsymbol{\theta} = \mathbf{c})}{\int_{\mathcal{R}_e} p(\boldsymbol{\theta} \mid \mathcal{H}_e)\,\mathbb{I}(\boldsymbol{\theta} = \mathbf{c})\,\mathrm{d}\boldsymbol{\theta}} = \frac{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathcal{H}_e)}{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathcal{H}_e)} = 1,$$

for $\boldsymbol{\theta} = \mathbf{c}$ and 0 otherwise.

and marginal Bayes factor,

$$\mathrm{BF}_{e0} = \frac{\mathrm{B}(\boldsymbol{\alpha})}{\mathrm{B}(\boldsymbol{\alpha}')} \left(\frac{1}{2}\right)^{x_2+x_3} \left(\frac{1}{3}\right)^{x_4+x_5+x_6} \frac{\mathrm{B}\left(\alpha_1',\, \alpha_2' + \alpha_3' - 1,\, \alpha_4' + \alpha_5' + \alpha_6' - 2\right)}{\mathrm{B}\left(\alpha_1,\, \alpha_2 + \alpha_3 - 1,\, \alpha_4 + \alpha_5 + \alpha_6 - 2\right)}$$

where we have used $\alpha_k' = \alpha_k + x_k$. Note that this result has also been established for a specific case, albeit for a more general set of hypotheses, in Mulder, Wagenmakers, and Marsman (in press). What the above analysis of the Bayes factor for the mixed hypotheses $\mathcal{H}_m$ shows is that we are, in general, able to factor the hypotheses and associated likelihoods. This factorization is beneficial since it allows us to compute Bayes factors for parametric constraints with the methods described in the main text, even if these constraints are mixed. Intuitively, parameters that vary freely in both hypotheses do not affect the resulting Bayes factor, since the associated part of the marginal likelihood can be split off from both the mixed and encompassing hypotheses.

8.4 Disclosures

8.4.1 Data, and Code

Readers can access the data from the empirical example, our online appendix, and the R code all analyses (including the creation of all figures), in our OSF folder at: `https://osf.io/59tce/`. The `R` package `multibridge` which implements the proposed bridge sampling method can be downloaded from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=multibridge`.

8.4.2 Ethical approval

The study was approved by the local ethics board of the University of Amsterdam.

8.5 Empirical Application: Memory of Negative Life Events

In this section we investigate the precision and efficiency of the estimation methods when applied to a real data set published in Uhlenhuth et al. (1974). Specifically, we conduct a Bayesian reanalysis of Haberman's sample to test whether the reported negative life events decrease over time as a function of forgetting. We test this inequality-constrained hypothesis against the encompassing hypothesis without constraints:

$$\begin{aligned}
&\mathcal{H}_r : \theta_1 > \theta_2 > \cdots > \theta_{18} \\
&\mathcal{H}_e : \theta_1, \theta_2, \cdots, \theta_{18}.
\end{aligned}$$

8.5.1 Method

We obtained the Bayes factor using the bridge sampling approach, the conditional encompassing method, and the unconditional encompassing method. To assess the precision and efficiency we computed Bayes factors in favor of $\mathcal{H}_r$ 100 times for the same data set and for each method and recorded the respective values and the runtime to produce a result. We assigned a uniform prior distribution to our parameters of interest, such that we could compute the prior probability of the constraint, $p(\boldsymbol{\theta} \in \mathcal{R} \mid \mathcal{H}_e)$, analytically. For the bridge sampling method, we drew $20,000$ samples from the constrained posterior distribution. For the conditional encompassing method the marginal probabilities of each constraint holding were estimated using $40,000$ draws from the posterior distribution, resulting in a total of $40,000 \times 18$ draws. For the unconditional encompassing method, we drew 5 million samples from the unconstrained posterior distribution.

8.5.2 Results

The estimated Bayes factors BF_{re} are displayed in Figure 8.2. Bayes factors based on the bridge sampling method and the conditional encompassing method are centered around the same value ($M = 168.88$ and $M = 168.55$, respectively); however, the bridge sampling estimates varied far less ($SD = 1.873$) than the estimates produced by the conditional encompassing method ($SD = 22.23$). To understand the reasons for these differences in variability, we investigated the autocorrelation and the influence of chain length on the Bayes factor estimates, but could not identify a consistent pattern. We suspect that the variability stems from the Monte Carlo error that increases with each sequential evaluation of the individual constraint. If it were possible to estimate the conditional probabilities more efficiently the variability in the estimates might reduce. Such an improved algorithm has been developed by Gu et al. (2019) for continuous variables but is not yet available for categorical data.

Regarding computational efficiency, the bridge sampling method had the lowest runtimes with a mean of $M = 29.11 (SD = 0.39)$ seconds. The conditional encompassing method on the other hand had mean runtimes of $M = 375.84$ ($SD = 5.04$) seconds, which is more than 6 minutes to estimate one Bayes factor, compared to less than half a minute for the bridge sampling method. In sum, the empirical example demonstrates that the bridge sampling routine outperforms both the conditional encompassing method and the unconditional encompassing method. The bridge sampling estimates are considerably more precise than those of the conditional encompassing method, and are obtained more quickly. The unconditional encompassing method fails to estimate any Bayes factor altogether.

8.6 Discussion

In this chapter we describe a precise, scalable, and efficient bridge sampling routine to estimate Bayes factors for inequality constrained hypotheses on multinomial data. Bridge sampling is a promising alternative to current methods that sample

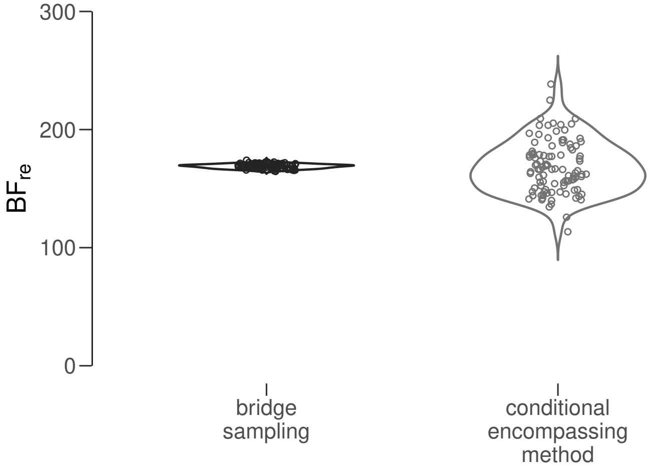


Figure 8.2: Bayes factors for the bridge sampling method (black), the conditional encompassing method (dark grey), and the unconditional encompassing method (light grey) for the test of an order-restriction in Haberman's (1978) data on the reporting of negative life events. Each dot represents one Bayes factor estimate in favor of $\mathcal{H}_r$ obtained by the respective method. The bridge sampling method yields more precise Bayes factor estimates than the conditional encompassing method; the unconditional encompassing method fails to estimate any Bayes factor.

from the unconstrained parameter space and hence may yield imprecise results and long runtimes.

The main reason why the bridge sampling method achieves relatively high precision –even for a model with many categories– is that it does not sample from the unconstrained or increasingly restricted parameter space. Instead, bridge sampling combines the draws from the restricted target distribution with samples from a proposal distribution to estimate the marginal likelihood efficiently. As a result, the precision of the bridge sampling estimate does not depend on the prior probability of the constraint, but rather depends on the similarity between the proposal distribution and the target distribution. Meng and Schilling (2002, p. 584) note that by using more sophisticated methods (e.g., by using warp bridge sampling) to create more overlap between the proposal distribution and the target distribution "[...] we can achieve better and better estimation efficiency based on the same set of draws, and it seems there is no lower bound on the Monte Carlo error". To achieve sufficient overlap between the two distributions, we applied random variable transformation and used the method of moments to construct a suitable proposal distribution.

Compared to existing methods, the bridge sampling routine requires more effort to implement. As with the conditional encompassing method, researchers who wish to use bridge sampling to evaluate inequality constrained hypotheses need to implement a Gibbs sampling algorithm to draw samples from the constrained prior and posterior distribution. In addition, functions must be implemented to perform the required variable transformations and to apply the bridge sampling algorithm. In order to maximize the accessibility of the proposed method, we developed the R package `multibridge` which can be downloaded from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=multibridge`. In the near future we also plan to make the analysis available in the user-friendly statistical software program JASP (JASP Team, 2021). Using JASP does not require any programming experience whatsoever.

At this point we would like to address some discussion points that repeatedly arise in the context of estimating Bayes factors for order restrictions. First, we chose to compare the restricted hypothesis to the encompassing hypothesis in which all parameters are free to vary. This comparison was central in the early work on the encompassing prior approach (e.g., Klugkist et al., 2005). However, this comparison can be critiqued since the encompassing hypothesis overlaps with the restrictive hypothesis (Morey & Rouder, 2011; but see Lee et al., 2019). An alternative comparison pits the restricted hypothesis against its complement (i.e., the part of the encompassing hypothesis that excludes the restricted hypothesis; e.g., Gu et al., 2019; Heck & Davis-Stober, 2019; Hoijtink, 2011; Mulder et al., 2012). Although our implementation is designed to compare against the encompassing hypothesis by default, researchers can easily obtain the comparison to the complement by exploiting the fact that the Bayes factor is transitive (i.e., dividing BF_{re} by $\text{BF}_{\neg re}$).

Secondly, our method involves a Dirichlet prior distribution, which allows researchers to specify values for the concentration parameters. Alternative Bayesian approaches are fully automatic in the sense that the prior distribution is determined by (part of) the sample data (e.g., fractional Bayes factors and adjusted fractional Bayes factors; for details see Böing-Messing & Mulder, 2016; Gu et al., 2019; Mulder, 2014; Mulder et al., 2021; Mulder, Hoijtink, & Klugkist, 2010). These alternative approaches have not yet been applied to multinomial models.

The method proposed here is relatively general and may be extended to problems of higher dimension and increasing sophistication. For instance, the bridge sampling framework could be expanded to multinomial models with complex linear restrictions (e.g., Heck & Davis-Stober, 2019). This would allow researchers to test more complex hypotheses, such as ordinal expectations on the size ratio of the parameters of interest (e.g., $\mathcal{H}_r : \theta_1 > 3 \times \theta_2$), on the differences between category proportions (e.g., $\mathcal{H}_r : (\theta_1 - \theta_2) < (\theta_3 - \theta_4)$), or on odds ratios for data that are summarized in contingency tables (e.g., $\mathcal{H}_r : \frac{\theta_1}{(\theta_1+\theta_2)} < \frac{\theta_3}{(\theta_3+\theta_4)}$). Another generalization of the presented methods concerns the application to hierarchical models, for cases where participants repeatedly choose a response option and therefore category proportions are nested within participants. The bridge sampling routine may also benefit the field of psychometrics, for instance, when evaluating ordinal rating scales (Schnuerch, Haaf, Sarafoglou, & Rouder, 2021), or ordinal item-response theory models (possible applications include Haaf et al., 2020; Karabatsos, 2001; Karabatsos & Sheu, 2004; J. Myung et al., 2005; Tijmstra & Bolsinova, 2019; Tijmstra et al., 2015; see also Meng & Schilling, 1996 for an early use of bridge sampling in the context of IRT modeling). Finally, in the field of cognitive psychology, the bridge sampling routine could be used to test ordinal hypotheses in multinomial processing tree models (e.g., W. Batchelder & Riefer, 1991; Erdfelder et al., 2009; Erdfelder, Hu, Rouder, & Wagenmakers, 2020; Kuhlmann, Erdfelder, & Moshagen, 2019) and in discrete choice models (T. R. Johnson, 2007).

Here we used bridge sampling to evaluate order constraints for the multinomial model, which is associated with categorical data. However, bridge sampling might also benefit the evaluation of order constraints in models associated with continuous data, for instance regression models or analysis of variance (ANOVA) models. The key to using bridge sampling for these applications is to bring the integral over the restricted parameter space into a form that works with standard bridge sampling. For the current application to multinomial data, we used a series of transformations to ensure that (1) the support of one parameter does not depend on other parameters and (2) that the support of the parameters are in a range that matches the one of the multivariate normal proposal distribution. If one can accomplish these goals for any particular problem, then standard bridge sampling immediately applies. As this strategy can be applied to a wide range of problems, we expect bridge sampling to be generalized to other models in the future.

Our results demonstrate that bridge sampling offers considerable improvements in precision and efficiency over existing methods. As our empirical application showed, for multinomial models it is common to have a relatively high number of categories (i.e., $K > 10$) which can easily lead to extreme values of the Bayes factors, if the data either speak for or against the restriction. In other disciplines, such as microbiology, we even find multinomial models with up to $K = 46$ categories, as a study of the relationship between gut microbiome and BMI showed (Song et al., 2020). In these scenarios we believe that the benefit of the bridge sampling routine is particularly apparent. To conclude, the bridge sampling routine of estimating Bayes factors for inequality constraints in multinomial models constitutes a promising tool to evaluate ordinal expectations reliably and efficiently.

8.A Simulation Study: Accuracy of Estimation Methods

To illustrate the accuracy of the estimation methods we conducted two simulation studies. The first simulation study features eight different data sets, given in Table 8.1, for which it is possible to obtain the exact normalizing constants of the restricted prior and posterior distributions (and hence the Bayes factors). These data sets were relatively small with 5 and 6 categories. The normalizing constant of the restricted prior distribution was readily available as we assigned a uniform Dirichlet prior on the model parameters. The exact computations for the restricted posterior distributions exploited the fact that integrating an order restriction expressed in the stick-breaking parameterization amounts to integrating a polynomial whenever the Dirichlet parameters are integers (for details see the online appendix). The exact Bayes factors were then compared to the estimated Bayes factors from the bridge sampling method, the conditional encompassing method, and the unconditional encompassing method.

The second simulation study features five different data sets for which the exact normalizing constants could not be obtained. These data sets featured 18 categories. We compared the variability of the Bayes factor estimates from the bridge sampling method to those from the conditional encompassing method. The comparison did not include the unconditional encompassing method because the prior probability of a sample obeying the restriction is minuscule. As a result, the Bayes factors from this method are liable to be staggeringly overestimated as outlined in the main text.

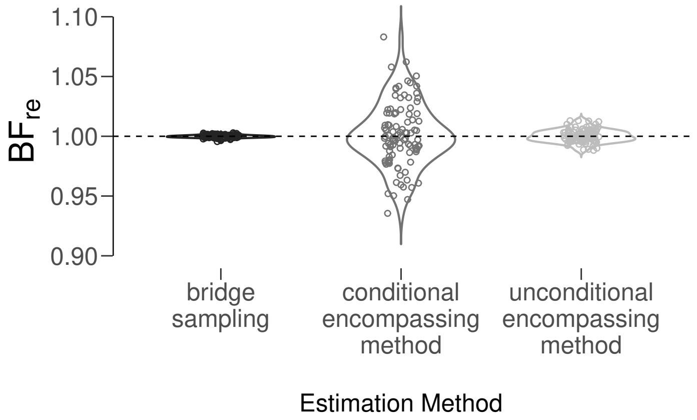
8.A.1 Models with a small number of categories

8.A.1.1 Methods

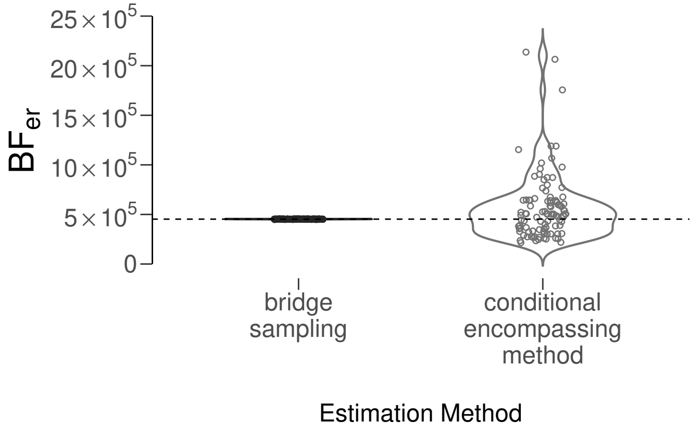
The eight data sets and exact results are summarized in Table 8.1. To quantify accuracy, we estimated the Bayes factors 100 times using the bridge sampling method, the conditional encompassing method, and the unconditional encompassing method. For all data sets, we estimated the Bayes factor in favor of the inequality-constrained hypothesis $\mathcal{H}_r$ that the probabilities of each category are increasing against the encompassing hypothesis $\mathcal{H}_e$ that allows all probabilities to vary freely:

$$\begin{aligned}\mathcal{H}_r &: \theta_1 < \theta_2 < \cdots < \theta_K \\ \mathcal{H}_e &: \theta_1, \theta_2, \cdots, \theta_K.\end{aligned}$$

For the bridge sampling method, we drew $20,000$ samples from the constrained posterior distribution. For the conditional encompassing method the marginal probabilities of each constraint holding were estimated using $40,000$ draws from the posterior distribution, resulting in a total of $200,000$ draws for $\mathbf{x_1}$ and $\mathbf{x_3} - \mathbf{x_7}$, and $240,000$ draws for $\mathbf{x_2}$ and $\mathbf{x_8}$. For the unconditional encompassing method, we drew 5 million samples from the unconstrained posterior distribution.

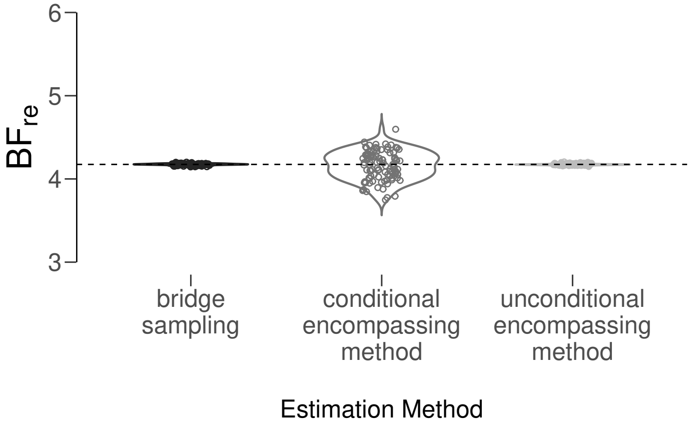


Distribution of Bayes factors for each estimation method for $\mathbf{x_1}$ for which the exact Bayes factor BF_{re} is 1.

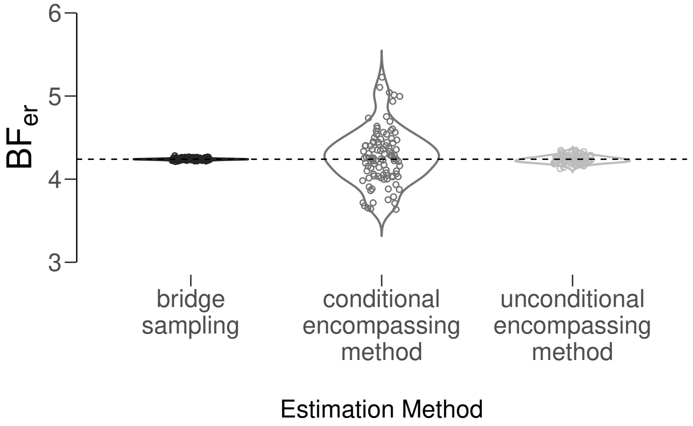


Distribution of Bayes factors for each estimation method for $\mathbf{x_2}$ for which the exact Bayes factor BF_{er} is $452,373$. For these data, the unconditional encompassing method does not succeed in estimating any Bayes factor

Figure 8.3: Violin plots display the estimated Bayes factors for the bridge sampling method (black), the conditional encompassing method (dark grey), and the unconditional encompassing method (light grey) for data sets $\mathbf{x_1}$ and $\mathbf{x_2}$. The dashed horizontal line indicates the exact Bayes factor. Note that the y-axis always shows the Bayes factor in favor of the preferred hypothesis.

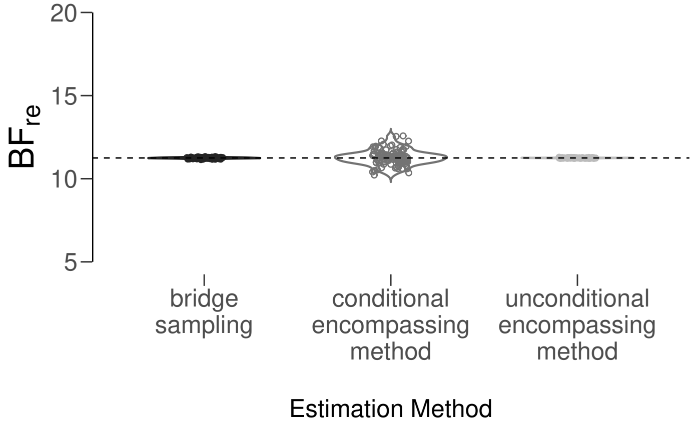


Distribution of Bayes factors for each estimation method for $\mathbf{x_3}$ for which the exact Bayes factor BF_{re} is 4.17.

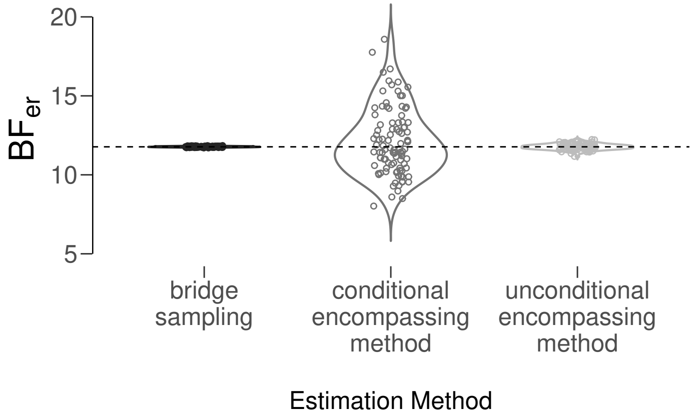


Distribution of Bayes factors for each estimation method for $\mathbf{x_4}$ for which the exact Bayes factor BF_{er} is 4.24.

Figure 8.4: Violin plots display the estimated Bayes factors for the bridge sampling method (black), the conditional encompassing method (dark grey), and the unconditional encompassing method (light grey) for data sets $\mathbf{x_3}$ and $\mathbf{x_4}$. The dashed horizontal line indicates the exact Bayes factor. The Bayes factor estimates of the conditional encompassing method and the encompassing prior method are less variable if the data provides evidence for the restricted hypothesis (top) than if the data provides evidence for the encompassing hypothesis (bottom).

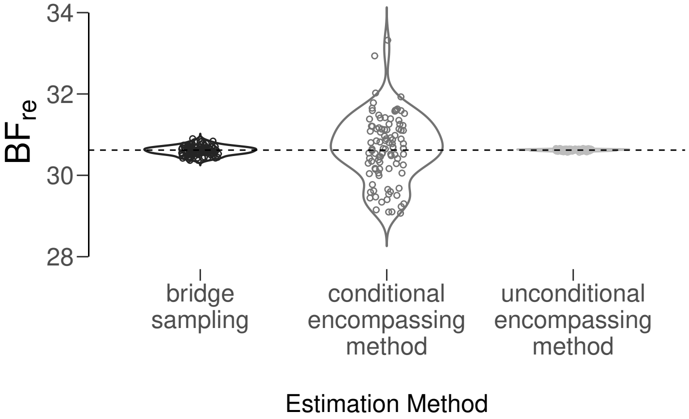


Distribution of Bayes factors for each estimation method for $\mathbf{x_5}$ for which the exact Bayes factor BF_{re} is 11.25.

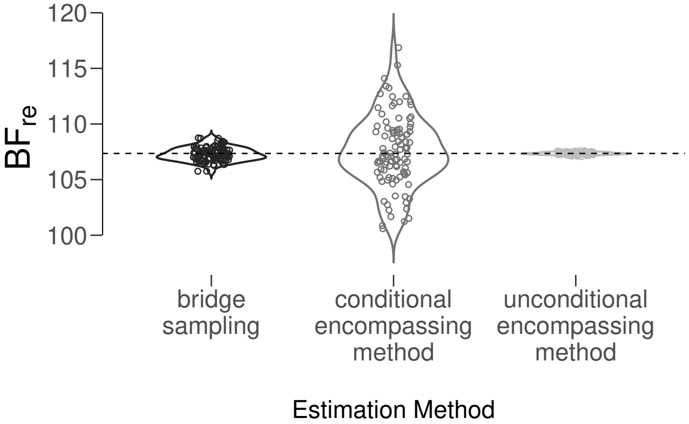


Distribution of Bayes factors for each estimation method for $\mathbf{x_6}$ for which the exact Bayes factor BF_{er} is 11.78.

Figure 8.5: Violin plots display the estimated Bayes factors for the bridge sampling method (black), the conditional encompassing method (dark grey), and the unconditional encompassing method (light grey) for data sets $\mathbf{x_5}$ and $\mathbf{x_6}$. The dashed horizontal line indicates the exact Bayes factor. The Bayes factor estimates of the conditional encompassing method and the encompassing prior method are less variable if the data provides evidence for the restricted hypothesis (top) than if the data provides evidence for the encompassing hypothesis (bottom).



Distribution of Bayes factors for each estimation method for $\mathbf{x_7}$ for which the exact Bayes factor BF$_{re}$ is 30.62.



Distribution of Bayes factors for each estimation method for $\mathbf{x_8}$for which the exact Bayes factor BF$_{re}$ is 107.35.

Figure 8.6: Violin plots display the estimated Bayes factors for the bridge sampling method (black), the conditional encompassing method (dark grey), and the unconditional encompassing method (light grey) for data sets $\mathbf{x_7}$ and $\mathbf{x_8}$. The dashed horizontal line indicates the exact Bayes factor.

Table 8.1: Data Sets, Exact Normalizing Constant of the Restricted Posterior Distribution, and Corresponding Bayes Factors (Rounded to Two Decimals) in Favor of and Against the Inequality-Constrained Hypotheses that the Parameters are Increasing.

Observations	$p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)$	BF$_{er}$	BF$_{re}$
$\mathbf{x_1} = (0,0,0,0,0)'$	0.0083	1	1
$\mathbf{x_2} = (18,15,12,9,6,3)'$	3.07023×10^{-9}	452,373	2.21×10^{-6}
$\mathbf{x_3} = (3,6,8,7,7)'$	0.0347872	0.24	4.17
$\mathbf{x_4} = (3,6,9,6,3)'$	0.00196566	4.24	0.24
$\mathbf{x_5} = (3,6,9,12,10)'$	0.0937483	0.089	11.25
$\mathbf{x_6} = (3,6,9,8,2)'$	0.000707877	11.78	0.085
$\mathbf{x_7} = (3,6,9,12,15)'$	0.255149	0.033	30.62
$\mathbf{x_8} = (3,6,9,12,15,18)'$	0.149099	0.01	107.35

Note. The exact normalizing constant of the restricted prior distribution is 0.008333 for $\mathbf{x_1}$ and $\mathbf{x_3} - \mathbf{x_7}$, and 0.001389 for $\mathbf{x_2}$ and $\mathbf{x_8}$.

8.A.1.2 Results

Figures 8.3 – 8.6 show violin plots that display the Bayes factors for the three estimation methods for the eight data sets. Two results stand out in this simulation: First, in general, all estimation methods approximate the exact Bayes factor, with the conditional encompassing method showing the highest variability. Second, the advantage of bridge sampling becomes most evident for data sets that show evidence for the encompassing hypothesis. Especially for data set $\mathbf{x_2}$, which provides extreme evidence against the inequality-constrained hypothesis, bridge sampling is able to accurately estimate the exact Bayes factor, whereas the conditional encompassing method yields highly variable results and the encompassing prior method fails to estimate any realistic Bayes factor at all: none of the posterior draws were consistent with the restrictive hypothesis, yielding a Bayes factor of 0 for all 100 estimates (bottom panel in Figure 8.3). For $\mathbf{x_6}$ the variability of the Bayes factors in the conditional encompassing method might even lead to different statistical decisions: the range of the Bayes factors is between 8 (which is considered moderate evidence) to 18 (which is considered strong evidence, see bottom panel in Figure 8.5).

8.A.2 Models with a higher number of categories

8.A.2.1 Methods

To further understand how the behavior of the bridge sampling method and the conditional encompassing method differ with increasing model size, we estimated Bayes factors for a model with $K = 18$ categories as applied to five additional

Table 8.2: Data Sets, Bayes Factor Types and Mean Bayes Factors For Models with 18 Categories.

Data Set	Bayes factor	Bridge Sampling	Conditional Encompassing Method
$\mathbf{x_9}$	BF_{re}	$M = 1.00\,[0.98, 1.01]$	$M = 0.99\,[0.77, 1.32]$
$\mathbf{x_{10}}$	BF_{re}	$M = 3.63\,[3.55, 3.71]$	$M = 3.75\,[2.00, 7.40]$
$\mathbf{x_{11}}$	BF_{er}	$M = 3.02\,[2.95, 3.08]$	$M = 3.19\,[1.84, 5.82]$
$\mathbf{x_{12}}$	BF_{re}	$M = 10.96\,[10.72, 11.20]$	$M = 11.02\,[4.94, 17.25]$
$\mathbf{x_{13}}$	BF_{er}	$M = 15.23\,[14.87, 15.55]$	$M = 17.03\,[7.68, 38.33]$.

Note. Square brackets indicate the minimum and maximum Bayes factor estimate.

data sets:

$$
\begin{aligned}
\mathbf{x_9} &: (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)' \\
\mathbf{x_{10}} &: (1,2,3,4,5,6,7,8,9,9,8,7,6,5,4,3,5,6)' \\
\mathbf{x_{11}} &: (1,2,3,4,5,6,7,8,9,9,8,7,6,5,4,4,4,3)' \\
\mathbf{x_{12}} &: (1,2,3,4,5,6,7,8,9,9,8,7,6,5,4,3,6,7)' \\
\mathbf{x_{13}} &: (1,2,3,4,5,6,7,8,9,9,8,7,6,5,3,2,4,4)'
\end{aligned}
$$

As in the first simulation study we drew $20,000$ samples from the constrained posterior distribution for the bridge sampling method. For the conditional encompassing method the marginal probabilities of each constraint holding were estimated using $40,000$ draws from the posterior distribution, resulting in a total of $720,000$ draws.

8.A.2.2 Results

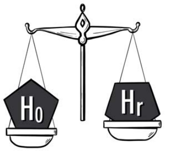
For each of the five data sets, Table 8.2 shows the means and range of the Bayes factors for the two estimation methods. The overall results are consistent with those from the first simulation study. Specifically, the bridge sampling method provides estimates that are less variable than those of the conditional encompassing method. Here again, for the data sets that provide evidence in favor of the encompassing hypothesis (i.e., $\mathbf{x_{11}}$ and $\mathbf{x_{13}}$), the variability of the Bayes factors in the conditional encompassing method might lead to different statistical decisions: the Bayes factors for $\mathbf{x_{11}}$ range between 1.8 (which is considered anecdotal evidence) to 5.8 (which is considered moderate evidence) and the Bayes factors for $\mathbf{x_{13}}$ range between 7.7 (which is considered moderate evidence) to 38 (which is considered strong evidence). For comparison, with bridge sampling the estimated Bayes factor for $\mathbf{x_{13}}$ ranged between 14.9 and 15.5. In general, for both methods the variability in the estimates increases with the strength of the evidence, either in favor or against the inequality-constrained hypothesis.

8.A.3 Conclusion

In two simulation studies we assessed the accuracy of the bridge sampling method, the conditional encompassing method, and the unconditional encompassing method. In the first simulation study, we obtained the exact Bayes factor for eight data sets and then estimated the Bayes factor using the three estimation methods. In the second simulation study, we specified a larger model and estimated Bayes factors for five additional data sets using bridge sampling and the conditional encompassing method. The first simulation shows that the unconditional encompassing method is the most accurate for four of the eight data sets. However, the performance of this method depends heavily on the size of the model: for a model with 18 categories the unconditional encompassing method could not be applied anymore. The method also quickly deteriorates when the data show evidence against the inequality-constrained hypothesis. In contrast, the conditional encompassing method is more responsive to fluctuations in evidence. To improve the accuracy of the method one could increase the number of samples. However, one should take into account that an increase in the number of samples comes at the expense of runtime, which is already many times higher than that of the other two methods.

Based on the recommendation of one of the reviewers we used the bridge sampling method to compute Bayes factors for models with 30 and 50 categories, in order to assess runtime and check whether accuracy decreases with increasing dimensionality. Even in these relatively extreme scenarios the variability of the bridge sampling estimates remained relatively low: in the 30 category scenario the mean Bayes factor was $M = 11.01$ and ranged between 10.55 and 11.45 whereas in the 50 category scenario the mean Bayes factor was $M = 9.51$ and ranged between 8.71 and 10.24. Regarding computational efficiency, the model with 30 categories took an average of about 65 seconds to compute a Bayes factor, whereas the model with 50 categories took an average of about 1 minute and 43 seconds to compute a Bayes factor. The increase in runtime between the 30 and the 50 category scenarios was largely due to the increase in time it takes to sample from the restricted distribution.

Overall, the bridge sampling routine shows a relatively good trade-off between accuracy and efficiency. The variability of the estimates remain in an acceptable range and bridge sampling outperforms the other methods especially when the data provides evidence against the inequality-constraint and when the models feature many categories. The reliability of bridge sampling, which was already on display in the empirical application, was again confirmed in these simulation studies. At this point, we would like to refer the interested reader to our online appendix for a more extended simulation study. This additional simulation study further describes under which conditions the unconditional encompassing method and sometimes even the conditional encompassing method fail to estimate a realistic Bayes factor.



Chapter 9

multibridge: An R Package to Evaluate Informed Hypotheses in Binomial and Multinomial Models

Abstract

The **multibridge** `R` package allows a Bayesian evaluation of informed hypotheses $\mathcal{H}_r$ applied to frequency data from an independent binomial or multinomial distribution. **multibridge** uses bridge sampling to efficiently compute Bayes factors for the following hypotheses concerning the latent category proportions $\boldsymbol{\theta}$: (a) hypotheses that postulate equality constraints (e.g., $\theta_1 = \theta_2 = \theta_3$); (b) hypotheses that postulate inequality constraints (e.g., $\theta_1 < \theta_2 < \theta_3$ or $\theta_1 > \theta_2 > \theta_3$); (c) hypotheses that postulate mixtures of inequality constraints and equality constraints (e.g., $\theta_1 < \theta_2 = \theta_3$); and (d) hypotheses that postulate mixtures of (a)–(c) (e.g., $\theta_1 < (\theta_2 = \theta_3), \theta_4$). Any informed hypothesis $\mathcal{H}_r$ may be compared against the encompassing hypothesis $\mathcal{H}_e$ that all category proportions vary freely, or against the null hypothesis $\mathcal{H}_0$ that all category proportions are equal. **multibridge** facilitates the fast and accurate comparison of large models with many constraints and models for which relatively little posterior mass falls in the restricted parameter space. This chapter describes the underlying methodology and illustrates the use of **multibridge** through fully reproducible examples.

This chapter has been submitted for publication as Sarafoglou, A., Aust, F., Wagenmakers, E. J., & Haaf, J. M. (2021). multibridge: An R Package To Evaluate Informed Hypotheses in Binomial and Multinomial Models. Available as *PsyArXiv preprint*: `https://psyarxiv.com/qk4cy`.

The most common way to analyze categorical variables is to conduct either binomial tests, multinomial tests, or chi-square goodness of fit tests. These tests compare the encompassing hypothesis to a null hypothesis that all underlying category proportions are either exactly equal, or follow a specific distribution. Accordingly, these tests are suitable when theories predict either the invariance of all category proportions or specific values. For instance, chi-square goodness of fit tests are commonly used to test Benford's law, which predicts the distribution of leading digits in empirical datasets (Benford, 1938; Newcomb, 1881). Often, however, the predictions that researchers are interested in are of a different kind. Consider for instance the weak-order mixture model of decision-making (Regenwetter & Davis-Stober, 2012). The theory predicts that individuals' choice preferences are weakly ordered at all times, that is, if they prefer choice A over B and B over C then they will also prefer A over C (Regenwetter et al., 2011)—a well-constrained prediction of behavior. The theory is, however, silent about the exact values of each choice preference. Hence, the standard tests that compare $\mathcal{H}_e$ to $\mathcal{H}_0$ are unsuited to test the derived predictions. Instead, the predictions need to be translated into an informed hypothesis $\mathcal{H}_r$ that reflects the predicted ordinal relations among the parameters. Only then is it possible to adequately test whether the theory of weakly-ordered preference describes participants' choice behavior. Of course, researchers may be interested in more complex hypotheses, including ones that feature combinations of equality constraints, inequality constraints, and unconstrained category proportions. For instance, Nuijten et al. (2016) hypothesized that articles published in social psychology journals would have higher error rates than articles published in other psychology journals. As in the previous example, the authors had no expectations about the exact error rate distribution across journals. Here, again, the standard tests are inadequate. Generally, by specifying informed hypotheses researchers and practitioners are able to "add theoretical expectations to the traditional alternative hypothesis" (Hoijtink et al., 2008, p. 2) and thus test hypotheses that relate more closely to their theories (Haaf et al., 2019; Rijkeboer & van den Hout, 2008).

In the Bayesian framework, researchers may test hypotheses of interest by means of Bayes factors (Jeffreys, 1935; Kass & Raftery, 1995). Bayes factors quantify the extent to which the data change the prior model odds to the posterior model odds, that is, the extent to which one hypothesis outpredicts the other. Specifically, Bayes factors are the ratio of marginal likelihoods of the respective hypotheses. For instance, the Bayes factor for the informed hypothesis versus the encompassing hypothesis is defined as:

$$\text{BF}_{re} = \frac{\overbrace{p(\mathbf{x} \mid \mathcal{H}_r)}^{\substack{\text{Marginal likelihood} \\ \text{under } \mathcal{H}_r}}}{\underbrace{p(\mathbf{x} \mid \mathcal{H}_e)}_{\substack{\text{Marginal likelihood} \\ \text{under } \mathcal{H}_e}}},$$

where the subscript r denotes the informed hypothesis and e denotes the encompassing hypothesis. Several available `R` packages compute Bayes factors for informed hypotheses. For instance, the package **multinomineq** (Heck & Davis-

Stober, 2019) evaluates informed hypotheses for multinomial models as well as models that feature independent binomials. The package **BFpack** (Mulder et al., 2021) evaluates informed hypotheses for statistical models such as univariate and multivariate normal linear models, generalized linear models, special cases of linear mixed models, survival models, and relational event models. The package **BAIN** (Gu et al., 2019) evaluates informed hypotheses for structural equation models. Outside of `R`, the Fortran 90 program **BIEMS** (Mulder et al., 2012) evaluates informed hypotheses for multivariate linear models such as MANOVA, repeated measures, and multivariate regression. All these packages rely on one of two implementations of the encompassing prior approach (Klugkist et al., 2005; Sedransk et al., 1985) to approximate order constrained Bayes factors: the unconditional encompassing method (Hoijtink, 2011; Hoijtink et al., 2008; Klugkist et al., 2005) and the conditional encompassing method (Gu et al., 2014; Laudy, 2006; Mulder, 2014, 2016; Mulder et al., 2009). Even though the encompassing prior approach is currently the most common method to evaluate informed hypotheses, it becomes increasingly unreliable and inefficient as the number of restrictions increases or the parameter space of the restricted model decreases (Sarafoglou, Haaf, et al., 2021).

As alternative to the encompassing prior approach, Sarafoglou, Haaf, et al. (2021) recently proposed a bridge sampling routine (C. H. Bennett, 1976; Meng & Wong, 1996) that computes Bayes factors for informed hypotheses more reliably and efficiently. This routine is implemented in **multibridge** (`https://CRAN.R-project.org/package=multibridge`) and is suitable to evaluate inequality constraints for multinomial and binomial models. When an informed hypothesis includes mixtures of equality and inequality constraints, the core functions in **multibridge** split the hypothesis to compute Bayes factors separately for equality constraints (for which the Bayes factor has an analytic solution) and inequality constraints (for which the Bayes factor is estimated using bridge sampling). The core functions of **multibridge**, that is `mult_bf_informed` and `binom_bf_informed`, return the Bayes factor estimate in favor of or against the informed hypothesis (see Table 9.2 for a summary of the basic required arguments of the two core functions). In addition, users can visualize the posterior parameter estimates under the encompassing hypothesis using the `plot`-method, or get more detailed information on how the Bayes factor is composed using the `summary`-method. For hypotheses that include mixtures between equality and inequality constrained hypotheses the `bayes_factor` method separately returns the Bayes factor for the equality constraints and the conditional Bayes factor for the inequality constraints given the equality constraints. The informed hypothesis can be conveniently specified using a string or character vector. Furthermore, the transitivity property of Bayes factors can be used to test two informed hypotheses against each other (see Example 1 for an illustration). The general workflow of **multibridge** is illustrated in Figure 9.1. A list of all currently available functions and data sets is given in Table 9.1.

This chapter showcases how the proposed bridge sampling routine by Sarafoglou, Haaf, et al. (2021) can be applied in a user-friendly way with **multibridge**. In the remainder of this chapter, we will describe the Bayes factor identity for informed hypotheses in binomial and multinomial models, and briefly describe

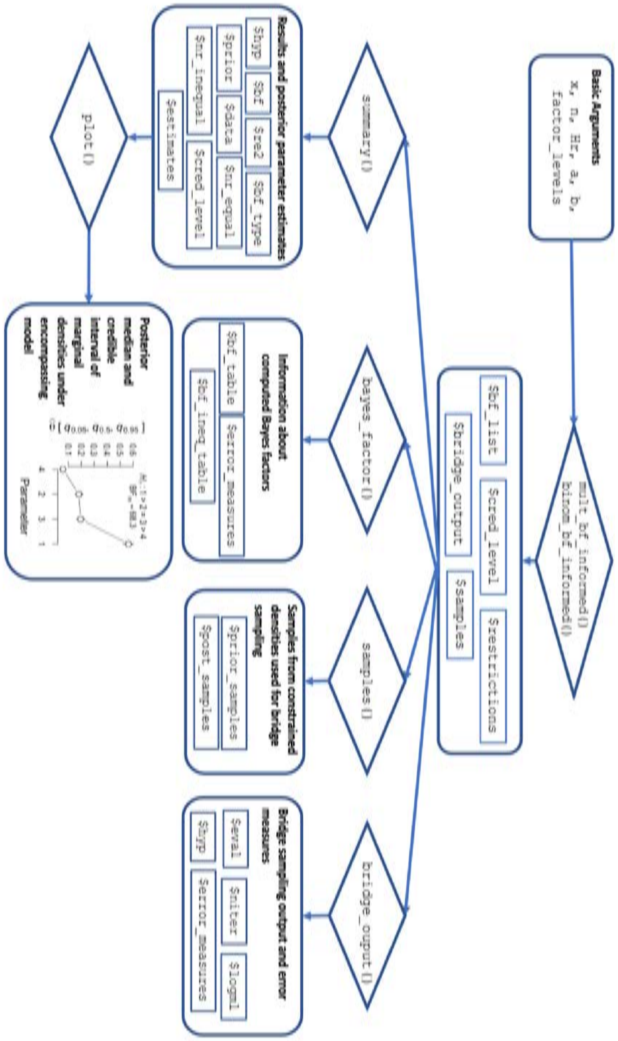


Figure 9.1: The **multibridge** workflow. When calling `mult_bf_informed` or `binom_bf_informed`, the user specifies the data values (`x` and `n` for binomial models and `x` for multinomial models, respectively), the informed hypothesis (`Hr`), the α and β parameters of the binomial prior distributions (`a` and `b`) or the concentration parameters for the Dirichlet prior distribution (`a`), respectively, and the category labels of the factor levels (`factor_levels`). The functions then return the estimated Bayes factor for the informed hypothesis relative to the encompassing or the null hypothesis. Based on these results different S3 methods can be used to get more detailed information on the individual components of the analysis (e.g., `summary`, `bayes_factor`), and parameter estimates of the encompassing distribution (`plot`).

Table 9.1: Core functions available in **multibridge**.

Function Name(s)	Description
`mult_bf_informed`	Evaluates informed hypotheses on multinomial parameters.
`mult_bf_inequality`	Estimates the marginal likelihood of a constrained prior or posterior Dirichlet distribution.
`mult_bf_equality`	Computes Bayes factor for equality constrained multinomial parameters using the standard Bayesian multinomial test.
`mult_tsampling`	Samples from constrained prior or posterior Dirichlet density.
`lifestresses`, `peas`	Data sets associated with informed hypotheses in multinomial models.
`binom_bf_informed`	Evaluates informed hypotheses on binomial parameters.
`binom_bf_inequality`	Estimates the marginal likelihood of constrained prior or posterior beta distributions.
`binom_bf_equality`	Computes Bayes factor for equality constrained binomial parameters.
`binom_tsampling`	Samples from constrained prior or posterior beta densities.
`journals`	Data set associated with informed hypotheses in binomial models.
`generate_restriction_list`	Encodes the informed hypothesis.

the bridge sampling method. Then, we illustrate the core functions of **multibridge** package using two examples and end with a brief summary.

9.1 Methods

In this section we formalize multinomial models and models that feature independent binomial probabilities as they have been implemented in **multibridge**. In the multinomial model, we assume that the vector of observations $\mathbf{x}$ in the K categories follows a multinomial distribution in which the parameters of interest, $\boldsymbol{\theta}$, represent the underlying category proportions. Since the K categories are dependent, the vector of probability parameters is constrained to sum to one, such that $\sum_{k=1}^{K}(\theta_1, \cdots, \theta_K) = 1$. Therefore, a suitable choice for a prior distribution for $\boldsymbol{\theta}$ is the Dirichlet distribution with concentration parameter vector $\boldsymbol{\alpha}$:

$$x_1, \cdots, x_K \sim \text{Multinomial}(\sum_{k=1}^{K} x_k, \theta_1, \cdots, \theta_K) \tag{9.1}$$

$$\theta_1, \cdots, \theta_K \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K), \tag{9.2}$$

where $\boldsymbol{\alpha}$ can be interpreted as vector of *a priori* category counts. The formalization of the model for independent binomial probabilities is similar since the multinomial model above constitutes a generalization of the binomial model (for $K \geq 2$). In the binomial model, we assume that the elements in the vector of successes $\mathbf{x}$ and the elements in the vector of total number of observations $\mathbf{n}$ in the K categories follow independent binomial distributions. As in the multinomial model, the parameter vector of the binomial success probabilities $\boldsymbol{\theta}$ contains the underlying category proportions, however, in this model we assume that categories are independent which removes the sum-to-one constraint. Therefore, a suitable choice for a prior distribution for $\boldsymbol{\theta}$ is a vector of independent beta distributions with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$x_1 \cdots x_K \sim \prod_{k=1}^{K} \text{Binomial}(\theta_k, n_k) \tag{9.3}$$

$$\theta_1 \cdots \theta_K \sim \prod_{k=1}^{K} \text{Beta}(\alpha_k, \beta_k), \tag{9.4}$$

where $\boldsymbol{\alpha}$ can be interpreted as vector of *a priori* successes that observations fall within the various categories and $\boldsymbol{\beta}$ can be interpreted as vector of *a priori* failures.

9.2 Bayes factor

multibridge features two different methods to compute Bayes factors: one method computes Bayes factors for equality constrained parameters and one method computes Bayes factors for inequality constrained parameters. Both methods will be outlined below. In cases where informed hypotheses feature mixtures between inequality and equality constraints, we compute the overall Bayes factor BF_{re} by multiplying the individual Bayes factors for both constraint types. This is motivated by the fact that the Bayes factor for mixtures will factor into a Bayes factor for the equality constraints and a conditional Bayes factor for the inequality constraints given the equality constraints (see Sarafoglou, Haaf, et al., 2021, for the proof).

9.2.1 The Bayes Factor For Equality Constraints

In **multibridge** the Bayes factor for the equality constraints can be computed analytically both for binomial and multinomial models using the functions `binom_bf_equality` and `mult_bf_equality`. For binomial models, assuming that the all binomial probabilities in a model are exactly equal, the Bayes factor is defined as:

$$\text{BF}_{0e} = \frac{\prod_{k=1}^{K} \text{B}(\alpha_k,\, \beta_k)}{\prod_{k=1}^{K} \text{B}(\alpha_k + x_k,\, \beta_k + n_k - x_k)} \times \frac{\text{B}(\alpha_+ + x_+ + 1,\, \beta_+ + n_+ - x_+ + 1)}{\text{B}(\alpha_+ + 1,\, \beta_+ + 1)},$$

where $\text{B}(\cdot)$ denotes the beta function and $\alpha_+ = \sum_{k=1}^{K} \alpha_k$, $\beta_+ = \sum_{k=1}^{K} \beta_k$, $x_+ = \sum_{k=1}^{K} x_k$ and $n_+ = \sum_{k=1}^{K} n_k$. If all binomial probabilities in a model are assumed to be exactly equal *and* equal to a predicted value θ_0, the Bayes factor is defined as:

$$\text{BF}_{0e} = \frac{\prod_{k=1}^{K} \text{B}(\alpha_k,\, \beta_k)}{\prod_{k=1}^{K} \text{B}(\alpha_k + x_k,\, \beta_k + n_k - x_k)} \times \theta_0^{x_+} (1 - \theta_0)^{n_+ - x_+}.$$

Note that **multibridge** only supports the specification of one predicted value for all binomial probabilities. The package does not support the specification of different predicted values for different binomial probabilities. The reason for this is theoretical: we believe that such hypotheses are better tested using a hierarchical structure (thus modeling the binomial probabilities as dependent).

For multinomial models, assuming that all category proportions in a model are equality constrained, the Bayes factor BF_{0e} is defined as:

$$\text{BF}_{0e} = \frac{\text{B}\left(\alpha_1,\, \ldots,\, \alpha_K\right)}{\text{B}\left(\alpha_1 + x_1,\, \ldots,\, \alpha_K + x_K\right)} \times \frac{\text{B}(\boldsymbol{\alpha} + \mathbf{x})}{\text{B}(\boldsymbol{\alpha})} \times \prod_{k=1}^{K} \theta_{0k}^{x_k},$$

where θ_{0k} represent the predicted category proportions. When all category proportions are assumed to be exactly equal all θ_{0k} are set to $\frac{1}{K}$. Otherwise, $\boldsymbol{\theta}_0$ is replaced with the user-specified predicted values.

9.2.2 The Bayes Factor For Inequality Constraints

To approximate the Bayes factor for informed hypotheses, Klugkist et al. (2005) derived an identity that defines the Bayes factor BF_{re} as the ratio of proportions of posterior and prior parameter space consistent with the restriction. This identity forms the basis of the encompassing prior approach. Recently, Sarafoglou, Haaf, et al. (2021) highlighted that these proportions can be reinterpreted as the marginal likelihoods (i.e., the normalizing constants) of the constrained posterior and constrained prior distribution:

$$\text{BF}_{re} = \frac{\overbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x},\, \mathcal{H}_e)}^{\substack{\text{Marginal likelihood of} \\ \text{constrained posterior distribution}}}}{\underbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)}_{\substack{\text{Marginal likelihood of} \\ \text{constrained prior distribution}}}}. \tag{9.5}$$

The benefit of reinterpreting the identity by Klugkist et al. (2005) is that we can estimate the Bayes factor by utilizing numerical sampling methods such as bridge sampling. For that we only need to be able to sample from the constrained densities. Crucially, when using bridge sampling, it does not matter how small the constrained parameter space is in proportion to the encompassing density. This gives the method a decisive advantage over the encompassing prior approach in terms of accuracy and efficiency especially (1) when binomial and multinomial models with moderate to high number of categories (i.e., $K > 10$) are evaluated and (2) when relatively little posterior mass falls in the constrained parameter space.

The bridge sampling algorithm implemented in **multibridge** estimates one marginal likelihood at the time (cf., Gronau et al., 2017; Overstall & Forster, 2010). Specifically, we separately estimate the marginal likelihood for the constrained prior distribution and the marginal likelihood of the constrained posterior distribution. Here we describe how to estimate the marginal likelihood for the constrained prior distribution; the steps presented can then be applied accordingly to the posterior distribution. It should be noted that the bridge sampling algorithm implemented in **multibridge** is an adapted version of the algorithm implemented in the `R` package **bridgesampling** (Gronau et al., 2020) and allows for the specification of informed hypotheses on probability vectors.[1] The bridge sampling identity for the marginal likelihood of the constrained prior distribution is defined as:

$$p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e) = \frac{\mathbb{E}_{g(\boldsymbol{\theta})}\left(p(\boldsymbol{\theta} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)h(\boldsymbol{\theta})\right)}{\mathbb{E}_{\text{prior}}\left(g(\boldsymbol{\theta})h(\boldsymbol{\theta})\right)}, \tag{9.6}$$

where the term $h(\boldsymbol{\theta})$ refers to the bridge function proposed by Meng and Wong (1996), $g(\boldsymbol{\theta})$ refers to a so-called proposal distribution, and $p(\boldsymbol{\theta} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)$ is the part of the prior parameter space under the encompassing hypothesis that is in accordance with the constraint. To estimate the marginal likelihood, bridge sampling requires samples from the target distribution, that is, the constrained Dirichlet distribution for multinomial models and constrained beta distributions for binomial models, and samples from the proposal distribution which in principle can be any distribution with a known marginal likelihood; in **multibridge** the proposal distribution is the multivariate normal distribution. Samples from the target distribution are generated using the Gibbs sampling algorithms proposed by Damien and Walker (2001). For binomial models, we apply the suggested Gibbs sampling algorithm for constrained beta distributions. In the case of the multinomial models, we apply an algorithm that simulates values from constrained Gamma distributions which are then transformed into Dirichlet random variables. To sample efficiently from these distributions, **multibridge** provides a `C++` implementation of this algorithm. Samples from the proposal distribution are generated using the standard `rmvnorm`-function from the `R` package **mvtnorm** (Genz et al., 2020).

The efficiency of the bridge sampling method is optimal only if the target and proposal distribution operate on the same parameter space and have sufficient

[1] In addition, the function to compute the relative mean square error for bridge sampling estimates in **multibridge** is based on the code of the `error_measures`-function from the **bridgesampling** package.

overlap. We therefore probit transform the samples of the constrained distributions to move the samples from the probability space to the entire real line. Subsequently, we use half of these draws to construct the proposal distribution using the method of moments. Details on the probit transformations are provided in Appendix A.

The numerator in Equation 9.6 evaluates the unnormalized density for the constrained prior distribution with samples from the proposal distribution. The denominator evaluates the normalized proposal distribution with samples from the constrained prior distribution. Using this identity, we obtain the bridge sampling estimator for the marginal likelihood of the constrained prior distribution by applying the iterative scheme proposed by Meng and Wong (1996):

$$\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)} \approx \frac{\frac{1}{N_2}\sum_{m=1}^{N_2} \frac{\ell_{2,m}}{s_1\ell_{2,m} + s_2 p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)}}}{\frac{1}{N_1}\sum_{n=1}^{N_1} \frac{1}{s_1\ell_{1,n} + s_2 p(\boldsymbol{\theta}^*_{\boldsymbol{n}} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)}}},$$

where N_1 denotes the number of samples drawn from the constrained distribution, that is, $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta} \mid \mathcal{H}_r)$, N_2 denotes the number of samples drawn from the proposal distribution, that is $\tilde{\boldsymbol{\theta}} \sim g(\boldsymbol{\theta})$, $s_1 = \frac{N_1}{N_2+N_1}$, and $s_2 = \frac{N_2}{N_2+N_1}$. The quantities $\ell_{1,n}$ and $\ell_{2,m}$ are defined as follows:

$$\ell_{1,n} = \frac{q_{1,1}}{q_{1,2}} = \frac{p(\boldsymbol{\theta}^*_{\boldsymbol{n}} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta}^*_{\boldsymbol{n}} \in \mathcal{R}_r)}{g(\boldsymbol{\xi_n}^*)}, \tag{9.7}$$

$$\ell_{2,m} = \frac{q_{2,1}}{q_{2,2}} = \frac{p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \mid \mathcal{H}_e)\mathbb{I}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \in \mathcal{R}_r)}{g(\tilde{\boldsymbol{\xi}}_{\boldsymbol{m}})}, \tag{9.8}$$

where $\boldsymbol{\xi_n}^* = \Phi^{-1}\left(\frac{\boldsymbol{\theta}^*_{\boldsymbol{n}} - \mathbf{l}}{\mathbf{u} - \mathbf{l}}\right)$, and $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} = ((\mathbf{u} - \mathbf{l})\Phi(\tilde{\boldsymbol{\xi}}_{\boldsymbol{m}}) + \mathbf{l})\,|J|)$. The quantity $q_{1,1}$ refers to the evaluations of the constrained distribution for constrained samples and $q_{1,2}$ refers to the proposal distribution evaluated at the probit-transformed samples from the constrained distribution, respectively. The quantity $q_{2,1}$ refers to evaluations of the constrained distribution at the inverse probit-transformed samples from the proposal distribution and $q_{2,2}$ refers to the proposal evaluations for samples from the proposal, respectively. Note that the quantities $\ell_{1,n}$ and $\ell_{2,m}$ have been adjusted to account for the necessary parameter transformations to create overlap between the constrained distributions and the proposal distribution. **multibridge** runs the iterative scheme until the tolerance criterion suggested by Gronau et al. (2017) is reached, that is:

$$\frac{\mid \hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)} - \hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)} \mid}{\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)}} \leq 10^{-10}.$$

The sampling from the target and proposal distribution, the transformations and computational steps are performed automatically within the core functions of

multibridge. The user only needs to provide the functions with the data, a prior and a specification of the informed hypothesis. As part of the standard output of `binom_bf_informed` and `mult_bf_informed`, the functions return the bridge sampling estimate for the log marginal likelihood of the target distribution, its associate relative mean square error, the number of iterations, and the quantities $q_{1,1}$, $q_{1,2}$, $q_{2,1}$, and $q_{2,2}$.

9.3 Usage and Examples

In the following, we will outline two examples on how to use **multibridge** to compare an informed hypothesis to a null or encompassing hypothesis. The first example concerns multinomial data and the second example concerns independent binomial data. Additional examples are available as vignettes (see `vignette(package = "multibridge")`). The two core functions of **multibridge**—`mult_bf_informed` and the `binom_bf_informed`— can be illustrated schematically as follows:

```
mult_bf_informed(x, Hr, a, factor_levels)
binom_bf_informed(x, n, Hr, a, b, factor_levels)
```

To compute a Bayes factor, the core functions require the observed counts, the informed hypothesis, the parameters of the prior distribution under $\mathcal{H}_e$, and the category labels. An overview of these arguments are provided in Table 9.2.

The package also includes S3 methods that, among other things, summarize the results, plot the parameter estimates under $\mathcal{H}_e$, or extract the Bayes factors. Table 9.3 summarizes all S3 methods currently available in **multibridge**.

9.3.1 Disclosures

9.3.1.1 Availability of data and code

The source code of the R package is available at: `https://github.com/ASarafoglou/multibridge/`. In addition, readers can access the code for reproducing all analyses and plots via our project folder on the Open Science Framework: `https://osf.io/2wf5y/`.

9.3.1.2 Ethical Approval

This is a methodological contribution which requires no ethical approval.

9.3.2 Example 1: Applying A Benford Test to Greek Fiscal Data

The first-digit phenomenon, otherwise known as Benford's law (Benford, 1938; Newcomb, 1881) states that the expected proportion of leading digits in empirical data can be formalized as follows: for any given leading digit $d, d = (1, \cdots, 9)$ the expected proportion is approximately equal to

$$\mathbb{E}_{\theta_d} = \log_{10}((d+1)/d).$$

Table 9.2: To estimate the Bayes factor in favor for or against the specified informed hypothesis, the user provides the core functions `mult_bf_informed` and `binom_bf_informed` with the basic required arguments listed below.

Argument	Description
`x`	`numeric`. Vector with data (for multinomial models) or a vector of counts of successes, or a two-dimensional table (or matrix) with 2 columns, giving the counts of successes and failures, respectively (for binomial models).
`n`	`numeric`. Vector with counts of trials. Must be the same length as `x`. Ignored if `x` is a matrix or a table. Included only in `binom_bf_informed`.
`Hr`	`string` or `character`. String or vector with the user specified informed hypothesis. Parameters may be referenced by the specified `factor_levels` or by numerical indices.
`a`	`numeric`. Vector with concentration parameters of Dirichlet distribution (for multinomial models) or α parameters for independent beta distributions (for binomial models). Must be the same length as `x`. Default sets all parameters to 1.
`b`	`numeric`. Vector with β parameters. Must be the same length as `x`. Default sets all β parameters to 1. Included only in `binom_bf_informed`.
`factor_levels`	`character`. Vector with category labels. Must be the same length as `x`.

This means that in an empirical data set, numbers with smaller leading digits are more common than numbers with larger leading digits. Specifically, a number has leading digit 1 in 30.1% of the cases, and leading digit 2 in 17.61% of the cases; leading digit 9 is the least frequent digit with an expected proportion of only 4.58% (see Table 9.4 for an overview of the expected proportions). Empirical data for which this relationship holds include population sizes, death rates, baseball statistics, atomic weights of elements, and physical constants (Benford, 1938). In contrast, artificially generated data, such as telephone numbers, do in general not obey Benford's law (Hill, 1995). Given that Benford's law applies to empirical data but not artificially generated data, a so-called Benford test can be used in fields like accounting and auditing to check for indications for poor data quality (for an overview, see e.g., Durtschi, Hillison, & Pacini, 2004; M. Nigrini, 2012; M. J. Nigrini & Mittermaier, 1997). Data that do not pass the Benford test, should raise audit risk concerns, meaning that it is recommended that they undergo additional follow-up checks (Nigrini, 2019).

Below we discuss four possible Bayesian adaptations of the Benford test. In a first scenario we simply conduct a Bayesian multinomial test in which we test the point-null hypothesis $\mathcal{H}_0$ which predicts a Benford distribution against the encompassing hypothesis $\mathcal{H}_e$. In a second scenario we test the null hypothesis

Table 9.3: S3 methods available in **multibridge**.

Function Name(s)	S3 Method	Description
`mult_bf_informed`, `binom_bf_informed`	`print`	Prints model specifications and descriptives.
	`summary`	Prints and returns the Bayes factor and associated hypotheses for the full model, and all equality and inequality constraints.
	`plot`	Plots the posterior median and credible interval of the parameter estimates of the encompassing model. Default sets credible interval to 95%.
	`bayes_factor`	Contains all Bayes factors and log marginal likelihood estimates for inequality constraints.
	`samples`	Extracts prior and posterior samples from constrained densities (if bridge sampling was applied).
	`bridge_output`	Extracts bridge sampling output and associated error measures.
	`restriction_list`	Extracts restriction list and associated informed hypothesis.
`mult_bf_inequality`, `binom_bf_inequality`	`print`	Prints the bridge sampling estimate for the log marginal likelihood and the corresponding percentage error.
	`summary`	Prints and returns the bridge sampling estimate for the log marginal likelihood and associated error terms.

against an alternative hypothesis, denoted as $\mathcal{H}_{r1}$, which predicts a decreasing trend in the proportions of leading digits. The hypothesis $\mathcal{H}_{r1}$ exerts considerably more constraint than $\mathcal{H}_e$ and provides a more sensitive test if our primary goal is to test whether data comply with Benford's law or whether the data follow a similar but different trend. In the next two scenarios, our main goal is to identify fabricated data. The third scenario therefore tests the null hypothesis against the hypothesis that all proportions occur equally often. This hypothesis $\mathcal{H}_{r2}$ could be considered if it is suspected that the data were generated randomly. In a fourth scenario we test the null hypothesis against a hypothesis which predicts a trend that is characteristic for manipulated data. This hypothesis, which we denote as $\mathcal{H}_{r3}$, could be derived from empirical research on fraud or be based on observed patterns from former fraud cases. For instance, Hill (1995) instructed students to produce a series of random numbers; in the resulting data the proportion of the leading digit 1 occurred most often and the digits 8 and 9 occurred least often which is consistent with the general pattern of Benford's law. However, the proportion for the remaining leading digits were approximately equal. Note that the predicted distribution derived from Hill (1995) is not currently used as a test to detect fraud. However, for the sake of simplicity, if we assume that this pattern could be an indication of manipulated auditing data, the Bayes factor BF_{0r3} would quantify the evidence of whether the proportion of first digits resemble authentic or fabricated data.

9.3.2.1 Data and Hypothesis

The data we use to illustrate the computation of Bayes factors were originally published by the European statistics agency *Eurostat* and served as basis for reviewing the adherence to the Stability and Growth Pact of EU member states. Rauch et al. (2011) conducted a Benford test on data related to budget deficit criteria, that is, public deficit, public dept and gross national products. The data used for this example features the proportion of first digits from Greek fiscal data in the years between 1999 and 2010; a total of $N = 1{,}497$ numerical data were included in the analysis. We choose this data, since the Greek government deficit and debt statistics states has been repeatedly criticized by the European Commission in this time span (European Commision, 2004, 2010). In particular, the commission has accused the Greek statistical authorities to have misreported deficit and debt statistics. For further details on the data set see Rauch et al. (2011). The observed and expected proportions are displayed in Table 9.4; the expected proportions versus the posterior parameter estimates under the encompassing hypothesis are displayed in Figure 9.2.

In this example, the parameter vector of the multinomial model, $\theta_1, \cdots, \theta_K$, reflects the probabilities of a leading digit in the Greek fiscal data being a number from 1 to 9. The hypotheses introduced above can then be formalized as follows. The null hypothesis specifies that the proportions of first digits obeys Benford's law:

$$\mathcal{H}_0 : \boldsymbol{\theta}_0 = (0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051, 0.046).$$

Table 9.4: Observed counts, observed proportions, and expected proportions of first digits in the Greek fiscal data set. The total sample size was N = 1,497 observations. Note that the observed proportions and counts deviate slightly from those reported in Rauch et al. (2011) (probably due to rounding errors).

Leading digit	Observed Counts	Observed Proportions	Expected Proportions: Benford's Law
1	509	0.340	0.301
2	353	0.236	0.176
3	177	0.118	0.125
4	114	0.076	0.097
5	77	0.051	0.079
6	77	0.051	0.067
7	53	0.035	0.058
8	73	0.049	0.051
9	64	0.043	0.046

This null hypothesis can then be tested against each of the following four alternative hypotheses:

$$\begin{aligned}
\mathcal{H}_e &: \boldsymbol{\theta} \sim \text{Dirichlet}(\mathbf{1}),\\
\mathcal{H}_{r1} &: \theta_1 > \theta_2 > \theta_3 > \theta_4 > \theta_5 > \theta_6 > \theta_7 > \theta_8 > \theta_9,\\
\mathcal{H}_{r2} &: \boldsymbol{\theta}_0 = \left(\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}\right),\\
\mathcal{H}_{r3} &: \theta_1 > (\theta_2 = \theta_3 = \theta_4 = \theta_5 = \theta_6 = \theta_7) > (\theta_8,\ \theta_9).
\end{aligned}$$

The comparison of any two informed hypotheses with one another follows from the fact that Bayes factors are transitive. For instance, the Bayes factor comparison between $\mathcal{H}_0$ and $\mathcal{H}_{r1}$ can be obtained by first computing BF_{e0} and BF_{er1}, and then dividing out the common hypothesis $\mathcal{H}_e$:

$$\text{BF}_{0r1} = \frac{\text{BF}_{e0}}{\text{BF}_{er1}}.$$

An overview of the relative plausibility of all $M = 5$ models simultaneously may be obtaining by presenting the posterior model probabilities $p(\mathcal{H}_i \mid x)$ (Berger & Molina, 2005). Denoting the prior model probability for model $\mathcal{H}_i$ by $p(\mathcal{H}_i)$, the posterior model probability for $\mathcal{H}_0$ is given by:

$$p(\mathcal{H}_0 \mid \mathbf{x}) = \frac{\dfrac{p(\mathbf{x} \mid \mathcal{H}_0)}{p(\mathbf{x} \mid \mathcal{H}_e)} \times p(\mathcal{H}_0)}{\displaystyle\sum_{i=1}^{M} \dfrac{p(\mathbf{x} \mid \mathcal{H}_i)}{p(\mathbf{x} \mid \mathcal{H}_e)} \times p(\mathcal{H}_i)}.$$

When all hypotheses are equally likely *a priori*, this simplifies to:

$$p(\mathcal{H}_0 \mid \mathbf{x}) = \frac{\text{BF}_{0e}}{\text{BF}_{0e} + \text{BF}_{r1e} + \text{BF}_{r2e} + \text{BF}_{r3e} + \text{BF}_{ee}}.$$

9.3.2.2 Method

Both BF_{0e} and BF_{r2e} may be readily computed by means of a Bayesian multinomial test which is implemented in the function `mult_bf_equality`. This function requires (1) a vector with observed counts, (2) a vector with concentration parameters of the Dirichlet prior distribution under $\mathcal{H}_e$, and (3) the vector of expected proportions under $\mathcal{H}_0$ and under $\mathcal{H}_{r2}$. We do not incorporate specific expectations about the distribution of leading digits in the Greek fiscal data and therefore set all concentration parameters under $\mathcal{H}_e$ to 1 (i.e., we assign $\boldsymbol{\theta}$ a uniform Dirichlet prior distribution).

```
# Observed counts
x <- c(509, 353, 177, 114,  77,  77,  53,  73,  64)
# Prior specification for Dirichlet prior distribution under H_e
a <-  rep(1, 9)
# Expected proportions for H_0 and H_r2
p0  <- log10((1:9 + 1)/1:9)
pr2 <- rep(1/9, 9)
# Execute the analysis
results_H0_He   <- mult_bf_equality(x = x, a = a, p = p0)
results_Hr2_He  <- mult_bf_equality(x = x, a = a, p = pr2)

logBFe0  <- results_H0_He$bf$LogBFe0
logBFer2 <- results_Hr2_He$bf$LogBFe0
```

The hypotheses $\mathcal{H}_{r1}$ and $\mathcal{H}_{r3}$ contain inequality constraints, and this necessitates the use of the function `mult_bf_informed` to compute the Bayes factors BF_{r1e} and BF_{r3e}. This function requires (1) a vector with observed counts, (2) a vector with concentration parameters of the Dirichlet prior distribution under $\mathcal{H}_e$, (3) labels for the categories of interest (i.e., leading digits), and (4) the informed hypothesis $\mathcal{H}_{r1}$ or $\mathcal{H}_{r3}$ (e.g., as a string). In addition to the basic required arguments, we use two additional arguments here. The first argument sets the Bayes factor type, that is, whether the output should print the Bayes factor in favor of the informed hypothesis (i.e., BF_{re}) or in favor of the encompassing hypothesis (i.e., BF_{er}). It is also possible to compute the log Bayes factor in favor of the hypothesis, which is the setting we choose for this example. The purpose of the second argument `seed` is to make the results reproducible:

```
# Observed counts
x <- c(509, 353, 177, 114,  77,  77,  53,  73,  64)
# Prior specification for Dirichlet prior distribution under H_e
a <-  rep(1, 9)
# Labels for categories of interest
```

Table 9.5: Prior model probabilities, posterior model probabilities, and Bayes factors for five rival accounts of first digit frequencies in the Greek fiscal data set.

Hypothesis	$p(\mathcal{H}_{\cdot})$	$p(\mathcal{H}_{\cdot} \mid \mathbf{x})$	$\log(\text{BF}_{\cdot 0})$
$\mathcal{H}_0$	0.2	1.27×10^{-11}	0
$\mathcal{H}_{r1}$	0.2	0.9994	25.09
$\mathcal{H}_e$	0.2	0.0006	17.67
$\mathcal{H}_{r3}$	0.2	9.46×10^{-79}	-154.57
$\mathcal{H}_{r2}$	0.2	2.71×10^{-212}	-462.06

```
factor_levels <- 1:9
# Specifying the informed hypotheses as a string
Hr1 <- c('1 > 2 > 3 > 4 > 5 > 6 > 7 > 8 > 9')
Hr3 <- c('1 > 2 = 3 = 4 = 5 = 6 = 7 > 8 > 9')
# Execute the analysis
results_He_Hr1 <- mult_bf_informed(x = x, Hr = Hr1, a = a,
                                   factor_levels = factor_levels,
                                   bf_type = 'LogBFer', seed = 2020)
logBFer1 <- summary(results_He_Hr1)$bf
results_He_Hr3 <- mult_bf_informed(x = x, Hr = Hr3, a = a,
                                   factor_levels = factor_levels,
                                   bf_type = 'LogBFer', seed = 2020)
logBFer3 <- summary(results_He_Hr3)$bf
```

We may now exploit transitivity to compare all alternative hypotheses to the Benford null hypothesis $\mathcal{H}_0$. We also compute the posterior model probabilities for all hypotheses. The results are shown in Table 9.5.

The results indicate strong support for $\mathcal{H}_{r1}$ –the model in which the proportions are assumed to decrease monotonically– over all other models. The log Bayes factor of $\mathcal{H}_{r1}$ against Benford's law $\mathcal{H}_0$ is an overwhelming 25.09; the evidence for $\mathcal{H}_{r1}$ is even stronger when it is compared against models that feature equality constraints (i.e., $\mathcal{H}_{r2}$ and $\mathcal{H}_{r3}$). Finally, $\mathcal{H}_{r1}$ also outperforms model $\mathcal{H}_e$, the unconstrained model in which all parameters are free to vary. The latter result demonstrates how a parsimonious model that makes precise predictions can be favored over a model that is more complex (e.g., Jefferys & Berger, 1992). The strong Bayes factor support for $\mathcal{H}_{r1}$ translates to a relatively extreme posterior model probability of 0.9994.

To summarize, the data offer overwhelming support for hypothesis $\mathcal{H}_{r1}$, which postulates a decreasing trend in the digit proportions. This model outperformed both simpler models (e.g., the Benford model) and a more complex model in which the proportions were free to vary. Detailed follow-up analyses are needed to discover why the data follow a monotonically decreasing pattern but not any of the two specific patterns that were put to the test (M. J. Nigrini, 2019).

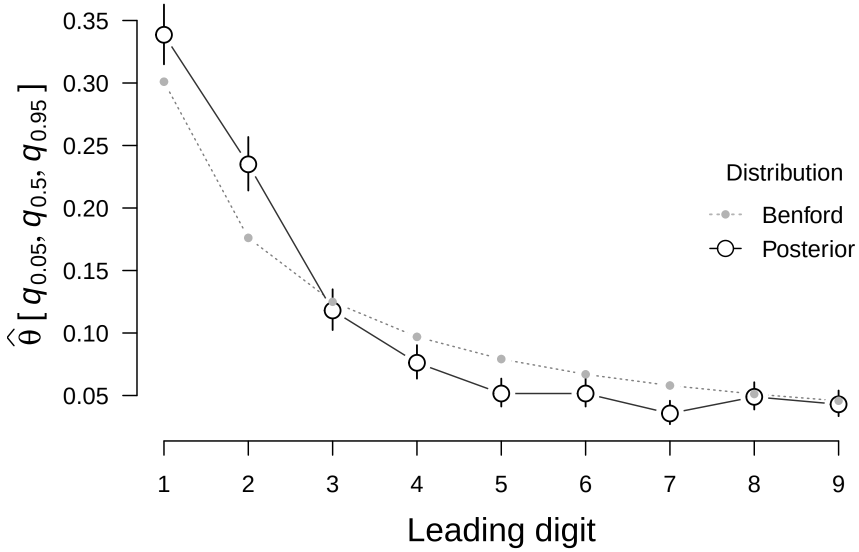


Figure 9.2: Predictions from Benford's law (in grey) show together with the posterior medians (black circles) for the category proportions estimated under the encompassing model $\mathcal{H}_e$. The circle skewers show the 95% credible intervals. Only three of nine intervals encompass the expected proportions, suggesting that the data do not follow Benford's law. This plot was created using the `plot`-S3-method for `summary.bmult` objects in **multibridge**.

9.3.3 Example 2: Prevalence of Statistical Reporting Errors

This section illustrates how **multibridge** may be used to evaluate models for independent binomial data rather than multinomial data. Our example concerns the prevalence of statistical reporting errors across eight different psychology journals. In any article that uses null hypothesis significance testing, there is a chance that the reported test statistic and degrees of freedom do not match the reported p-value, possibly because of copy-paste errors. To flag these errors, Epskamp and Nuijten (2016) developed the R package `statcheck`, which scans the PDF of a given scientific article and automatically detects statistical inconsistencies. This package allowed Nuijten et al. (2016) to estimate the prevalence of statistical reporting errors in the field of psychology. In total, the authors investigated a sample of 30,717 articles (which translates to over a quarter of a million p-values) published in eight major psychology journals between 1985 to 2013: *Developmental Psychology* (DP), the *Frontiers in Psychology* (FP), the *Journal of Applied Psychology* (JAP), the *Journal of Consulting and Clinical Psychology* (JCCP), *Journal of Experimental Psychology: General* (JEPG), the *Journal of Personality and Social Psychology* (JPSP), the *Public Library of Science* (PLoS), *Psychological Science* (PS).

Based on several background assumptions, Nuijten et al. (2016) predicted that the proportion of statistical reporting errors is higher for articles published in the *Journal of Personality and Social Psychology* (JPSP) than for articles published in the seven other journals.

9.3.3.1 Data and Hypothesis

Here we reuse the original data published by Nuijten et al. (2016), which we also distribute with the package **multibridge** under the name `journals`.

```
data(journals)
```

The Nuijten et al. (2016) hypothesis of interest, $\mathcal{H}_r$, states that the prevalence for statistical reporting errors is higher for JPSP than for the other journals.[2] We will consider two specific versions of the Nuijten et al. (2016) $\mathcal{H}_r$ hypothesis. The first hypothesis, $\mathcal{H}_{r1}$, stipulates that JPSP has the highest prevalence of reporting inconsistencies, whereas the other seven journals share a prevalence that is lower. The second hypothesis, $\mathcal{H}_{r2}$, also stipulates that JPSP has the highest prevalence of reporting inconsistencies, but does not commit to any particular structure on the prevalence for the other seven journals.

The **multibridge** package can be used to test $\mathcal{H}_{r1}$ and $\mathcal{H}_{r2}$ against the null hypothesis $\mathcal{H}_0$ that all eight journals have the same prevalence of statistical reporting errors. In addition, we will compare $\mathcal{H}_{r1}$, $\mathcal{H}_{r2}$, and $\mathcal{H}_0$ against the encompassing hypothesis $\mathcal{H}_e$ that makes no commitment whatsoever about the prevalence of reporting inconsistencies across the eight journals. In this example, the parameter vector of the binomial success probabilities, $\boldsymbol{\theta}$, reflects the probabilities that articles contain at least one statistical reporting inconsistency across journals. Thus, the above hypotheses can be formalized as follows:

$$\begin{aligned}
\mathcal{H}_0 &: \theta_{\text{JAP}} = \theta_{\text{PS}} = \theta_{\text{JCCP}} = \theta_{\text{PLOS}} = \theta_{\text{DP}} = \theta_{\text{FP}} = \theta_{\text{JEPG}} = \theta_{\text{JPSP}} \\
\mathcal{H}_{r1} &: (\theta_{\text{JAP}} = \theta_{\text{PS}} = \theta_{\text{JCCP}} = \theta_{\text{PLOS}} = \theta_{\text{DP}} = \theta_{\text{FP}} = \theta_{\text{JEPG}}) < \theta_{\text{JPSP}} \\
\mathcal{H}_{r2} &: (\theta_{\text{JAP}}, \theta_{\text{PS}}, \theta_{\text{JCCP}}, \theta_{\text{PLOS}}, \theta_{\text{DP}}, \theta_{\text{FP}}, \theta_{\text{JEPG}}) < \theta_{\text{JPSP}} \\
\mathcal{H}_e &: \theta_{\text{JAP}} \cdots \theta_{\text{JPSP}} \sim \prod_{k=1}^{K} \text{Beta}(\alpha_k, \beta_k).
\end{aligned}$$

9.3.3.2 Method

To compute the Bayes factor BF_{0r} we need to specify (1) a vector with observed successes (i.e., the number of articles that contain a statistical inconsistency), (2) a vector containing the total number of observations (i.e., the number of articles), (3) a vector with prior parameter α_k for each binomial proportion of the beta prior distribution under $\mathcal{H}_e$, (4) a vector with prior parameter β_k for each binomial proportion of the beta prior distribution under $\mathcal{H}_e$, (5) the category labels (i.e., journal names), and (6) the informed hypothesis $\mathcal{H}_{r1}$ or $\mathcal{H}_{r2}$ (e.g., as a string). We also change the Bayes factor type to `LogBFr0` so that the function

[2] Nuijten et al. (2016) did not report inferential tests because they had sampled the entire population. We do report inferential tests here because we wish to learn about the latent data-generating process.

Table 9.6: Prior model probabilities, posterior model probabilities, and Bayes factors for four hypotheses concerning the prevalence of statistical reporting errors across psychology journals.

Hypothesis	$p(\mathcal{H}_.)$	$p(\mathcal{H}_. \mid \mathbf{x})$	$\log(\text{BF}_{.0})$
$\mathcal{H}_0$	0.25	1.6073×10^{-69}	0
$\mathcal{H}_{r2}$	0.25	0.8814	158.28
$\mathcal{H}_e$	0.25	0.1186	156.27
$\mathcal{H}_{r1}$	0.25	1.9517×10^{-37}	73.88

returns the log Bayes factor in favor for the informed hypothesis compared to the null hypothesis. Since we have no specific expectations about the distribution of statistical reporting errors in any given journal, we set all parameters α_k and β_k to one which corresponds to uniform beta distributions. With this information, we can now conduct the analysis with the function `binom_bf_informed`.

```
# Since percentages are rounded to two decimal values, we round the
# articles with an error to obtain integer values
x <- round(journals$articles_with_NHST  *
             (journals$perc_articles_with_errors/100))
# Total number of articles
n <- journals$articles_with_NHST
# Prior specification for beta prior distributions under H_e
a <- rep(1, 8)
b <- rep(1, 8)
# Labels for categories of interest
journal_names <- journals$journal

# Specifying the informed Hypothesis
Hr1 <- c('JAP = PS = JCCP = PLOS = DP = FP = JEPG < JPSP')
Hr2 <- c('JAP , PS , JCCP , PLOS , DP , FP , JEPG < JPSP')

# Execute the analysis for Hr1
results_H0_Hr1 <- binom_bf_informed(x = x, n = n, Hr = Hr1, a = a, b = b,
                                    factor_levels = journal_names,
                                    bf_type = 'LogBFr0', seed = 2020)
# Execute the analysis for Hr2
results_H0_Hr2 <- binom_bf_informed(x = x, n = n, Hr = Hr2, a = a, b = b,
                                    factor_levels = journal_names,
                                    bf_type = 'LogBFr0', seed = 2020)

LogBFe0  <- results_H0_Hr1$bf_list$bf0_table[['LogBFe0']]
LogBFr10 <- summary(results_H0_Hr1)$bf
LogBFr20 <- summary(results_H0_Hr2)$bf
```

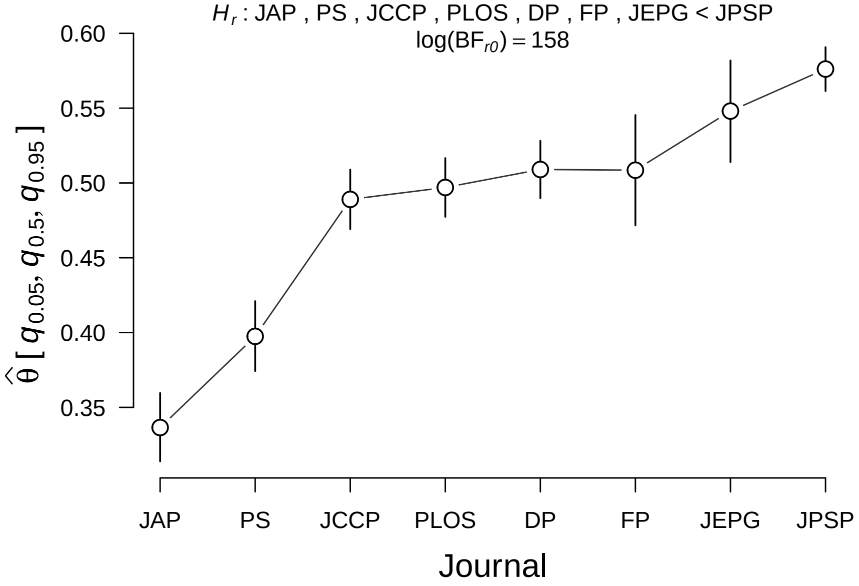


Figure 9.3: Posterior medians for the prevalence of statistical reporting inconsistencies across eight psychology journals, as obtained using the encompassing model. The circle skewers show the 95% credible intervals. Analysis based on data from Nuijten et al. (2016). This plot was created using the `plot`-S3-method for `summary.bmult` objects.

As the evidence is extreme in all four cases, we again report all Bayes factors on the log scale. The Bayes factor log(BF$_{r20}$) indicates overwhelming evidence for the informed hypothesis that JPSP has the highest prevalence for statistical reporting inconsistencies compared to the null hypothesis that the statistical reporting errors are equal across all eight journals; log(BF$_{r20}$) = 158.28. For a clearer picture about the ordering of the journals we can investigate the posterior distributions for the prevalence rates obtained under the encompassing model.

```
LogBFe0  <- results_H0_Hr1$bf_list$bf0_table[['LogBFe0']]
LogBFr10 <- summary(results_H0_Hr1)$bf
LogBFr20 <- summary(results_H0_Hr2)$bf

plot(summary(results_H0_Hr2), xlab = "Journal")
```

The posterior medians and 95% credible intervals are returned by the `summary`-method and are shown in Figure 9.3. The figure strongly suggests that the prevalence of reporting inconsistencies is not equal across all eight journals. This impression may be quantified by comparing the null hypothesis $\mathcal{H}_0$ to the encompassing hypothesis $\mathcal{H}_e$. The corresponding Bayes factor equals log(BF$_{e0}$) = 156.27, which confirms that the data dramatically undercut the null hypothesis that the prevalence of statistical reporting inconsistencies is equal across journals.

The data offer most support for the Nuijten hypothesis $\mathcal{H}_{r2}$, which posits that JPSP has the highest prevalence but does not commit to any restriction on the prevalences for the remaining seven journals. This hypothesis may be compared to the encompassing hypothesis $\mathcal{H}_e$, which yields $\log(\text{BF}_{r2e}) = 2.01$. This means that the observed data are $\exp(2.01) \approx 7.45$ times more likely under $\mathcal{H}_{r2}$ than under $\mathcal{H}_e$; this is moderate evidence for the restriction suggested by Nuijten et al. (2016). Under equal prior probability for the models, this Bayes factor translates to a posterior probability on $\mathcal{H}_e$ of 0.119, an amount that researchers may deem too large to discard in an all-or-none fashion.

To summarize, the data provide moderate evidence for the hypothesis stated by Nuijten et al. (2016) that the prevalence of statistical reporting inconsistencies in JPSP is higher than that in seven other psychology journals.

9.4 Summary

The `R` package **multibridge** facilitates the estimation of Bayes factors for informed hypotheses in both multinomial and independent binomial models. The efficiency gains of **multibridge** are particularly pronounced when the parameter restrictions are highly informative or when the number of categories is large.

multibridge supports the evaluation of informed hypotheses that feature equality constraints, inequality constraints, and free parameters, as well as mixtures between them. Moreover, users can choose to test the informative hypothesis against an encompassing hypothesis that lets all parameters vary freely or against the null hypothesis that states that category proportions are exactly equal. Beyond the core functions currently implemented in **multibridge**, there are several natural extensions we aim to include in future versions of this package. For instance, to compare several models with each other we plan to implement functions that compute the posterior model probabilities. Another extension is to facilitate the specification of hierarchical binomial and multinomial models which would allow users to analyze data where responses are nested within a higher-order structure such as participants, schools, or countries. Hierarchical multinomial models can be found, for instance, in source memory research where people need to select a previously studied item from a list (e.g., Arnold, Heck, Bröder, Meiser, & Boywitt, 2019). In addition, we aim to enable the specification of informed hypotheses that are more complex, including hypotheses on the size ratios of the parameters (e.g., $\theta_1 < 2 \times \theta_2$) of interest or the difference between category proportions such that informed hypotheses can also be specified on odds ratios (e.g., $\frac{\theta_1}{(\theta_1+\theta_2)} < \frac{\theta_3}{(\theta_3+\theta_4)}$).

9.A Transforming an Ordered Probability Vector to the Real Line

The bridge sampling routine in **multibridge** uses the multivariate normal distribution as proposal distribution, which requires moving the target distribution $\boldsymbol{\theta}$ to the real line. Crucially, the transformation needs to retain the ordering of the parameters, that is, it needs to take into account the lower bound l_k and the upper bound u_k of each θ_k. To meet these requirements, **multibridge** uses a probit transformation, as proposed in Sarafoglou, Haaf, et al. (2021), and subsequently transforms the elements in $\boldsymbol{\theta}$, moving from its lowest to its highest value. In the binomial model, we move all elements in $\boldsymbol{\theta}$ to the real line and thus construct a new vector $\boldsymbol{y} \in \mathbb{R}^K$. For multinomial models it follows from the sum-to-one constraint that the vector $\boldsymbol{\theta}$ is completely determined by its first $K-1$ elements, where θ_K is defined as $1 - \sum_{k=1}^{K-1} \theta_k$. Hence, for multinomial models we will only consider the first $K-1$ elements of $\boldsymbol{\theta}$ and we will transform them to $K-1$ elements of a new vector $\boldsymbol{y} \in \mathbb{R}^{K-1}$.

Let ϕ denote the density of a normal variable with a mean of zero and a variance of one, Φ denote its cumulative density function, and Φ^{-1} denote the inverse cumulative density function. Then for each element θ_k, the transformation is

$$\xi_k = \Phi^{-1}\left(\frac{\theta_k - l_k}{u_k - l_k}\right),$$

The inverse transformation is given by

$$\theta_k = (u_k - l_k)\Phi(\xi_k) + l_k.$$

To perform the transformations, we need to determine the lower bound l_k and the upper bound u_k of each θ_k. Assuming $\theta_{k-1} < \theta_k$ for $k \in \{2 \cdots, K\}$ the lower bound for any element in $\boldsymbol{\theta}$ is defined as

$$l_k = \begin{cases} 0 & \text{if } k = 1 \\ \theta_{k-1} & \text{if } 1 < k < K. \end{cases}$$

This definition holds for both binomial models and multinomial models. Differences in these two models appear only when determining the upper bound for each parameter. For binomial models, the upper bound for each θ_k is simply 1. For multinomial models, however, due to the sum-to-one constraint the upper bounds depend on the values of smaller elements as well as on the number of remaining larger elements in $\boldsymbol{\theta}$. To be able to determine the upper bounds, we represent $\boldsymbol{\theta}$ as unit-length stick which we subsequently divide into K elements (Frigyik et al., 2010; Stan Development Team, 2021). By using this so-called stick-breaking method we can define the upper bound for any θ_k as follows:

$$u_k = \begin{cases} \dfrac{1}{K} & \text{if } k = 1 \\ \dfrac{1 - \sum_{i<k} \theta_i}{ERS} & \text{if } 1 < k < K, \end{cases} \tag{9.9}$$

where $1-\sum_{i<k}\theta_i$ represents the length of the remaining stick, that is, the proportion of the unit-length stick that has not yet been accounted for in the transformation. The elements in the remaining stick are denoted as ERS, and are computed as follows:

$$ERS = K - 1 + k.$$

The transformations outlined above are suitable only for ordered probability vectors, that is, for informed hypotheses in binomial and multinomial models that only feature inequality constraints. However, when informed hypotheses also feature equality constrained parameters, as well as parameters that are free to vary we need to modify the formula. Specifically, to determine the lower bounds for any θ_k, we need to take into account how many parameters were set equal to it (denoted as e_k) and how many parameters were set equal to its preceding value θ_{k-1} (denoted as e_{k-1}):

$$l_k = \begin{cases} 0 & \text{if } k = 1 \\ \frac{\theta_{k-1}}{e_{k-1}} \times e_k & \text{if } 1 < k < K. \end{cases} \tag{9.10}$$

The upper bound for parameters in the binomial models still remains 1. To determine the upper bound for multinomial models we must, additionally for each element θ_k, take into account the number of free parameters that share common upper and lower bounds (denoted with f_k). The upper bound is then defined as:

$$u_k = \begin{cases} \frac{1-(f_k \times l_k)}{K} = \frac{1}{K} & \text{if } k = 1 \\ \left(\frac{1-\sum_{i<k}\theta_i - (f_k \times l_k)}{ERS}\right) \times e_k & \text{if } 1 < k < K \text{ and } u_k \geq \max(\theta_{i<k}), \\ \left(2 \times \left(\frac{1-\sum_{i<k}\theta_i - (f_k \times l_k)}{ERS}\right) - \max(\theta_{i<k})\right) \times e_k & \text{if } 1 < k < K \text{ and } u_k < \max(\theta_{i<k}). \end{cases} \tag{9.11}$$

The elements in the remaining stick are then computed as follows

$$ERS = e_k + \sum_{j>k} e_j \times f_j.$$

The rationale behind these modifications will be described in more detail in the following sections. In **multibridge**, information that is relevant for the transformation of the parameter vectors is stored in the generated `restriction_list` which is returned by the main functions `binom_bf_informed` and `mult_bf_informed` but can also be generated separately with the function `generate_restriction_list`. This restriction list features the sublist `inequality_constraints` which encodes the number of equality constraints collapsed in each parameter in `nr_mult_equal`. Similarly the number of free parameters that share common bounds are encoded under `nr_mult_free`.

9.A.1 Equality Constrained Parameters

In cases where informed hypotheses feature a mix of equality and inequality constrained parameters, we compute the Bayes factor BF_{re}, by multiplying the individual Bayes factors for both constraint types with each other:

$$\text{BF}_{re} = \text{BF}_{1e} \times \text{BF}_{2e} \mid \text{BF}_{1e},$$

where the subscript 1 denotes the hypothesis that only features equality constraints and the subscript 2 denotes the hypothesis that only features inequality constraints. To receive $\text{BF}_{2e} \mid \text{BF}_{1e}$, we collapse all equality constrained parameters in the constrained prior and posterior distributions into one category. This collapse has implications on the performed transformations.

When transforming the samples from the collapsed distributions, we need to account for the fact that the inequality constraints imposed under the original parameter values might not hold for the collapsed parameters. Consider, for instance, a multinomial model in which we specify the following informed hypothesis

$$\mathcal{H}_r : \theta_1 < \theta_2 = \theta_3 = \theta_4 < \theta_5 < \theta_6,$$

where samples from the encompassing distribution take the values $(0.05, 0.15, 0.15, 0.15, 0.23, 0.27)$. For these parameter values the inequality constraints hold since 0.05 is smaller than 0.15, 0.23, and 0.27. However, the same constraint does not hold when we collapse the categories θ_2, θ_3, and θ_4 into θ_*. That is, the collapsed parameter $\theta_* = 0.15 + 0.15 + 0.15 = 0.45$ is now larger than 0.23 and 0.27. In general, to determine the lower bound for a given parameter θ_k we thus need to take into account both the number of collapsed categories in the preceding parameter e_{k-1} as well as the number of collapsed categories in the current parameter e_k. Thus, lower bounds for the parameters need to be adjusted as follows:

$$l_k = \begin{cases} 0 & \text{if } k = 1 \\ \frac{\theta_{k-1}}{e_{k-1}} \times e_k & \text{if } 1 < k < K, \end{cases}$$

which leads to Equation 9.10. In this equation, e_{k-1} and e_k refer to the number of equality constrained parameters that are collapsed in θ_{k-1} and θ_k, respectively. In the example above, this means that to determine the lower bound for θ_* we multiply the preceding value θ_1 by three, such that the lower bound is $\left(\frac{0.05}{1}\right) \times 3 = 0.15$. In addition, to determine the lower bound of θ_5 we divide the preceding value θ_* by three, that is, $\left(\frac{0.45}{3}\right) \times 1 = 0.15$. Similarly, to determine the upper bound for a given parameter value θ_k, we need to multiple the upper bound by the number of parameters that are collapsed within it:

$$u_k = \begin{cases} \dfrac{1}{ERS} \times e_k & \text{if } k = 1 \\ \dfrac{1 - \sum_{i<k} \theta_i}{ERS} \times e_k & \text{if } 1 < k < K, \end{cases} \tag{9.12}$$

where $1 - \sum_{i<k} \theta_i$ represents the length of the remaining stick and the number of elements in the remaining stick are computed as follows: $ERS = \sum_k^K e_k$. For the example above, the upper bound for θ_* is $\dfrac{1 - 0.05}{5} \times 3 = 0.57$. The upper bound for θ_5 is then $\dfrac{(1 - 0.05 - 0.45)}{2} \times 1 = 0.25$.

9.A.2 Corrections for Free Parameters

Different adjustments are required for a sequence of inequality constrained parameters that share upper and lower bounds. Consider, for instance, a multinomial model in which we specify the informed hypothesis

$$\mathcal{H}_r : \theta_1 < (\theta_2\, , \theta_3) < \theta_4.$$

This hypothesis specifies that θ_2 and θ_3 have the shared lower bound θ_1 and the shared upper bound θ_4, however, θ_2 can be larger than θ_3 or vice versa. To integrate these cases within the stick-breaking approach one must account for these potential changes of order. For these cases, the lower bounds for the parameters remain unchanged. To determine the upper bound for θ_k, we need to subtract from the length of the remaining stick the lower bound from the parameters that are free to vary. However, only those parameters are included in this calculation that have not yet been transformed:

$$u_k = \begin{cases} \dfrac{1 - (f_k \times l_k)}{K} & \text{if } k = 1 \\ \dfrac{1 - \sum_{i<k} \theta_i - (f_k \times l_k)}{ERS} & \text{if } 1 < k < K, \end{cases} \tag{9.13}$$

where f_k represents the number of free parameters that share common bounds with θ_k and that have been not yet been transformed. Here, the number of elements in the remaining stick is defined as the number of all parameters that are larger than θ_k: $ERS = 1 + \sum_{j>k} f_j$. To illustrate this correction, assume that samples from the encompassing distribution take the values $(0.15, 0.29, 0.2, 0.36)$. The upper bound for θ_1 is simply $\frac{1}{4}$. For θ_2, we need to take into account that θ_2 and θ_3 share common bounds. To compute the upper bound for θ_2, we subtract from the length of the remaining stick the lower bound of θ_3: $\dfrac{1 - 0.15 - (1 \times 0.15)}{1+1} = 0.35$.

A further correction is required if a preceding free parameter (i.e., a parameter with common bounds that was transformed already) is larger than the upper bound of the current parameter. For instance, in our example the upper bound for θ_3 would be $\dfrac{1 - 0.44 - 0}{1+1} = 0.28$, which is smaller than the value of the preceding free parameter, which was 0.29. If in this case θ_3 would actually take on the value close to its upper bound, for instance $\theta_3 = 0.275$, then—due to the sum-to-one constraint—θ_4 would violate the constraint (i.e., $0.15 < (0.29\, , 0.275) \not< 0.285$). In these cases, the upper bound for the current θ_k needs to be corrected downwards. To do this, we subtract from the current upper bound the difference to the largest preceding free parameter. Thus, if $u_k < \max(\theta_{i<k})$, the upper bound becomes:

$$\begin{aligned} u_k &= u_k - (\max(\theta_{i<k}) - u_k) \qquad (9.14)\\ &= 2 \times u_k - \max(\theta_{i<k}). \qquad (9.15) \end{aligned}$$

For our example the corrected upper bound for θ_3 would become $2 \times 0.28 - 0.29 = 0.27$ which secures the proper ordering for the remainder of the parameters. If in this case θ_3 would take on the value close to its upper bound, for instance $\theta_3 = 0.265$, θ_4—due to the sum-to-one constraint—would take on the value 0.295 which would be in accordance with the constraint (i.e., $0.15 < (0.29, 0.265) < 0.295$).

Chapter 10

Theory-Informed Refinement of Bayesian MPT Modeling

Abstract

Multinomial processing tree (MPT) models are a broad class of statistical models used to test sophisticated psychological theories. The research questions derived from these theories often go beyond simple condition effects on parameters and involve ordinal expectations (e.g., bias is higher in one condition than another) or disordinal expectations (e.g., the effect reverses in one experimental condition). Here we argue that by refining common modeling practices, Bayesian hierarchical models are well suited to estimate and test these expectations. Concretely, we show that the default priors proposed in the literature lead to nonsensical predictions for individuals and the population distribution, leading to problems not only in model comparison but also in parameter estimation. Rather than relying on these priors, we argue that MPT modelers should determine priors that are consistent with their theoretical knowledge. In addition, we demonstrate how Bayesian model comparison may be used to test ordinal and disordinal interactions by means of Bayes factors. We apply the techniques discussed to two case studies with empirical data from Bell, Mieth, and Buchner (2015) and Symeonidou and Kuhlmann (2021).

This chapter is under preparation as Sarafoglou, A., Kuhlmann, B. G. Aust, F., & Haaf, J. M. Theory-Informed Refinement of Bayesian MPT Modeling.

10.1 Introduction

Multinomial processing tree (MPT) models are a broad class of statistical models to estimate probabilities of latent cognitive processes underlying observed behaviour (W. Batchelder & Riefer, 1999; Riefer & Batchelder, 1988). In psychology, MPT models are used to test sophisticated theories of memory, judgement and decision making, and reasoning (for a review on the literature, see Erdfelder et al., 2009). These sophisticated theories make predictions for data from experimental tasks and in many cases these predictions are very specific. For instance, based on aging theories, researchers may predict that memory retrieval is more affected by an experimental manipulation for older adults than for younger adults. This prediction is a specific ordinal interaction, however, specifying a statistical model for this prediction is not entirely trivial. We argue that while hierarchical Bayesian MPT modeling is well suited to test these nontrivial predictions, current Bayesian MPT modeling practices leave room for refinement. In particular, we will argue that Bayesian model comparison may be used to easily test specific ordinal interactions, and that the most commonly used priors (Klauer, 2010; Matzke, Dolan, Batchelder, & Wagenmakers, 2015) are simultaneously too vague and too informative for most applications. We will elaborate these arguments using a MPT model that instantiates a psychological theory of source memory, namely the 2-High-Threshold Source Monitoring (2HTSM) model.

Source memory captures a person's ability to remember contextual details that accompanied a piece of learned information (M. K. Johnson, Hashtroudi, & Lindsay, 1993). A typical paradigm to study source memory consists of a study phase and a test phase (e.g., W. Batchelder & Riefer, 1990; M. K. Johnson et al., 1993). In the study phase, participants are presented with items stemming from one of several sources (e.g., words spoken by a female or male voice). In the test phase, participants are presented with the learned items again along with new items. Participants must then decide for each item whether it is a new item or has been presented before, and if so, by what source.

While the source memory paradigm is fairly simple, multiple psychological processes are most likely at play. Suppose someone correctly identified an item as old, and also correctly identified the source of the item. Then this correct identification could be due to actual mnemonic information about the item and the source, or it could be due to guessing. Note that two guessing processes—about the item and the source—might be at play here. The interlocking of guessing and memory processes on several levels are the reason why MPT modeling is so popular in the source memory literature (Erdfelder et al., 2009).

Traditionally, MPT models are specified in the classical frequentist framework. However, Bayesian modeling has increasingly become the tool of choice in the MPT literature, since it facilitates the specification of complex models. For instance, it is reasonable to assume that cognitive processes such as guessing vary across individuals, an assumption that can be easily implemented in the Bayesian framework by extending MPT models as hierarchical models (Rouder & Lu, 2005; Rouder, Morey, & Pratte, 2017).

Within the Bayesian framework the specification of an MPT model requires three steps. The first step is routine for most MPT modelers: specifying the MPT model equations within one experimental condition. These equations formalize assumptions about the cascade of distinct cognitive processes that contribute to the behavior of interest. Each process is associated with a model parameter that controls the probability with which the process is engaged. These equations are often communicated as tree-like diagrams as in Figure 10.1. The second step is to specify the statistical model on parameters within one experimental condition. This step includes the formulation of general model assumptions, such as how variability of items or participants are modeled, but also the specification of adequate prior distributions on the MPT parameters. The third step concerns the expected effects on the model parameters across experimental conditions. In most cases, this step corresponds to the specification of the main research question as competing statistical models. Here, we will focus on steps two and three.

10.1.1 Specifying Theory-Informed Prior Distributions

Just as the model equation, the prior distributions implemented in step two are a crucial part of the theory. Since the model parameters correspond to psychological variables, their distributions identify values that are permissible, likely, unlikely, or non-permissible according to the theory (Lee & Vanpaemel, 2018). Parameter priors therefore formalize theoretical knowledge and need to be determined from the theory itself, based on prior literature, expert knowledge, or informed guesses (Lee, 2018; Lee & Vanpaemel, 2018; Stefan, Evans, & Wagenmakers, 2020; Stefan et al., 2019; Vanpaemel, 2010). However, researchers are often reluctant to utilize their own expertise in determining prior distributions for fear that they may spoil their model evaluation (Kass & Raftery, 1995), or other researchers could criticize the choice. However, instead of viewing prior specification as a burden or a necessary evil, many Bayesians have been advocating a change in perspective: when researchers work with quantitatively instantiated theories—which MPT models undoubtedly are—prior distributions along with model equations, are an opportunity to fully describe all aspects captured by theory (e.g., Dienes, 2011; Rouder, Morey, & Wagenmakers, 2016; Vanpaemel, 2010; Vanpaemel & Lee, 2012).

In this chapter, we show why the prior distributions proposed by Matzke, Dolan, et al. (2015) and based on Klauer (2010) may be problematic, despite their wide used in the MPT literature. These Matzke-Klauer priors were intentionally designed not to embody any psychological theory, resulting in diffuse distributions with a broad range on the latent space. Using the 2HTSM model (Bayen, Murnane, & Erdfelder, 1996) we will demonstrate that predictions made from mathematical models with diffuse priors on the latent space—even if they feature sophisticated model equations—are at odds with basic intuitions about possible data patterns. These priors are therefore inappropriate for testing theory, and may under some conditions even be problematic for parameter estimation.

10.1.2 Model Comparison for Bayesian Hierarchical MPT Models

Theories that are tested with MPT models are oftentimes quite sophisticated, thus requiring complex experimental designs. In turn, MPT models need to be specified to account for a rich set of predictions of experimental effects on cognitive parameters. These predictions often concern ordinal expectations (e.g., bias is higher in one condition than in another) or disordinal expectations (e.g., the effect reverses in one experimental condition) of multiple interaction effects.

In step three of model specification, these predictions are implemented. Traditionally, competing models are implemented in the frequentist framework, and an encompassing model is tested against a model with constraints on parameters across experimental conditions. These constraints can be equality constraints or ordinal constraints, and they are implemented by changing the likelihood of the model, for example using reparameterization (Klauer, Singmann, & Kellen, 2015; Knapp & Batchelder, 2004). More recently, Bayesian model comparison using Bayes factors has also gained traction in MPT modeling, mainly due to computational progress (Gronau, Wagenmakers, Heck, & Matzke, 2019). However, Bayes factor model comparison is not yet common practice, in part because it is challenging to specify MPT models that correspond to specific hypotheses and to evaluate to what extend they are supported by the data. The success of this endeavor depends entirely on how well the researcher succeeds in building their mathematical model. Here, we provide a simple solution to incorporate a set of equality and ordinal constraints on parameters across experimental conditions, and to testing these constraints using Bayes factors.

The structure of the chapter is as follows. First, we introduce the Bayesian implementation of the hierarchical 2HTSM model. Second, we describe how predictions from the 2HTSM model aid model specification, including the selection of appropriate prior distributions. Third, we show how to compare MPT models by means of the Bayes factor. The models under consideration include predictions about the order of processes across experimental predictions, that is, ordinal and disordinal interactions. We show that instead of reparameterizing the models to implement the constraints (i.e., modify their likelihood) the constraints can be implemented directly in the prior distributions. Finally, we illustrate our methods with two case studies using empirical data from Symeonidou and Kuhlmann (2021) and Bell et al. (2015).

10.2 The Two-High Threshold Model for Source Monitoring

We start with the first step MPT modelers go through to express their theory by a mathematical model. This step concerns the formulation of the model equation and the formulation of general model assumptions. The 2HTSM model proposed by Bayen et al. (1996) assumes four independent cognitive processes to contribute to a response in a source memory paradigm. According to the model, participants need to cross two thresholds to be able to fully remember an item and its source. To cross the first threshold, participants have to remember an item as old which is represented by the parameter D. The second threshold depends on the participants' source memory, that is, the probability to remember the source of an item. This probability is represented by parameter d. If either of these thresholds is not crossed, guessing processes will partially or fully determine the response

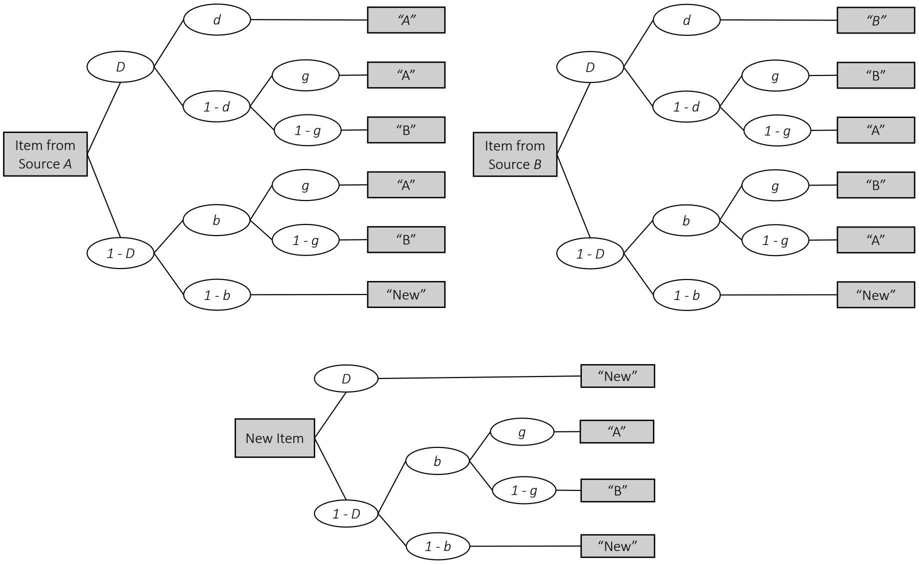


Figure 10.1: Tree architecture for a paradigm of the 2HTSM model. In a source memory task, participants are presented with items that they have previously learned and that either stem from source A, source B (top two trees), or are new items (bottom tree). They then have to distinguish previously learned items from new items and must decide for the previously learned items from which source they originate. The 2HTSM model assumes that participants responses' depend on four cognitive processes: item memory D, source memory d, item guessing b, and source guessing g.

behaviour. Parameter b describes the probability of correctly guessing whether an item has already been learned (item guessing), and parameter g describes the probability of correctly guessing the source of an item (source guessing). In Figure 10.1 we illustrate the tree architecture of the 2HTSM model. Note, however, that the architecture is typically adapted to the specific experimental paradigms used in a study (e.g., using the graphical model builder in Moshagen, 2010).

10.2.1 Specification Of The Statistical Model

The second step of model specification involves the specification of the statistical model within one experimental condition. This specification includes the treatment of participants and items. Arguably, when experimental materials are standardized and validated in pilot studies, item heterogeneity can be well controlled, justifying aggregation across items. The assumption of homogeneity of individuals on the other hand is more problematic (e.g., Rouder & Lu, 2005; Rouder, Lu, Morey, Sun, & Speckman, 2008; Webb & Lee, 2004, but see Matzke, Dolan, et al., 2015 and J. B. Smith & Batchelder, 2008). Since MPT parameters reflect psychological processes (e.g., memory performance) which depend on individual participant characteristics (e.g., age, response biases, stereotypes), it is useful to allow for individual differences in the model. Within the Bayesian framework, two model classes have been established to account for individual differences, the beta-MPT model (J. B. Smith & Batchelder, 2010), and the latent-trait model (Klauer, 2010).

The beta-MPT model assumes that individual level MPT parameters stem from independent group-level beta distributions. As MPT parameters are modeled in the probability space on the individual level and the group level, prior selection is intuitive. In contrast, the latent-trait model, transforms the parameter space to a latent continuous space. The benefit of a latent continuous space is that intuitions from generalized linear models are appropriate here, simplifying the development of regression models on specific parameters. In addition, the latent-trait model allows for the specification of a covariance matrix that models the correlation across participants and allows for more hierarchical shrinkage. Since hierarchical shrinkage is necessary to avoid overestimating individual differences, we will focus on the latent-trait model in the remainder of the chapter. The full mathematical specification of the latent-trait models discussed in this chapter are provided in Appendix A.

In the latent-trait approach, assuming item homogeneity, participant responses are aggregated over items in each experimental condition. The category frequencies are assumed to follow a multinomial distribution with the underlying category probabilities resulting from the MPT model equation. In this model, all parameters are probit-transformed to a latent space. Then, individual differences in MPT parameters are modelled in this unbounded (latent) parameter space. Specifically, it is assumed that the transformed parameters are normally distributed and may be correlated with other parameters (i.e., transformed parameters follow a multivariate normal distribution). The means of the multivariate distribution represent group-level parameters and the variance-covariance matrix determines the magnitude of individuals' deviations around said group-level parameters and their correlation.

10.2.2 The Problem with Default MPT Priors

After establishing the model equation and the statistical model, we now turn to determining adequate prior distributions for the model priors. Priors are needed on the group-level parameters as well as the variance-covariance matrix.

When translating source memory theory into MPT models, the priors we place on the multivariate normal distribution (i.e. means, variances, and covariances) deserve careful consideration. These parameters determine which MPT parameter values are deemed plausible both at the group level and the individual level. Thus, carefully chosen prior distributions should (1) be theoretically justified, (2) faithfully reflect expectations about group- and individual-level parameters in their original probability scale, and by extension (3) imply sensible predictions of group- and individual-level response rates.

However, these requirements are not met for seemingly vague and uninformative prior distributions such as those proposed by Klauer (2010) and Matzke, Dolan, et al. (2015) and implemented in the `R` package `TreeBUGS` (Heck, Arnold,

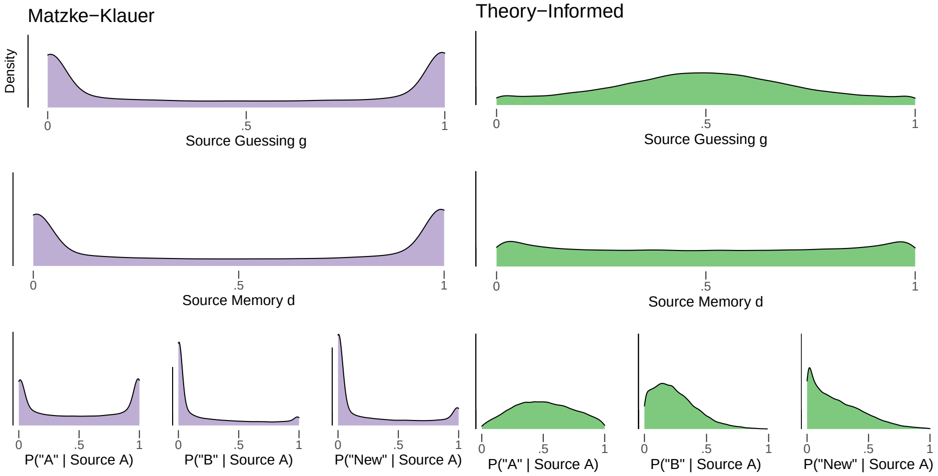


Figure 10.2: The 2HTSM model implemented with Matzke-Klauer priors on the group-level leads to nonsensical and extreme predictions on the individual level (left panel; purple). The right panel shows (right panel; green) predictions of the 2HTSM model with theory-informed prior distributions. The top rows show for one participant prior predictions for the source guessing and source memory parameter. The bottom row depicts for one participant the prior predictions of the probabilities of responding that an item from source A stemmed from source A (left), source B (middle), or is a new item (right).

& Arnold, 2018). On the contrary: they imply highly informative and nonsensical predictions particularly about individual-level parameter values and response rates. Note that the specifications in Klauer (2010), Matzke, Dolan, et al. (2015), and Heck et al. (2018) differ with respect to their vagueness. While specification in Klauer (2010) was the most vague in that it allowed for the most participant variability, Matzke, Dolan, et al. (2015) proposed priors that were slightly more constrained. Heck et al. (2018) implemented in `TreeBUGS` the model proposed by Matzke, Dolan, et al. (2015), but constrained participant variability even further. In the following, when discussing Matzke-Klauer priors, we refer to the default implementation in `TreeBUGS`.

Let us consider the Matzke-Klauer prior on the group-level means of the MPT parameters—a standard normal distribution. This prior is popular since it translates to a uniform distribution on the probability space implying that all values are equally likely a priori. However, since specific prior distributions can be derived from many psychological theories, the standard normal distribution is not ideal for many cases. For instance, if sources appeared equally often and were randomly assigned to items, source guessing g is most likely to be at or near chance level (i.e., .5) rather than strongly biased towards one source. The same applies to source memory d which is recollection based and difficult, thus not likely to be near 1.

A less obvious problem is a vague prior on the variance-covariance matrix. It seems natural to assume that vague priors on the group-level means in combination with high participant variability will lead to vague priors at the individual level. Yet, when moving from the latent space to the probability space, these vague group-level priors combine to highly informative priors for individual participants.

The left panel in Figure 10.2 illustrates the individual-level predictions of participants' source memory parameter and source guessing parameter as well as the predicted category probabilities to answer "A", "B", or "New" given that the correct source was Source A. These predictions indicate that the Matzke-Klauer priors place an outsized amount of prior probability mass on implausible extreme values, a perhaps unexpected result for many users of these models (Lee, 2018 illustrated a similar case in the field of psychophysics). The prior distribution on individual-level g parameters posits that a participant is most likely to either never guess correctly or to always guess incorrectly. Similarly, the prior distribution on individual-level d parameters posits that a participant is most likely to either have perfect source memory or no source memory at all. Based on these priors, any plausible group-level parameter values may correspond to (symmetric or asymmetric) bimodal distributions of individual participant parameters. These priors clearly do not express theoretically sensible expectations about individual participants' behavior.

Moreover, these priors are at odds with typical assumptions about the population distribution. A key motivation for using hierarchical models, such as the latent-trait model, is that they assume that participants belong to a relatively homogeneous population and that therefore the estimates of any one participant partially inform estimates of all other participants from the same population. In the original probability scale, the prior predictions of the Matzke-Klauer priors, however, implement an assumption that is antithetical to the assumption of a common population: a mixture of several different populations. The composition of this mixture is illustrated in Figure 10.3. The figure shows the multivariate prior distribution for the individual-level g and d parameters of the 2HTSM model. The left panel shows the distribution for Matzke-Klauer priors. Here, the density is localized in the corners of the plot implying four populations of participants: (1) perfect source memory and never guessing correctly; (2) perfect source memory and always guessing correctly; (3) no source memory and never guessing correctly; and (4) no source memory and always guessing correctly. (In general, the model will predict a mixture of 2^k populations, where k is the number of parameters.) This pattern seems undesirable; after all, priors in line with MPT-modelers' expectations would spread prior mass more evenly across all combinations of the two parameters instead of the extremes. Thus, the Matzke-Klauer priors neither yield sensible predictions for any single individual nor for the population distribution.

The predicted mixture distribution works against the generally desired hierarchical shrinking of individual parameter estimates (i.e., partial pooling) and instead leads to prior shrinking (i.e., estimates are pushed the extremes). As a result, the peaked and extreme prior distributions requires more data to be overpowered. In fields such as memory research, the problem may be particularly serious. As memory capacity is limited, participants are presented often with no more than 30 items per source (which is one reason why data in this field are often aggregated; Chechile, 2009). In scarce data environment, extreme priors may influence posterior estimates, especially the individual estimates.

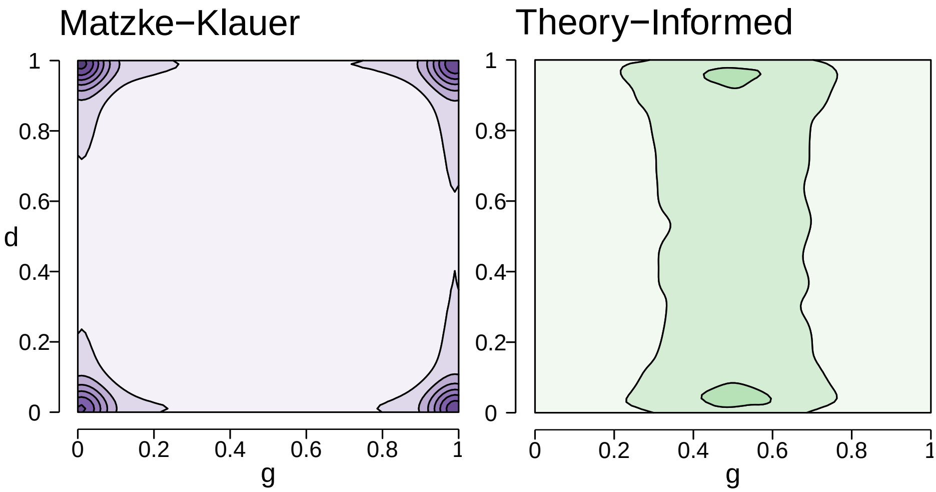


Figure 10.3: Illustration of the bivariate population distribution of participants for the source memory (d) and source guessing (g) parameters of the 2HTSM model with Matzke-Klauer priors (left panel; purple) and theory-informed priors (right panel; green). Darker colors indicates a higher density. When assigning Matzke-Klauer priors, the model predicts a mixture of four populations. Each population realizes one extreme combination of guessing and source memory. In contrast, theory-based priors cover the space of possible values for d and g more evenly, that is, extreme values are favored less.

10.2.3 Refinement No. 1: Determine Theory-Informed Parameter Priors

To prevent parameter priors from jeopardizing the modeling process, researchers need to pay due attention to their specification (Barnard, McCulloch, & Meng, 2000; Lee & Vanpaemel, 2018). Part of the model specification should be to place appropriate theory-based restrictions on the priors for group-level MPT parameters and on the variability between participants encoded in the covariance matrix. Importantly, we do not advertise any particular alternative default priors. As Vanpaemel (2010, p. 495) states: “No formal guidelines about how to capture theory into a prior exist, just like there are no formal guidelines about how to capture theory into a model equation. Model building crucially depends on the skill and the creativity of the modeler and cannot be automated.” Instead of proposing default priors, we advice using visualizations of prior predictions at the group and individual level to guide the development of appropriate models (Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019; Schad, Betancourt, & Vasishth, 2021; Wagenmakers et al., 2021). An example for this practice in MPT research can be found, for instance, in Gronau et al. (2019).

Based on the above considerations, we have made the following adjustments to the Matzke-Klauer priors for the 2HTSM model. First, we replaced the uniform prior distribution for the guessing parameter with a prior distribution that mildly favors values around the nominal guessing level. Additionally, to counteract the extreme predictions about the population distribution, we placed stronger constraints on the participant variability encoded in the covariance matrix. To determine the exact prior distributions we proceeded as follows. First, we explored a set of prior distributions that adequately implemented our assumption that participants are relatively similar, resulting in 48 plausible prior settings. Then, we conducted a small simulation study in which we visualized the predictions for data from these priors. We settled on priors that yielded individual-level predictions that were uninformative and had little bimodality. The set of priors that resulted in the best model predictions is the one depicted on the right panel in Figure 10.2 and 10.3. In this model, small deviations from the group-level means are favored over large ones, resulting in a more homogeneous population. In addition, our priors favor moderate to low correlations between MPT parameters in the range of $[-0.5, 0.5]$ over anything more extreme. By imposing these restrictions, the predictions of the model are now better in line with domain knowledge.

10.3 Testing Theory-Informed Ordinal Predictions

The previous section outlined the role of theory-informed prior distributions for the general model specification. This section explains how to make use of theory-informed prior distributions for model comparisons in the context of a particular study design. That is, we further adjusted parameter priors across experimental conditions so that they conform to theory-derived hypotheses. In particular, we focus on hypotheses on specific orderings of parameters across conditions, that is, ordinal or disordinal interactions.

A disordinal interaction was predicted, for instance, in Bell et al.'s study on biases. Specifically, the authors investigated how appearance-based biases would affect person memory. The authors were interested in a directional model which predicted that unexpected information should be remembered easier than expected information. Furthermore, this effect should be larger if the unexpected information was negative than if the unexpected information was positive.

An example of ordinal interactions that extend over several factors can be found in the study by Symeonidou and Kuhlmann (2021). Symeonidou and Kuhlmann (2021) proposed a test which measures source memory more accurately than common testing methods. The authors predicted therefore that their test would outperform the standard test. In addition, they predicted that this added benefit was influenced by how often subjects were allowed to repeat source-item pairs and by the types of sources used. Overall, then, predictions were made about the ordinal relations between one main effect and two interactions (i.e., the benefits of the novel test should be *larger* in one condition than another).

Frequentist approaches test ordinal and disordinal interactions through reparametrization of the MPT parameters (Knapp & Batchelder, 2004) but so far have only been applied for non-hierarchical models and interactions between no more than two factors (e.g., Kuhlmann et al., 2019; Moshagen, 2010; see Klauer et al., 2015 for an application to confidence or Likert scales). In principle, within the Bayesian framework disordinal interactions can be implemented through reparametrization. However, this technique adds a layer of complexity to the assignment of appropriate prior distributions, since the reparameterized model requires adjusted priors that are coherent on the original scale (Heck & Wagenmakers, 2016). Moreover, so far the literature on Bayesian MPT models has focused primarily on parameter estimation methods but not on model comparison by means of the Bayes factor (Jeffreys, 1935; Kass & Raftery, 1995).

10.3.1 Refinement No. 2: Specify Ordinal Expectations As Competing Statistical Models

Computing the Bayes factor for cognitive models is computationally complicated. However, when expectations are expressed in the parameter priors directly and they concern either point null hypotheses or directional hypotheses (e.g., interactions), the problem can be greatly simplified. The methods discussed here have three important advantages over the frequentist method. First, there is no need to reparameterize the MPT parameters in order to represent interaction effects, which facilitates the interpretation of the estimates and the assignment of prior distributions. Second, the methods are suited for hierarchical models, thus taking into account participant heterogeneity and hierarchical shrinkage. Third, the methods are able to test theories directly. Interaction effects such as the ones predicted in Bell et al. (2015) and Symeonidou and Kuhlmann (2021) are typically tested in a traditional ANOVA approach. That is, seeking to reject the null hypothesis that there is no interaction, which carries no information about the validity of the model. Only when subsequently analyzing the contrasts between the conditions it becomes evident whether or not the data adhere to the predicted pattern. To maximize efficiency and theoretical information, however, it is desirable to test the predicted pattern simultaneously. The model comparison method described here is able to do so.

For model comparisons between a point null and the encompassing model (i.e., a model that imposes no constraints on the parameters), the Bayes factor simplifies to the Savage-Dickey density ratio (Dickey, 1971; Dickey & Lientz, 1970). For model comparisons between a directional prediction and the encompassing model, the Bayes factor simplifies to the unconditional encompassing Bayes factor (Gelfand et al., 1992; Klugkist et al., 2005; Sedransk et al., 1985).

The Savage-Dickey Bayes factor is defined as the ratio of prior and posterior density under the encompassing model at the point of interest. To illustrate the approach consider Bell et al.'s null model predicting that source memory is equal for unexpected and expected information in both experimental conditions. That is, for each condition the difference between the source-level parameters is zero.

10.3.1.1 Testing Equality Constraints

We first explain how to quantify the evidence for the prediction that source memory is equal for unexpected and expected information in one experimental condition. For that, we need the prior and posterior density of the difference between the source memory parameters for unexpected and expected information. Then, we compute the height of this density at zero. If the height of the density at zero is larger for the prior density than for the posterior density, it implies evidence against the hypothesis. In this case, the data have caused more density to be allocated to a different part of the distribution. If the height of the density at the zero point is lower for the prior density than for the posterior density, this is evidence in favor of the hypothesis. In this case, the data have caused values around the zero point to become more likely.

In some cases (mostly for prior densities) the height of the distribution at the point of interest is available in closed form and can thus be computed directly. In cases where the distributions cannot be obtained in closed form (i.e., posterior densities) the density at the point of interest needs to be approximated using Markov Chain Monte Carlo (MCMC) samples. The approximated density can then be calculated, for instance, using logspline nonparametric density estimates (Kooperberg, 2020; Stone, Hansen, Kooperberg, & Truong, 1997), or by fitting a normal distribution to the MCMC samples using the methods of moments (Morey, Rouder, Pratte, & Speckman, 2011). In the study by Bell et al. (2015) we can compute the Savage-Dickey Bayes factor for each experimental condition. Assuming that these conditions are independent, the Bayes factor for the joint point null model to the encompassing model is the product of these quantities.

Computing the Savage-Dickey Bayes factor, requires that the point null model can be derived from truncating the prior distribution of the encompassing model. Furthermore, the test-relevant parameters should be independent of all other parameters in the model (Heck, 2019; Wetzels et al., 2010). In MPT research this requirement is met when (1) expectations concern group-level MPT parameters (e.g., group-level source memory under different experimental conditions) and (2) independent priors are assigned to these parameters, as is the case of the current instance of the 2HTSM model. Caution should be taken when testing parameters at a lower level, for instance, individual-level MPT parameters.

10.3.1.2 Testing Ordinal Constraints

Let us turn to directional hypotheses. The unconditional encompassing Bayes factor is defined as the proportion of prior and posterior density in agreement with the prediction. The Bayes factor identity of this method is a generalization of the Savage-Dickey density ratio (Wetzels et al., 2010). For this method, it is not necessary to approximate the prior or posterior densities. Instead, the unconditional encompassing method is a simple counting method. The unconditional encompassing Bayes factor is defined as the ratio of the sample proportions of the prior and posterior draws of the encompassing model that match the restriction. Thus, only the number of prior and posterior MCMC samples obtained from the encompassing model satisfying the constraint need to be counted.

Consider Bell et al.'s directional model that unexpected information should be remembered easier than expected information and that this effect should be larger in one experimental condition than the other. When we have MCMC samples available across all conditions, we can evaluate whether the prediction is true for each iteration. Concretely whether the difference in the source memory parameter for unexpected information and expected information is positive (implying that unexpected information is remembered easier) and that the difference between the source memory parameters is larger in one experimental condition than the other.

Here again, if the proportion of prior samples in agreement with the constraint is larger than the of proportion of samples in agreement with the constraint, it implies evidence against the hypothesis. Vice versa, if the proportion of prior samples in agreement with the constraint is smaller than the of proportion of posterior samples in agreement with the constraint, it implies evidence in favor of the hypothesis.

10.4 Case studies

In this section we illustrate parameter estimation and model comparison using theory-informed priors for two published studies on source memory. The first study by Bell et al. (2015) embedded source memory in a theory on schema-congruence. The second study by Symeonidou and Kuhlmann (2021) examined whether source memory improves when participants are assessed with a test newly developed by the authors and designed to facilitate source retrieval compared to the currently used source memory test.

10.4.1 Disclosures

10.4.1.1 Availability of Data and Code

Readers can access the R code to reproduce all analyses (including the prior simulation study and the creation of all figures), in our OSF folder at: `https://osf.io/bpgc5/`. The data needed to run the reanalysis from Bell et al. (2015) were kindly provided by the authors and can be accessed in in OSF folder. The data needed to run the reanalysis from Symeonidou and Kuhlmann (2021) are openly available on the OSF repository of the authors at `https://osf.io/6nzjs/`.

10.4.1.2 Ethical Approval

This is a non-empirical study and therefore did not require ethical approval.

10.4.2 Models

Both case studies implemented slightly adapted versions of the 2HTSM model. In contrast to the model presented in Figure 10.1, Bell et al. (2015) assumed that the item memory D and the source memory d for sources A and B are not necessarily equivalently good. Thus, their model estimated these parameters separately for the respective source. In contrast, item memory for new items was defined as the average item memory of source A and source B, that is, $D_{\text{New}} = (D_{\text{A}} + D_{\text{B}})/2$. In addition, the authors estimated separate MPT parameters in each of the two within-subjects conditions. Thus, for each participant, twelve MPT parameters were estimated, that is, two pairs of item memory and source memory parameters and two guessing parameters in each condition.

Symeonidou and Kuhlmann (2021) adhered to the structure of Figure 10.1, but estimated separate MPT parameters in each of the four between-subjects conditions and the two within-subjects conditions. Thus, for each participant, eight MPT parameters were estimated, two pairs of item memory and source memory parameters and two pairs of guessing parameters.

We implemented the hierarchical 2HTSM model with the modifications outlined in the previous sections and detailed in the appendix. That is, we assumed item homogeneity, assigned a Normal$(0, 0.28)$ prior distribution to the group-level guessing parameters, a LKJ(1) prior distribution to the correlation matrix, and a Gamma$(2, 3)$ prior distribution to the individual shift parameters.

10.4.3 Method

Parameter estimation and model comparison was based on 30,000 samples from the posterior distribution of the encompassing model with 6,000 samples as burn-in. In both case studies, the relevant comparison was between the null model and the restricted model expressing ordinal predictions. In addition, we report the Bayes factor between the encompassing model and the restricted model as a "bookend" comparison. Bookends in model comparison have been suggested as a proxy of model adequacy, that is, a model is only considered adequate if it can outperform the most restrictive model (i.e., the null model) as well as the most vague model (i.e., the encompassing model; Lee et al., 2019).

We derived the Bayes factors for these comparisons as follows. The quantities related to the prior distributions, that is the height of the prior density at the point of interest and the proportion of the prior parameter space that conformed to the constraints, were available in analytic form. This follows from the fact that the prior distributions of the group-level source memory parameters are known and simple (i.e., standard normal distributions). For instance, assume a null model which predicts that in source memory is the same in the within-subjects conditions. The distribution of this difference yields a Normal$(0, \sqrt{2})$ distribution (i.e., difference between two standard normal distributions) and evaluating this density at zero yields 0.282. Similarly, assume a restricted model which predicts that source memory is greater in one within-subjects condition than the other. Since the Normal$(0, \sqrt{2})$ distribution is symmetric at zero, half of the prior mass satisfies the prediction, yielding 0.5 as the proportion of prior density in accordance with the constraint. The quantities related to the posterior distributions were not available in analytic form and instead were approximated using MCMC samples. The posterior samples were used to approximate the logspline nonparametric density estimates which were evaluated at the points of interest, and to calculate the proportion of draws in agreement with the constraints.

With the Savage-Dickey Bayes factor and the unconditional encompassing Bayes factor at hand, the desired comparison between the null and the restricted hypothesis was obtained through transitivity. For the main hypothesis (i.e., the hypothesis concerning group-level parameters) we computed the Bayes factor 100 times to quantify the uncertainty of the Bayes factor estimates (i.e., repeatedly sampled from the posterior distribution). Bayes factors used to illustrate a concept (i.e., Matzke-Klauer Bayes factors, and Bayes factors assessing individual differences) were computed only once.

10.4.4 Reanalysis of Bell et al. 2015

To illustrate our approach for the case of disordinal interactions, we conducted a reanalysis of the data from Experiment 1 and 2 presented in Bell et al. (2015). Both experiments investigated how appearance-based first impressions affect person memory. Figure 10.4 shows a schematic illustration of the experimental setup. Participants were instructed to memorize person information, that is, pictures of faces and short behaviour descriptions. Importantly, the faces were chosen to be either pleasant looking or disgusting looking. Similarly the behaviour descriptions were categorized as being either pleasant or disgusting. In the subsequent test phase, participants were presented with pictures of faces again. Their task was to indicate whether they had seen these faces before and if so, whether their behaviour had been pleasant or disgusting.

10.4.4.1 Data

The individual-level data for the study were kindly provided by the authors. The experiments feature data from 138 and 114 participants, respectively. Each participant was instructed to learn 40 face-behaviour pairs randomly drawn from an item pool, with 10 falling into each of the four cells of the experimental design. In the test phase, 40 additional faces were introduced. Thus, each participant provided a total of 80 data points.

10.4.4.2 Hypothesis

The authors hypothesized that inconsistent information should be memorized easier than consistent information and that the effect should be larger in the disgusting behaviour condition than in the pleasant behaviour condition (see Figure 10.5). This hypothesis was based on the findings by Bell, Buchner, Kroneisen, and Giang (2012) who suggested the existence of a cognitive mechanism which emphasizes events that contradict expectations. The predicted interaction was based on the assumption of a negativity or threat bias, that is, unexpected negative information would be remembered easier than unexpected positive information (e.g., Bell & Buchner, 2010). We compared the restricted model to the null model which predicts that the source memory parameters are equal for both behaviour descriptions and the encompassing model which makes no predictions (i.e., all parameters are free to vary).

10.4.4.3 Results

For both experiments, the group-level parameter estimates and the individual-level parameter estimates are displayed in Figure 10.6. For Experiment 1, the data suggests weak evidence in favor for the null model relative to the restricted model. The Bayes factor centers around 2.86 and ranges from 2.60 to 3.18. This result is consistent with the Bayes factor estimate obtained from the Matzke-Klauer prior. Using this prior, we obtained a Bayes factor of 3.03 in favor of the null hypothesis.

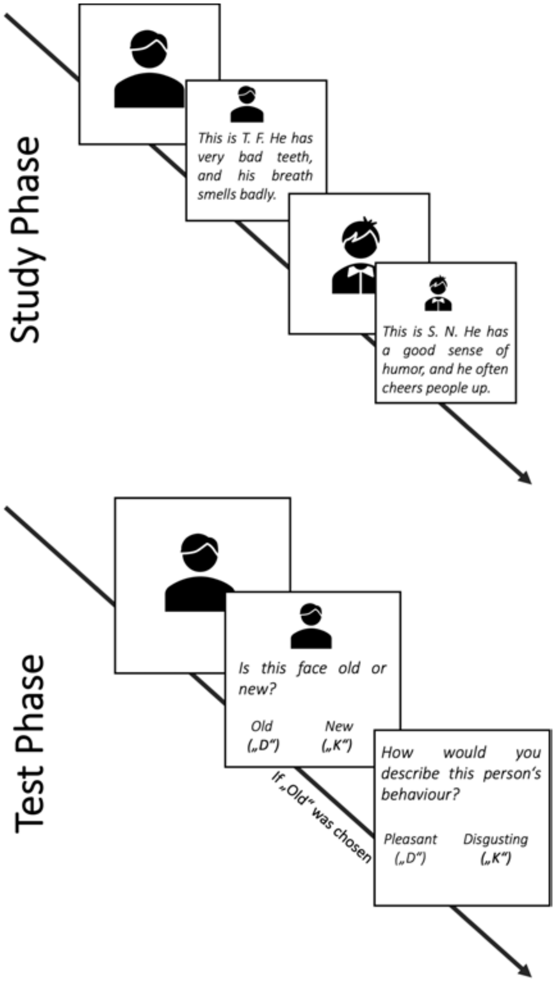

Study Phase
This is T. F. He has very bad teeth, and his breath smells badly.
This is S. N. He has a good sense of humor, and he often cheers people up.
Test Phase
Is this face old or new?
Old („D")
New („K")
If „Old" was chosen
How would you describe this person's behaviour?
Pleasant („D")
Disgusting („K")

Figure 10.4: Schematic illustration of the experimental procedure in Bell et al. (2015) relevant to the reanalysis. In the study phase (top), participants were shown faces and behaviour descriptions of persons that were either disgusting or pleasant. In the test phase (bottom), participants were presented with faces again. If participants indicated that they had already seen the faces in the study phase, they had to indicate whether the person's behavior was disgusting or pleasant.

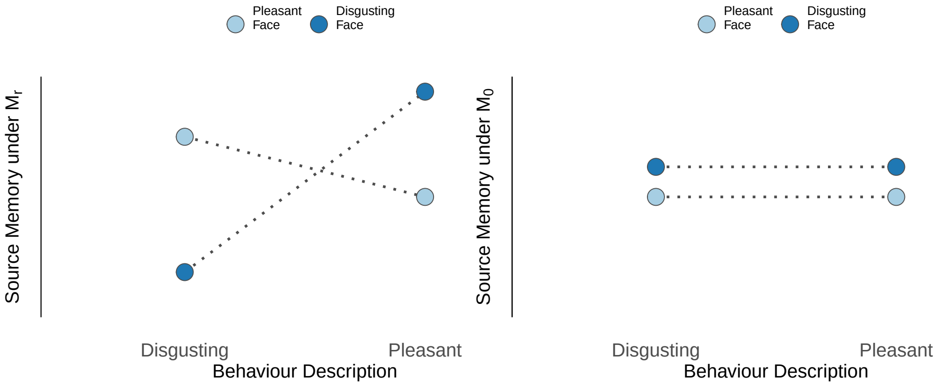


Figure 10.5: Schematic representation of the group-level source memory parameter d under the restricted model (left) and the null model (right). The restricted model makes predictions both about the ordering of the source memory parameters within the behaviour description conditions but also in within the faces conditions. The null hypothesis predicts that the source memory should be equal for both behaviour descriptions.

Table 10.1: Estimates for the group-level source memory parameter d for the data of Experiment 1 and 2 in Bell et al. (2015). The column "Reported" shows the estimates as reported in the original manuscript using frequentist estimation on aggregated data. The columns "Theory-Informed" and "Matzke-Klauer" show the median estimates and 95% credible intervals when using Bayesian hierarchical approaches.

			Source Memory d		
	Face	Behavior	Theory-Informed	Reported	Matzke-Klauer
Exp. 1	Pleasant	Pleasant	.28 [.07, .48]	.35 [.20, .50]	.21 [.02, .43]
		Disgusting	.30 [.15, .41]	.31 [.21, .41]	.31 [.17, .41]
	Disgusting	Pleasant	.41 [.29, .50]	.49 [.43, .55]	.42 [.28, .51]
		Disgusting	.19 [.02, .40]	.03 [.00, .30]	.09 [.00, .31]
Exp. 2	Pleasant	Pleasant	.18 [.02, .37]	.05 [.00, .30]	.14 [.01, .33]
		Disgusting	.30 [.15, .41]	.40 [.31, .49]	.31 [.17, .41]
	Disgusting	Pleasant	.44 [.32, .52]	.47 [.40, .55]	.45 [.34, .53]
		Disgusting	.14 [.01, .33]	.10 [.00, .32]	.08 [.00, .26]

Note. Data of reported estimates are extracted from Figure 3 and Figure 4 of the original manuscript.

Although we cannot draw a firm conclusion, the data suggest some evidence for the null model in that unexpected information and expected information are recalled equally well in Experiment 1. Thus this analysis does not support (but also does not clearly contradict) the results from the frequentist analysis of the aggregate data reported by Bell et al. (2015) which suggested the presence of a disordinal interaction. This discrepancy may be partially explained by the fact that the parameter estimates from the hierarchical analysis are very uncertain, as can be seen in the left panel in Figure 10.6. Although the point estimates descriptively conform to the predicted disordinal interaction, the large uncertainty in the estimates leads to inconclusive evidence. In the bookend comparison between the restricted and the encompassing model, the data are uninformative. The Bayes factor centers around 1.12 and ranges from 1.05 to 1.20. The restricted model can thus not outperform the encompassing model.

For Experiment 2 the data suggests weak evidence in favor for the restricted model relative to the null model. The Bayes factor centers around 2.31 and ranges from 1.92 to 2.66. These estimates are somewhat higher than the Bayes factor estimate of 0.87 we receive using the Matzke-Klauer prior which suggests no evidence. Thus for Experiment 2, the evidence is again inconclusive. As for the bookend model, the data again are uninformative. The Bayes factor centers around 1.71 and ranges from 1.63 to 1.79.

Table 10.1 summarizes the estimates for the source memory parameter, obtained from theory-informed prior distributions, the reported estimates in the original manuscript, and from Matzke-Klauer priors. The reported estimates—which were based on frequentist estimation on aggregated data—suggest less variability in the estimates compared to the Bayesian hierarchical models. The two Bayesian models are more similar to each other. However, one interesting observation concerns the source-memory estimates regarding the pairing of disgusting face and disgusting behaviour in both experiments: here the median of the theory-informed prior is much higher than the estimates from the other approaches. However, it should also be noted that all estimates have credible/confidence intervals with lower bounds close to zero. Regarding the source-guessing parameter (Table 10.2), Experiment 1 shows a discrepancy in the parameters in the pleasant face condition. The estimates obtained from the theory-informed priors and the reported estimates are better in agreement and larger than the estimates obtained from Matzke-Klauer priors. In the disgusting face condition the two Bayesian models yield similar estimates that are smaller than the reported ones.

10.4.4.4 Assessing Individual Differences

When predicting a specific pattern of effects on cognitive parameters, researchers might be interested in whether the patterns observed on the aggregate level also generalize to the individual level (Haaf & Rouder, 2019; W. Miller J. amd Schwarz, 2018). That is, whether for the biases on the population-level, are exhibited by all individuals within the population.

Another benefit of Bayesian hierarchical models is that they can not only estimate overall effects, but are also suited to test effects on the individual-level. Note that the studies by Bell et al. (2015) and Symeonidou and Kuhlmann (2021) were

Table 10.2: Estimates for the group-level source guessing parameter g for the data of Experiment 1 and 2 in Bell et al. (2015). The column "Reported" shows the estimates as reported in the original manuscript using frequentist estimation on aggregated data. The columns "Theory-Informed" and "Matzke-Klauer" show the median estimates and 95% credible intervals when using Bayesian hierarchical approaches.

		Source Guessing g		
	Face	Theory-Informed	Reported	Matzke-Klauer
Exp. 1	Pleasant	.34 [.26, .42]	.36 [.30, .42]	.31 [.24, .40]
	Disgusting	.68 [.61, .75]	.75 [.69, .80]	.70 [.62, .77]
Exp. 2	Pleasant	.36 [.30, .44]	.30 [.24, .37]	.35 [.28, .42]
	Disgusting	.66 [.59, .72]	.67 [.61, .73]	.68 [.61, .73]

Note. Data of reported estimates are extracted from Table 4 in the original manuscript.

designed to test group-level hypotheses, that is, the authors did not pose research questions concerning individual-level effects. As will become apparent below, the individual-level parameter estimates are also highly uncertain which hampers statistical inference. Nevertheless, researchers interested in assessing individual-level effects might find the following demonstration valuable.

To study individual effects, we can assess whether descriptively the point estimates depicted at the bottom panel in Figure 10.6 crossed the diagonal line contrary to the prediction. For Experiment 1, based on the median point estimates, 55.80% of participants (i.e., 77/138) showed the predicted effect for pleasant faces, that is, they remembered the pleasant face better when it was paired with disgusting rather than with pleasant information. For disgusting faces, this was the case for 91.30% of participants (i.e., 126/138). However, when accounting for the uncertainty of the estimates (i.e., 80% credible intervals), this number dropped to 0% of participants in the pleasant face condition and 13% of participants (i.e., 18/138) in the disgusting face condition.

For Experiment 2, we see a similar pattern. For pleasant faces, 99.10% of participants (i.e., 113/114) showed the predicted effect. For disgusting faces, this was the case for all participants. Again, when accounting for the 80% credible intervals, this number dropped to 0% in the pleasant face condition and to 40.40% (i.e., 46/114) in the disgusting face condition. These results once again highlight the extent to which estimated values are subject to uncertainty.

10.4.4.5 A Principled Test To Evaluate Individual Differences

Although the descriptive statistics give an insight into whether or not the individuals show the predicted effect, assessing whether credible intervals cross the diagonal is not a principled test of whether all participants show the predicted pattern. Moreover, it yields only limited information about whether people out-

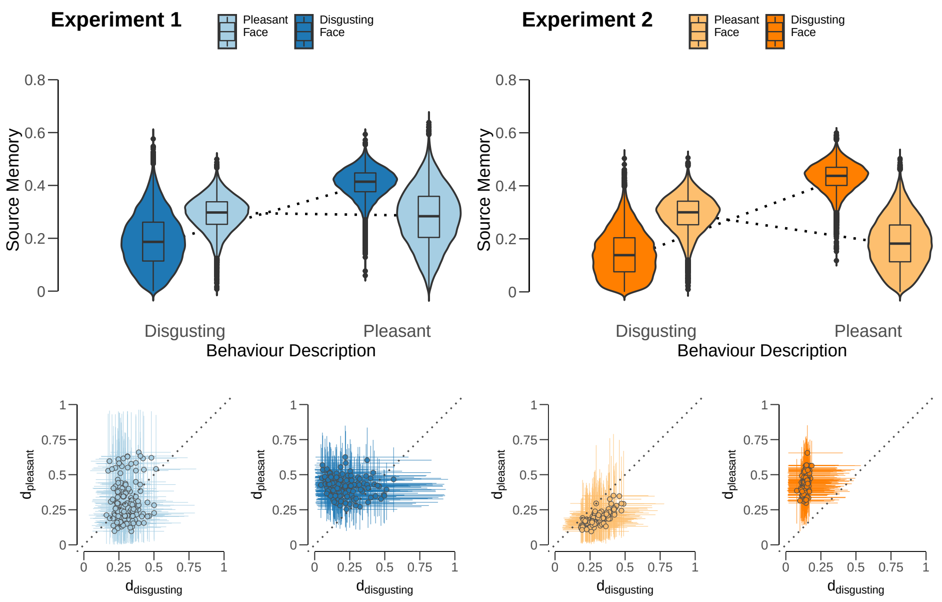


Figure 10.6: Violin plots of the estimated source memory parameters in Experiment 1 (left panel; blue) and Experiment 2 (right panel; orange) of Bell et al. (2015). In the top panels, we illustrate the group-level parameters. The bottom panel illustrates the comparison between the source memory parameters in the two behavior description conditions at the individual level. The dots represent the median estimate, the error bars the 80% credible intervals.

side of this sample would have an effect in the same direction. For a detailed discussion of the issues with this approach we refer interested readers to Haaf and Rouder (2019) and Thiele, Haaf, and Rouder (2017). Instead of counting the number of participants who show an effect in one or the other direction, we apply the model comparison approach developed by Haaf and Rouder (2017), and compare an individual-constrained model where *every* participant shows an effect in the predicted direction with an encompassing model where this constraint is not obeyed. If the encompassing model is supported, further research could shed light on the conditions under which the test is beneficial or not (Haaf & Rouder, 2017).

To examine individual differences, the same restrictions that we have previously imposed at the group-level can be applied at the individual level. That is, we now test whether the predicted biases in person memory are present in every participant at the same time. This model makes a very risky prediction: the a priori probability that all participants show the effect is only approximately 2 in ten thousand in both Experiments. In this case, the risk-taking does not pay off. Since the posterior probability that all participants show the effect is smaller than the prior probability (i.e., 1 in a million for Experiment 1 and 3 in ten thousand for Experiment 2) the hypothesis that all participants show the effect simultaneously is not supported. For Experiment 1, the data suggest strong evidence in favor for the encompassing model with a Bayes factor of 230. For Experiment 2 the data suggested inconclusive evidence. The Bayes factor in favor for the encompassing model relative to the restricted model is 0.922.

10.4.4.6 Sensitivity of individual-level Bayes factor to group-level priors

So far, in this case study the differences in theory-informed priors compared to Matzke-Klauer priors led to only negligible deviations in group-level estimates. The conclusions that one would draw from the Bayes factors of both priors differed only in experiment 2 and were not particularly large there either (weak evidence versus no evidence). The massive influence of the two priors, however, becomes apparent when evaluating the individual effects.

For Experiment 1, the Matzke-Klauer the data suggested very strong evidence *in favor of* the restricted model relative to the encompassing model with of 71. By comparison, theory-informed priors lead to opposite conclusion and provided extreme evidence *against* the restrictive model. For Experiment 2 while the Bayes factor of theory-informed priors was inconclusive, the Matzke-Klauer Bayes factor in favor of the restricted model was 144 suggesting extreme evidence.

These vastly different Bayes factors are a result of extreme participant populations predicted by Matzke-Klauer priors. The mixture of several populations causes an extreme low a priori probability that all participants will show the predicted effect simultaneously, approximately only 2 in ten million for Experiment 1 and 2 (which is one thousand times less than for theory-informed priors). The implication of this low a priori probability is that a small number of posterior samples consistent with the constraint is already sufficient to suggest evidence in favor of the restricted model. And indeed: since the posterior probability increased at least relative prior probability (i.e., 117 and 312 in 10 million) the restricted model is favored in both experiments.

When estimating individual-level parameters, the two priors also diverge with Matzke-Klauer priors appearing to cause undesirable prior shrinkage instead of hierarchical shrinkage. Figure 10.7 compares the individual estimates of source memory and guessing from the two models in the disgusting face-disgusting behaviour condition in Experiment 1. The dashed line corresponds to identical estimates from the two priors. As can be seen, the estimates for source guessing parameters (right plot) do not deviate much from this line. For the source memory parameter (left plot), on the other hand, Matzke-Klauer priors pull the parameters closer to the extremes than theory-informed priors. Here, estimates are mostly below 0.5, and are therefore pulled towards zero.

10.4.5 Reanalysis of Symeonidou and Kuhlmann (2021)

The reanalysis of the data by Symeonidou and Kuhlmann (2021) illustrate our approach for the case of ordinal interactions. Within the classical source memory paradigm, participants are asked to memorize a combination of source and content information. In standard source-monitoring test participants are then presented with only the content and need to correctly recall the source (M. K. Johnson et al., 1993). Symeonidou and Kuhlmann (2021) extended this paradigm by replac-

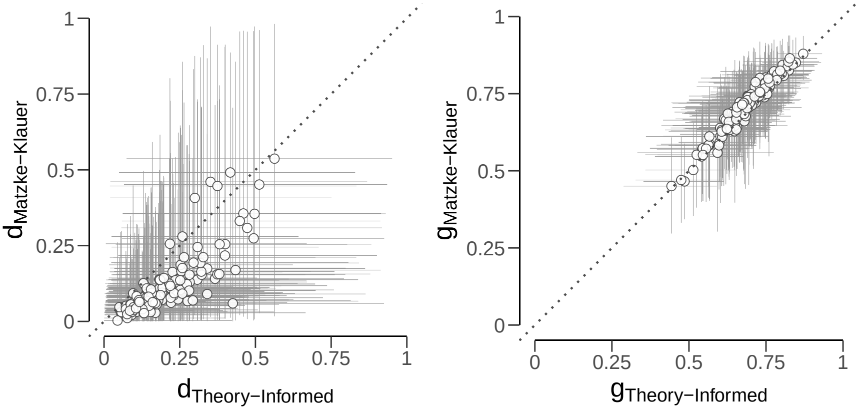


Figure 10.7: For the reanlaysis of Bell et al. (2015) Matzke-Klauer priors and theory-informed priors yield the same posterior estimates for guessing parameters (right). For the source-memory parameters (left), on the other hand, estimates from theory-informed priors are larger than Matzke-Klauer estimates.

ing the standard source-monitoring test with a reinstatement test. In this test, participants are presented with an item (i.e., a noun) and are asked to correctly identify the source it came from (i.e., which person spoke the noun). However, they also receive all possible source-content combinations, that is, the recordings of the noun by different voices, together with a picture and a name (see Figure 10.8). The authors predicted that this test would facilitate source retrieval and thus increase the probability to correctly recall the source. In their study, all participants took part in both test types (i.e., all participants completed a standard source-monitoring test followed by a reinstatement test). In addition, the authors manipulated between-participants the task difficulty (i.e., sources were either easy or hard to distinguish) and the encoding frequency (i.e, whether the source-item combination was repeated or not).

10.4.5.1 Data

The individual-level data for the study were shared by the authors on the OSF: https://osf.io/6nzjs/. The experiment features data from 146 participants across four between-subjects conditions. Each participant was instructed to learn 70 nouns, which were randomly drawn from a pool of 167 nouns. In each test phase, participants had to recognize all items and 35 new nouns were introduced as distractors. Thus, each participant provided a total of 210 data points.

10.4.5.2 Hypothesis

The hypotheses are schematically illustrated in Figure 10.9. The authors stated the following three hypotheses. First, they predicted a main effect for the test type. That is, source memory should be higher when participants completed a reinstatement test compared to the standard source-memory test. Second, the

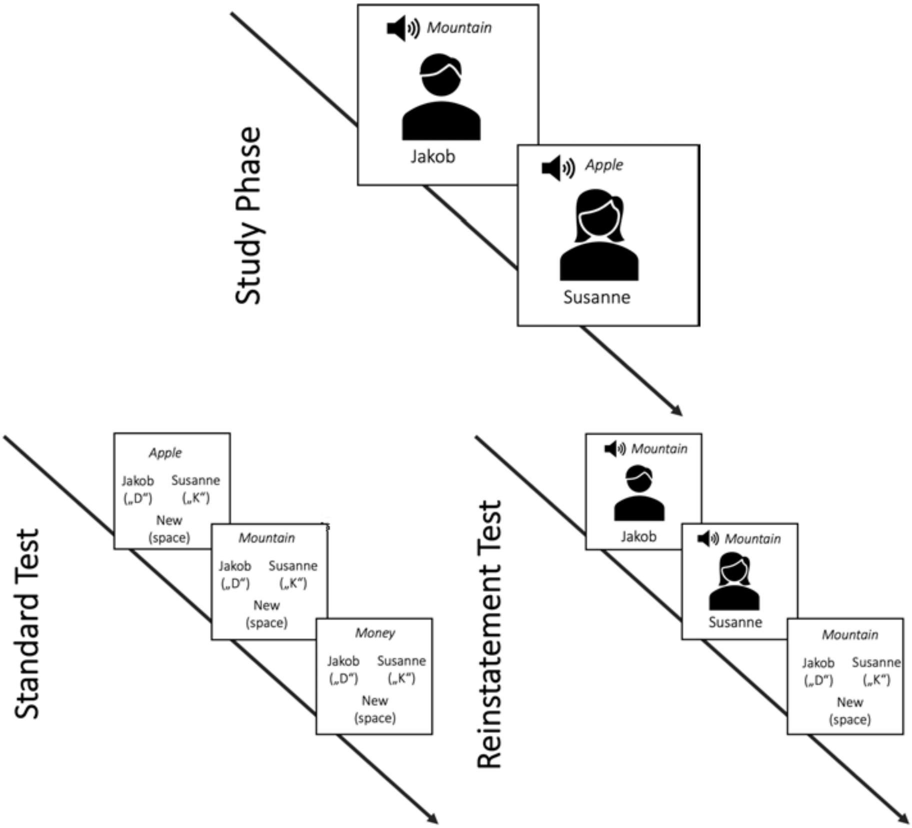


Figure 10.8: Schematic illustration of the experimental procedure in Symeonidou and Kuhlmann (2021) in the easy condition in which the items were not repeated. In the study phase (top), participants were presented with words spoken by either a male or a female voice together with a face and a name. In the standard test (bottom left), words were presented to them on the screen and the participants had to decide who had spoken them or whether they were new. In the reinstatement test (bottom right), the words were first spoken by both sources and the participants had to make the source decision afterwards. Stimuli were presented originally in German.

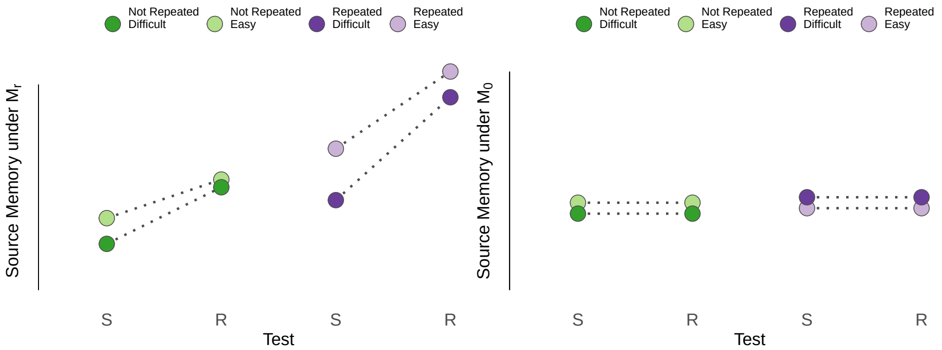


Figure 10.9: Schematic representation of how the restricted model (left) and the null model (right) predict the source memory parameter d on the group-level. The restricted model makes predictions about the ordering of the source memory parameters within the two test types. In addition, it predicts a specific interaction pattern between the test type and encoding frequency, as well es between the test type and the task difficulty. The null model predicts no source memory benefit for the reinstatement test.

authors predicted a specific interaction effect between test type and task difficulty. That is, source memory should be higher when participants completed a reinstatement test compared to the standard source-memory test and that this improvement would be greater when the sources are harder to distinguish (i.e., they are similar). Third, the authors predicted a specific interaction effect between test type and encoding frequency. That is, source memory should be higher when items were repeated compared to items that were not repeated and that this improvement is greater for the reinstatement test. This restricted model will be compared to the null model which predicts no source memory benefit for the reinstatement test and the encompassing model which makes no predictions (i.e., all parameters are free to vary).

10.4.5.3 Results

The group-level parameter estimates and the individual-level parameter estimates are displayed in Figure 10.10. The data suggest extreme evidence in favor of the restricted model relative to the null model. The Bayes factor estimate centers around 4196 and range from 2660 to 6425. Concerning the bookend comparison, the data similarly suggest very strong evidence in favor for the restricted hypothesis relative to the encompassing model. The Bayes factor estimates center around 74.49 and range from 72.12 to 76.45. When computing the Bayes factor using the Matzke-Klauer prior, the evidence is greatly reduced to a Bayes factor of 3.29 in favor of the restrictive model relative to the null model.

Table 10.3 and Table 10.4 summarize the estimates for the source memory and source guessing parameter respectively, obtained from theory-informed prior distributions, obtained from Matzke-Klauer priors, and the reported frequentist estimates in the original manuscript. The estimates using the two Bayesian ap-

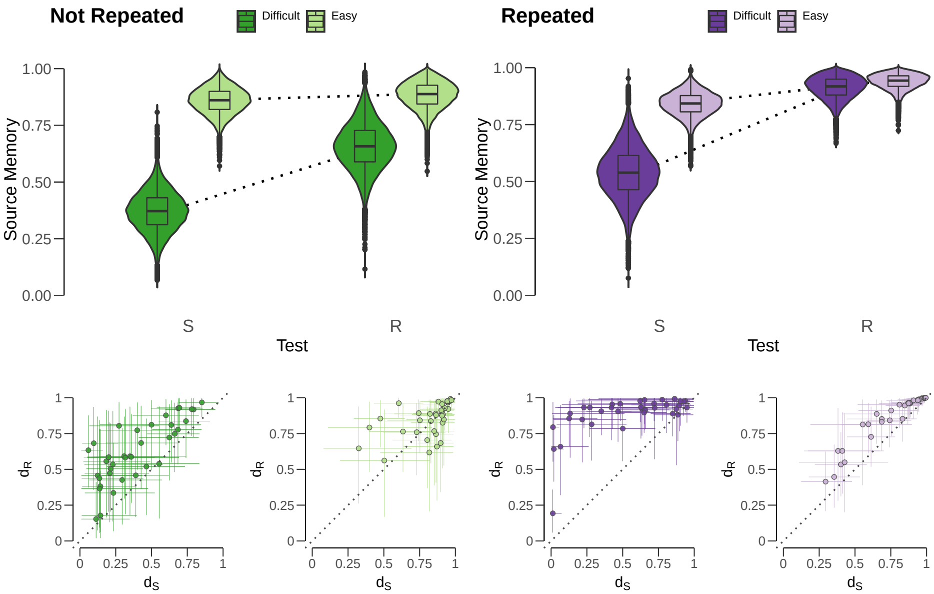


Figure 10.10: Violin plots of the estimated source memory parameters for the condition in which items were repeated once (left panel; green) or twice (right panel; purple) of Symeonidou and Kuhlmann (2021). In the top panels, we illustrate the group-level parameters. The bottom panel illustrates the comparison between the source memory parameters in for the standard and reinstatement test at the individual level. The dots represent the median estimate, the error bars the 80% credible intervals.

proaches largely converge and differences between the two are apparent only when inspecting the credible intervals. Compared to the Bayesian estimation methods, the frequentist approach leads to somewhat smaller estimates with the exception of the no-repetition difficult-task condition in the standard test where the estimates are somewhat larger.

10.4.5.4 Assessing Individual-Differences

As with the first case study, we may again assess whether participants showed the predicted reinstatement effect, that is, whether the individual 80% credible intervals are above the diagonal line at the bottom panel in Figure 10.10. When presented with the items once in the difficult condition all 37 participants showed the predicted effect, that is, they had a higher estimated source memory parameter with the reinstatement test compared to the standard test (depicted in dark green). In the easy condition that was 69.40 % of participants (i.e., 25/36; depicted in bright green). When items were repeated in the difficult condition 94.40% of participants (i.e., 34/36; depicted in dark purple) showed the predicted effect, that is, they had a higher estimated source memory parameter with the reinstatement test compared to the standard test. In the easy condition all 37 participants

Table 10.3: Estimates for the group-level source memory parameter d for the data of Symeonidou and Kuhlmann (2021). The column "Reported" shows the estimates as reported in the original manuscript using frequentist estimation on aggregated data. The columns "Theory-Informed" and "Matzke-Klauer" show the median estimates and 95% credible intervals when using Bayesian hierarchical approaches.

			Source Memory d		
Item Presented	Difficulty	Test	Theory-Informed	Reported	Matzke-Klauer
Once	Easy	Standard	.86 [.74, .96]	.82 [.45; .50]	.87 [.74; .98]
		Reinstate	.89 [.75, .98]	.86 [.77; .94]	.87 [.73; .99]
	Hard	Standard	.37 [.20, .55]	.42 [.34; .50]	.34 [.14; .54]
		Reinstate	.66 [.45, .87]	.65 [.54; .76]	.66 [.36; .94]
Twice	Easy	Standard	.84 [.72, .94]	.82 [.77; .88]	.84 [.71; .95]
		Reinstate	.94 [.86, .99]	.97 [.89; 1.04]	.95 [.86; 1.00]
	Hard	Standard	.54 [.32, .76]	.48 [.42; .55]	.55 [.27; .84]
		Reinstate	.92 [.80, .99]	.86 [.77; .95]	.94 [.83; 1.00]

Note. Default Bayes estimates based on Table D1 (Appendix D) in the original manuscript.

Table 10.4: Estimates for the group-level source guessing parameter g for the data of Symeonidou and Kuhlmann (2021). The column "Reported" shows the estimates as reported in the original manuscript using frequentist estimation on aggregated data. The columns "Theory-Informed" and "Matzke-Klauer" show the median estimates and 95% credible intervals when using Bayesian hierarchical approaches.

			Source Guessing g		
Item Presented	Difficulty	Test	Theory-Informed	Reported	Matzke-Klauer
Once	Easy	Standard	.50 [.46, .54]	.50 [.46, .53]	.50 [.46, .54]
		Reinstate	.53 [.49, .58]	.54 [.51, .57]	.54 [.49, .58]
	Hard	Standard	.48 [.45, .51]	.49 [.46, .52]	.48 [.45, .51]
		Reinstate	.46 [.43, .49]	.47 [.44, .49]	.46 [.43, .49]
Twice	Easy	Standard	.51 [.46, .55]	.51 [.48, .55]	.51 [.46, .55]
		Reinstate	.52 [.47, .57]	.52 [.49, .56]	.51 [.46, .56]
	Hard	Standard	.49 [.45, .53]	.48 [.42, .55]	.49 [.45, .52]
		Reinstate	.52 [.49, .56]	.53 [.50, .55]	.53 [.49, .56]

Note. Matzke-Klauer parameter estimates based on Table D1 (Appendix D) in the original manuscript.

showed the effect (depicted in bright purple). When accounting for the uncertainty of the estimates (i.e., 80% credible intervals), the numbers in the difficult no-repetition condition reduced to 16.20% (6/37) for difficult items and 5.60% (2/36) for easy items. When items were repeated this was 55.60% (20/36) for difficult items, 5.40% (2/37) for easy items.

Finally, we tested whether the reinstatement effect is present in each participant. The a priori probability that all participants show the effect was approximately 2 in ten thousand. In this case, again, the proportion of posterior samples in agreement with the constraints is smaller than the prior probability, approximately 2 in ten million, yielding a Bayes factor in favor for the encompassing model over the restrictive model of 8,267. This result may be somewhat surprising considering that the majority of individual median effects are in the predicted direction. However, there seemingly is not enough evidence across participants that *all of them* experience a benefit from the reinstatement test.

As in the first case study, Matzke-Klauer priors result in evidence pointing in the opposite direction from the evidence we received from theory-informed priors. In this case, no prior samples were in agreement with the constraints. Therefore, even though the number of posterior samples in agreement with the constraints was similar to the one obtained from theory informed priors (i.e., approximately 4 in ten million), the Bayes factor yield infinite evidence in favor the restricted model. And it would yield the same evidence no matter how high the posterior probability of the constraint!

Returning to the comparison of the commonly used Matzke-Klauer priors and our informed priors, we chose to visualize a condition for which Matzke-Klauer priors and theory-informed priors differ most, that is the reinstatement condition in which sources were similar and only presented once. Figure 10.11 compares for this condition the individual estimates of source memory and guessing from the two models. The dashed line highlights when the estimates are identical. For the source memory parameter (left plot), prior shrinkage of the Matzke-Klauer priors is apparent in the s-curve, indicating that Matzke-Klauer priors yield estimates pulled towards the extremes. For the guessing parameter (right plot) the Matzke-Klauer priors yield slightly more hierarchical shrinkage (corresponding to less extreme estimates). However, the guessing parameter estimates are fairly comparable. The differences between the estimates at the individual level illustrate that the problems associated with vague group-level priors mainly concern parameters that cannot be estimated with great precision.

10.5 Discussion

This chapter discussed two points of refinement for the specification and comparison of Bayesian MPT models. Specifically, we highlighted how to specify theory-informed predictions both in terms of plausible values for model parameters within experimental conditions and in terms of the rank ordering across experimental conditions. We did do so by using the 2HTSM model for source monitoring, however, our arguments and methods generalize to all MPT research. The aim of this work was to provide researchers with principles for specifying and

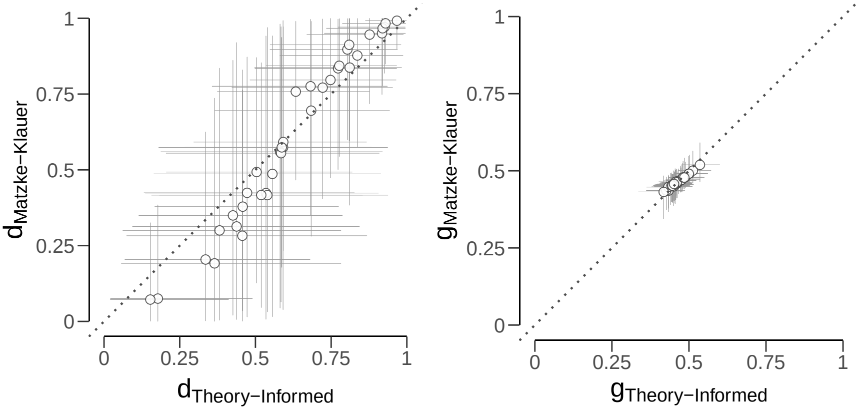


Figure 10.11: For the reanlaysis of Symeonidou and Kuhlmann (2021) Matzke-Klauer priors and theory-informed priors seem do differ systematically for source memory parameter (left) but not for source guessing (right). The figure illustrates parameters in the reinstatement condition in which sources were similar and only presented once.

comparing MPT models and to present methods that are simple to use.

To ensure plausible model predictions within experimental conditions, we argued for the need of theory-informed priors. Using prior-predictive checks, we illustrated that the priors proposed by Matzke, Dolan, et al. (2015) and Klauer (2010) and implemented in `TreeBUGS` (Heck et al., 2018) are uninformative and vague at the group level but are highly informative at the individual level and favor extreme values. These distributions predicted nonsensical response rates and—instead of a homogeneous population of participants—a mix of different participant populations. Since Matzke-Klauer priors favor extreme values, more data are needed to overwhelm them. In both case studies, the source memory parameter was still influenced by the priors, leading to prior shrinkage. Moreover, the priors massively affected Bayes factor estimates. In our case studies, Matzke-Klauer Bayes factors and Bayes factors obtained from theory-informed priors diverged, sometimes with extreme evidence for opposite hypotheses. This was especially the case when the hypothesis concerned individual differences.

However, prior shrinkage did not affect all parameters to the same extent, that is, prior shrinkage became apparent only in individual-level parameters. This is good news for MPT modelers: regarding parameter estimation on the group-level, the two priors discussed in this chapter largely converged to the same posterior estimates. However, when the goal is model comparison—even when the research question concerns only group-level parameters—Bayes factors obtained from Matzke-Klauer priors and theory-informed priors did not reliably yield the same results. In the second case study, theory-informed Bayes factor indicated very strong evidence in favor for the restricted model while the Matzke-Klauer Bayes factor indicated only moderate evidence.

Similarly, not all parameters at the individual level were affected by the prior shrinkage to the same extend, but mainly parameters that were subject to greater uncertainty. Importantly, this was the case for the case study with somewhat scarce data, but also for the case study that featured more data per participant. In MPT models, the tree architecture gives an indication on which parameters might be affected most: the more frequently parameters occur in the individual branches, the more information is available for estimation. For instance, Figure 10.1 shows that in the 2HTSM model, source guessing is featured in every branch as guessing can guide all responses. Thus, data from all trials can be used to estimate the parameter resulting in precise estimates. By contrast, source memory may be represented only in branches of previously presented items for which the source was correctly identified. MPT modellers should keep this in mind, especially when working with tree structures where the test-relevant parameters are the ones occurring in only a few branches. To inform these parameters, researchers could collect more data. However, in memory experiments, this is inherently difficult, since presenting participants with hundreds of source-items pairs for learning and retrieval is not possible. Thus, in paradigms with scarce data and/or a tree-architecture that does not sufficiently inform test-relevant parameters, the development of theory-informed prior distributions is especially important.

We would like to stress that we do not propose our priors as a new default. The appropriateness of the priors depend on the model and research design at hand. The specific prior we set on the correlation matrix, for instance, will make different predictions when more or less parameters are featured in the model. Furthermore, the priors chosen here, while making assumptions consistent with domain knowledge (e.g., source guessing centered at nominal guessing level), are far from perfect. As apparent in Figure 10.3 our priors on the source memory parameters are not completely flat but still assign a higher density to extreme values. This is due to our consideration to adapt the default priors only to the extent that the model made reasonable prior predictions. However, based on the prior predictions, it would be also legitimate to constrain the model priors even further. For instance, constraining the participant variability group-level further will lead to predictions that result in a more homogeneous population. Experienced MPT modelers will certainly be able to identify further improvements based on published literature or their own research, for instance, by centering distributions of MPT parameter on specific values.

In addition to specifying the statistical model within an experimental condition, we also discussed how to specify expected effects on model parameters across experimental conditions. Research questions in MPT research are often characterized as ordinal expectations in the form of ordinal and disordinal interactions, but so far it has been challenging to evaluate them. Commonly used methods to test ordinal expectations require the reparametrization of the MPT parameters (Knapp & Batchelder, 2004), are mainly applicable for non-hierarchical models, and are not suited to test these expectations directly.

As a refinement to current practices, we therefore suggested to compute Bayes factors using Savage-Dickey and the unconditional encompassing method (Gelfand et al., 1992; Klugkist et al., 2005; Sedransk et al., 1985). The methods discussed here have three considerable advantages over the current methods, as (1) they do not require the reparametrization of MPT parameters in order to represent interaction effects, (2) they are suited for hierarchical models, thus taking into account participant heterogeneity and hierarchical shrinkage, and (3) they are able to test theories directly. Furthermore, the methods allow researchers for testing of a wide variety of research questions, including the assessment of individual differences. Finally, the unconditional encompassing method relies simply on counting instances from the posterior distribution, and thus is intuitive and simple to use.

However, the unconditional encompassing method comes with some limitations. That is, the robustness of the Bayes factor depends largely on whether enough prior and posterior samples in agreement with the constraint can be drawn from the encompassing model in order to estimate the proportion of restricted parameter space reliably. That is, if the number of restricted parameters is large or the restricted parameter space decreases (e.g., if the data suggests extreme evidence against the restricted model) the Bayes factor results become unreliable (Sarafoglou, Haaf, et al., 2021). This issue might, for instance, occur for the assessment of individual differences in the previous section. As a first remedy, more samples can be drawn from the encompassing model to stabilize the Bayes factor estimates, but ultimately more efficient alternatives must be developed. Alternatives to the unconditional encompassing approach are the conditional encompassing method (Mulder et al., 2009) and the recently developed bridge sampling method to evaluate restricted models (Gronau et al., 2020, 2019; Sarafoglou, Haaf, et al., 2021). These methods have already been applied to test order constraints in multinomial models, but not yet to test these restrictions on the class of (hierarchical) MPT models (e.g., Heck & Davis-Stober, 2019; Sarafoglou, Aust, Marsman, Wagenmakers, & Haaf, 2021). A user-friendly implementation of these methods would be a key asset for Bayesian MPT modeling.

We presented various techniques to compute Bayes factors, however, Bayes factors are often criticized as they are sensitive to priors (e.g., Kass & Raftery, 1995). This is correct; as our case studies demonstrated, priors drastically affected the Bayes factors. Yet, the same applies to the other two steps of model specification. A different model equation will result in different Bayes factors, as will different hypotheses as they are incorporated in the model. In the formalization of theories, subjectivity comes into play at all stages of model specification–this is one of the characteristics of cognitive modeling and the construction of psychological theories in general. However, one should not equate subjectivity with randomness. The model equation, model assumptions, and parameter priors are not random: they result largely from theoretical considerations.

Our results show that Bayes factors depend strongly on the modeling choices of researchers and that thorough considerations about parameter priors play a crucial role in the analysis. However, from our own experiences, MPT modelers are often discouraged to inform their priors of fear of being accused of "Bayes factor hacking". Researchers can counteract this by well justifying their parameter priors and preregister them to ensure the confirmatory status of their analyses. In addition, researchers can determine through sensitivity analyses whether the findings are fragile or robust to aspects of the prior that are not fully justified by theory (I. J. Myung & Pitt, 1997; Sinharay & Stern, 2002). For instance, in our implementation of the 2HTSM model, our goal was to assign a prior to the source guessing parameter that was centered at the change level, however, we had no preconceived idea about the exact standard deviation of this distribution. In this case, a sensitivity analysis could examine whether the Bayes factor is robust against alternative plausible prior distributions, that is, whether different priors lead to diverging conclusions. When conducting sensitivity analyses, the alternative prior distributions should have the same theoretically justified properties as the prior distribution chosen in the analysis (i.e., a distribution centered around chance), make reasonable prior predictions, (Haaf & Rouder, 2017, 2019; Lee & Vanpaemel, 2018), and be preregistered along with the prior distribution chosen in the analysis. Sensitivity analyses are particularly justified when data are scarce (e.g., in memory research or research on clinical populations) and parameter priors can be expected to have a greater impact on the results.

10.5.1 Concluding Comments

Although prior specification is often considered a nuisance in Bayesian modeling, it offers MPT modelers the opportunity to make model evaluation a complete test of the theory. Heck et al.'s work has made it possible for many researchers to apply Bayesian MPT modeling to their data, but it also tempts researchers to rely entirely on the default settings of `TreeBUGS`. We hope that we have succeeded in drawing attention to potential problems with diffuse Matzke-Klauer priors and encouraged researchers to give the specification of priors the attention it deserves.

Even though the focus of this chapter was on source memory models, our arguments extend to all MPT research. MPT research is characterized by complex experimental designs and predictions often describe ordinal relations of parameters that span multiple factor levels. In combination with theory-informed priors, the specification of ordinal expectation brings MPT modelers closer to quantitatively describing and testing their theory to the fullest extent possible.

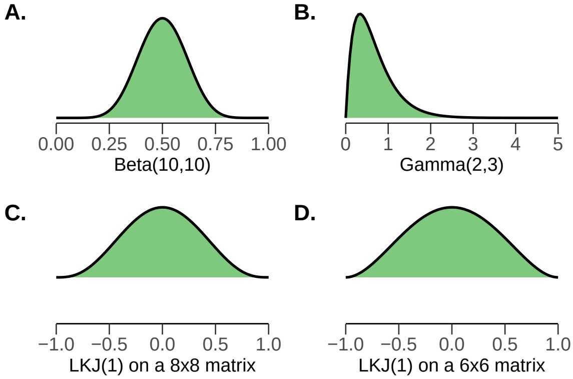


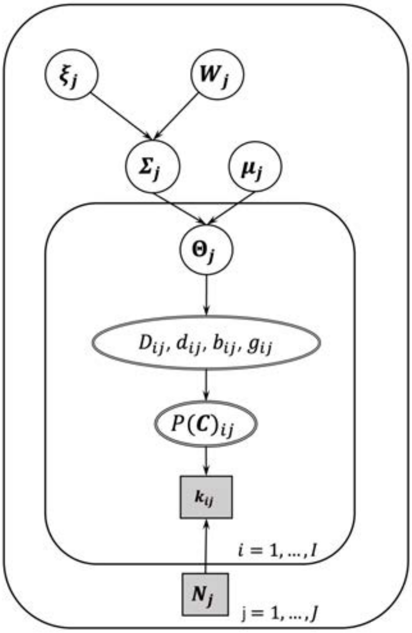
Figure 10.12: Theory-informed prior distributions for guessing parameters (A.), the standard deviations across participants (B.), and marginal correlation for LKJ(1) prior on 8x8 matrices (C.) and 6x6 matrices (D.).

10.A Model specifications

To model individual-level differences in the latent-trait approach responses are aggregated over items so that we receive a vector of category frequencies for each participant i ($i = 1, \cdots, I$) in each between-subjects experimental condition j ($j = 1, \cdots, J$). For convenience, we will drop subscript j in the following paragraphs. The individual-level MPT parameters are denoted as D_i, d_i, b_i and g_i. The function f_{MPT} that encodes the model equation translates the parameters into category probabilities $P(\mathbf{C})_i$, where a category corresponds, for instance, to the probability to answer "A" given that the correct source was Source A (i.e., $P(\text{“A”} \mid \text{Source A})_i$). Differences between the default 2HTSM model proposed in Matzke, Dolan, et al. (2015) and the theory-informed model used in this chapter lies in the prior distributions for the group-level parameters, that is, the prior distribution for the guessing parameter (panel A in Figure), the prior distribution on the standard deviations across participants (panel B), and the marginal correlation across participants (panel C and D) in Figure 10.12.

10.A.1 Default Prior Distributions

The graphical model using default prior distributions is illustrated in Figure 10.13. In this model, the vector containing the individual-level MPT parameters is probit-transformed into the vector $\boldsymbol{\theta_i}$. The vectors over all participants are then combined in a matrix $\mathbf{\Theta}$ and assigned a multivariate normal distribution as prior distribution with mean vector $\boldsymbol{\mu}$ and a covariance matrix $\mathbf{\Sigma}$. As proposed in Heck et al. (2018) and Matzke, Dolan, et al. (2015), the group-level MPT parameters are assumed to be independent so that each element in $\boldsymbol{\mu}$ is assigned an independent standard normal distribution. The covariance matrix is composed of a vector containing scaling parameters $\boldsymbol{\xi}$ and a matrix $\mathbf{W}$. In the original work by Matzke, Dolan, et al. (2015), each element in $\boldsymbol{\xi}$ is assigned a uniform distribution ranging from 0 to 100. In `TreeBUGS`, Heck et al. (2018) assign as default a uniform distribution ranging from 0 to 10. Lastly, the matrix $\mathbf{W}$ is assigned an inverse-Wishart as prior distribution with degrees of freedom $df = P + 1$, where P refers to the number of free participant parameters.



$$
\begin{aligned}
\mathbf{W} &\sim \text{Inverse Wishart}(df = P + 1) \\
\xi &\sim \text{Uniform}(0, 10) \\
\mathbf{\Sigma} &= \text{Diag}(\boldsymbol{\xi}) \times \mathbf{W} \times \text{Diag}(\boldsymbol{\xi}) \\
\\
\mu &\sim \text{Normal}(0, 1) \\
\\
\mathbf{\Theta} &\sim \text{Multivariate-Normal}(\boldsymbol{\mu}, \mathbf{\Sigma}) \\
D_i, d_i, b_i, g_i &= \phi(\boldsymbol{\theta_i}) \\
\\
P(\mathbf{C})_i &= f_{\text{MPT}}(D_i, d_i, b_i, g_i) \\
\\
\mathbf{k}_i &\sim \text{Multinomial}(P(\mathbf{C})_i, \mathbf{N})
\end{aligned}
$$

Figure 10.13: Graphical model for the default 2HTSM model as implemented in `TreeBUGS` with two within-subjects conditions and J between-subjects conditions. Index I refers to the total number of participants within each between-subjects condition.

10.A.2 Theory-Informed Prior Distributions

The graphical model using theory-informed prior distributions is illustrated in Figure 10.14. The adaptations from the classical latent-trait model are based on recommendations of Stan Development Team (2022) and Barnard et al. (2000) and implemented in Singmann (2019). In this model, the probit-transformed individual-level MPT parameters are assumed to be determined by the mean vector $\boldsymbol{\mu}$ featuring the group-level MPT parameters and by a matrix $\mathbf{\Delta}$ containing participants' individual deviations around the group mean. As in the default model, we assume that the elements in $\boldsymbol{\mu}$ are independent. We assign the group-level memory parameters standard normal distribution as prior. The guessing parameters are assigned a Normal$(0, 0.28)$ distribution as prior. The individual deviation matrix $\mathbf{\Delta}$ is obtained by drawing from a multivariate normal distribution with means $\mu = 0$. The variance-covariance matrix of this distribution is decomposed into a vector of standard deviations $\boldsymbol{\sigma}$ and a Cholesky-factorized correlation matrix $\mathbf{L}$, which are used to scale a matrix of standardized deviation $\tilde{\boldsymbol{\Delta}}$ from a standard normal distribution. We assigned each element in $\boldsymbol{\sigma}$ a Gamma$(2, 3)$ distributions, favoring small standard deviations over large ones. We further assign an LKJ prior distribution with shape $\eta = 1$ to the correlation matrix $\mathbf{L}$. This yields a marginal distribution favoring correlations among parameters in the range $[-0.5, 0.5]$ over more extreme correlations, Figure 10.12 panels C and D.

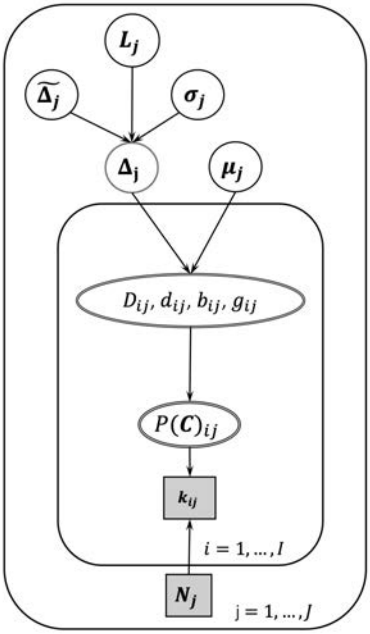


Figure 10.14: Graphical model for 2HTSM model with theory-informed priors with two within-subjects conditions and J between-subjects conditions. Index I refers to the total number of participants within each between-subjects condition.

$$\mathbf{L} \sim LKJ(\eta = 1)$$
$$\tilde{\delta} \sim \text{Normal}(0, 1)$$
$$\sigma \sim \text{Gamma}(2, 3)$$

$$\boldsymbol{\Delta} = (\text{Diag}(\boldsymbol{\sigma}) \times \mathbf{L}) \times \tilde{\boldsymbol{\Delta}})^{\top}$$

$$\mu_D, \mu_d, \sim \text{Normal}(0, 1)$$
$$\mu_b, \mu_g, \sim \text{Normal}(0, 0.28)$$

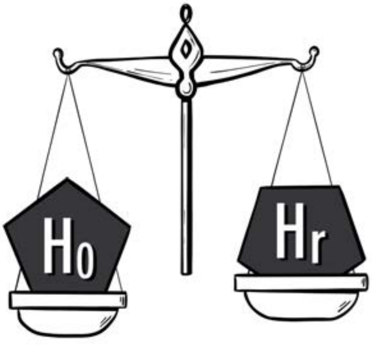
$$D_i, d_i, b_i, g_i = \phi(\mu + \delta_i)$$

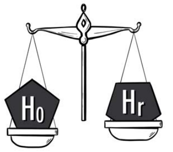
$$P(\mathbf{C})_i = f_{\text{MPT}}(D_i, d_i, b_i, g_i)$$

$$\mathbf{k}_i \sim \text{Multinomial}(P(\mathbf{C})_i, \mathbf{N})$$

Part III

Guidelines for Good Research Practices





Chapter 11

A Consensus–Based Transparency Checklist

Abstract

We present a comprehensive checklist that social and behavioural scientists can use to improve and document the transparency of their research. The checklist was created using a preregistered expert-consensus method, with guidance from 45 social and behavioural science journal editors and 18 open science advocates. The resulting checklist presents a consensus-based solution to a difficult task: identifying the most important steps needed for achieving transparent research in these fields. An accompanying online application allows users to complete the form and generate a report that they can submit with their manuscript and/or post to a public repository. Although this checklist was developed for social and behavioural researchers who conduct and report confirmatory research on primary data, other research approaches and disciplines should find value in it and adapt it to their field's needs.

This chapter is published as Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., ... & Wagenmakers, E. J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, *4*, 4–6.

We present a consensus-based checklist to improve and document the transparency of research reports in social and behavioural research. An accompanying online application allows users to complete the form and generate a report that they can submit with their manuscript or post to a public repository.

11.1 Good Science Requires Transparency

Ideally, science is characterized by a 'show me' norm, meaning that claims should be based on observations that are reported transparently, honestly and completely (Merton, 1973). When parts of the scientific process remain hidden, the trustworthiness of the associated conclusions is eroded. This erosion of trust affects the credibility not only of specific articles, but—when a lack of transparency is the norm—perhaps even entire disciplines. Transparency is required not only for evaluating and reproducing results (from the same data), but also for research synthesis and meta-analysis from the raw data and for effective replication and extension of that work. Particularly when the research is funded by public resources, transparency and openness constitute a societal obligation.

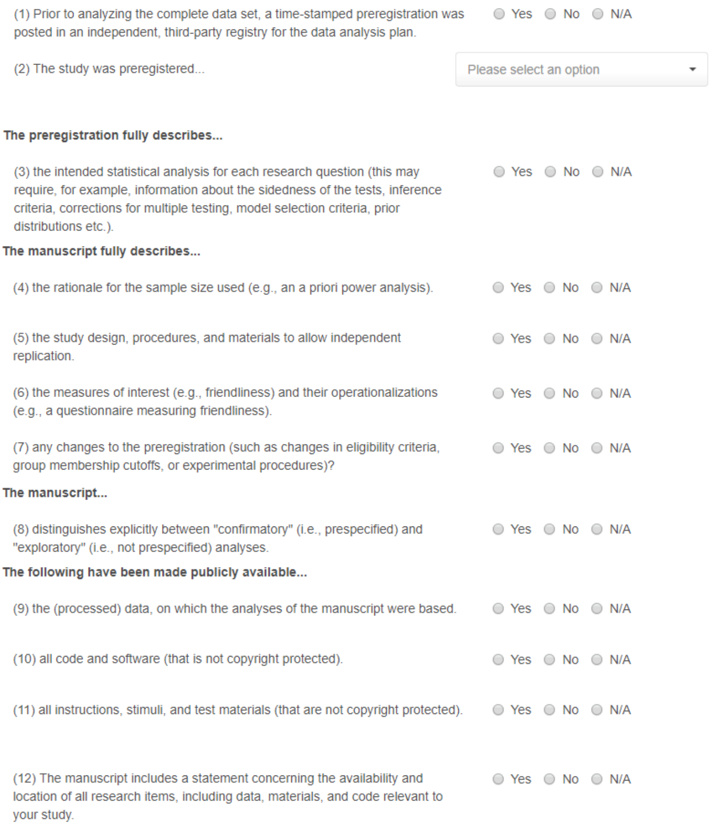
In recent years many social and behavioural scientists have expressed a lack of confidence in some past findings (Baker, 2016), partly due to unsuccessful replications. Among the causes for this low replication rate are underspecified methods, analyses and reporting practices. These research practices can be difficult to detect and can easily produce unjustifiably optimistic research reports. Such lack of transparency need not be intentional or deliberately deceptive. Human reasoning is vulnerable to a host of pernicious and often subtle biases, such as hindsight bias, confirmation bias and motivated reasoning, all of which can drive researchers to unwittingly present a distorted picture of their results.

11.1.1 The practical side of transparency

How can scientists increase the transparency of their work? To begin with, they could adopt open research practices such as study preregistration and data sharing (Chambers, 2013; Gernsbacher, 2018; Munafò et al., 2017). Many journals, institutions and funders now encourage or require researchers to adopt these practices. Some scientific subfields have seen broad initiatives to promote transparency standards for reporting and summarizing research findings, such as START, SPIRIT, PRISMA, STROBE and CONSORT (see `https://www.equator-network.org`). A few journals ask authors to answer checklist questions about statistical and methodological practices (e.g., the Nature Life Sciences Reporting Summary) (Campbell, 2013) and transparency (for example, Psychological Science). Journals can signal that they value open practices by offering 'badges' that acknowledge open data, code and materials (Kidwell et al., 2016). The Transparency and Openness Promotion (TOP) guidelines (Nosek et al., 2015), endorsed by many journals, promote the availability of all research items, including data, materials and code. Authors can declare their adherence to these TOP standards by adding a transparency statement in their articles (TOP Statement; Aalbersberg et al., 2018). Collectively, these somewhat piecemeal innovations illustrate a science-wide shift toward greater transparency in research reports.

11.2 Transparency Checklist

We provide a consensus-based, comprehensive transparency checklist that behavioural and social science researchers can use to improve and document the transparency of their research, especially for confirmatory work. The checklist reinforces the norm of transparency by identifying concrete actions that researchers can take to enhance transparency at all the major stages of the research process. Responses to the checklist items can be submitted along with a manuscript, providing reviewers, editors and, eventually, readers with critical information about the research process necessary to evaluate the robustness of a finding. Journals could adopt this checklist as a standard part of the submission process, thereby improving documentation of the transparency of the research that they publish.



(1) Prior to analyzing the complete data set, a time-stamped preregistration was posted in an independent, third-party registry for the data analysis plan. ○ Yes ○ No ○ N/A

(2) The study was preregistered... Please select an option ▾

The preregistration fully describes...

(3) the intended statistical analysis for each research question (this may require, for example, information about the sidedness of the tests, inference criteria, corrections for multiple testing, model selection criteria, prior distributions etc.). ○ Yes ○ No ○ N/A

The manuscript fully describes...

(4) the rationale for the sample size used (e.g., an a priori power analysis). ○ Yes ○ No ○ N/A

(5) the study design, procedures, and materials to allow independent replication. ○ Yes ○ No ○ N/A

(6) the measures of interest (e.g., friendliness) and their operationalizations (e.g., a questionnaire measuring friendliness). ○ Yes ○ No ○ N/A

(7) any changes to the preregistration (such as changes in eligibility criteria, group membership cutoffs, or experimental procedures)? ○ Yes ○ No ○ N/A

The manuscript...

(8) distinguishes explicitly between "confirmatory" (i.e., prespecified) and "exploratory" (i.e., not prespecified) analyses. ○ Yes ○ No ○ N/A

The following have been made publicly available...

(9) the (processed) data, on which the analyses of the manuscript were based. ○ Yes ○ No ○ N/A

(10) all code and software (that is not copyright protected). ○ Yes ○ No ○ N/A

(11) all instructions, stimuli, and test materials (that are not copyright protected). ○ Yes ○ No ○ N/A

(12) The manuscript includes a statement concerning the availability and location of all research items, including data, materials, and code relevant to your study. ○ Yes ○ No ○ N/A

Figure 11.1: The Shortened Transparency Checklist 1.0. After each section, the researchers can add free text if they find that further explanation of their response is needed. The full version of the checklist can be found at `http://www.shinyapps.org/apps/TransparencyChecklist/`.

We developed the checklist contents using a preregistered 'reactive-Delphi' expert consensus process (McKenna, 1994), with the goal of ensuring that the contents cover most of the elements relevant to transparency and accountability in behavioural research. The initial set of items was evaluated by 45 behavioural and social science journal editors-in-chief and associate editors, as well as 18 open-science advocates. The Transparency Checklist was iteratively modified by deleting, adding and rewording the items until a sufficiently high level of acceptability and consensus were reached and no strong counter arguments for single items were made (for the selection of the participants and the details of the consensus procedure see Supplementary Information). As a result, the checklist represents a consensus among these experts.

The final version of the Transparency Checklist 1.0 contains 36 items that cover four components of a study: preregistration; methods; results and discussion; and data, code and materials availability. For each item, authors select the appropriate answer from prespecified options. It is important to emphasize that none of the responses on the checklist is a priori good or bad and that the transparency report provides researchers the opportunity to explain their choices at the end of each section.

In addition to the full checklist, we provide a shortened 12-item version (Figure 11.1). By reducing the demands on researchers' time to a minimum, the shortened list may facilitate broader adoption, especially among journals that intend to promote transparency but are reluctant to ask authors to complete a 36-item list. We created online applications for the two checklists that allow users to complete the form and generate a report that they can submit with their manuscript and/or post to a public repository (Box 1). The checklist is subject to continual improvement, and users can always access the most current version on the checklist website; access to previous versions will be provided on a subpage.

This checklist presents a consensus-based solution to a difficult task: identifying the most important steps needed for achieving transparent research in the social and behavioural sciences. Although this checklist was developed for social and behavioural researchers who conduct and report confirmatory research on primary data, other research approaches and disciplines might find value in it and adapt it to their field's needs. We believe that consensus-based solutions and user-friendly tools are necessary to achieve meaningful change in scientific practice. While there may certainly remain important topics the current version fails to cover, nonetheless we trust that this version provides a useful to facilitate starting point for transparency reporting. The checklist is subject to continual improvement, and we encourage researchers, funding agencies and journals to provide feedback and recommendations. We also encourage meta-researchers to assess the use of the checklist and its impact in the transparency of research.

11.3 Disclosures

11.3.1 Data availability

All anonymized raw and processed data as well as the survey materials are publicly shared on the Open Science Framework page of the project: `https://osf.io/v5p2r/`. Our methodology and data-analysis plan were preregistered before the project. The preregistration document can be accessed at: `https://osf.io/v5p2r/registrations`.

11.3.2 Supplemental Information

Supplemental information to this chapter can be accessed via `https://www.nature.com/articles/s41562-019-0772-6#MOESM1`.

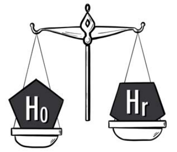
Box 1. Online applications and the benefits of the transparency checklist

Online applications for the checklist:

1. `http://www.shinyapps.org/apps/TransparencyChecklist/` for the complete, 36-item version
2. `http://www.shinyapps.org/apps/ShortTransparencyChecklist/` for the shortened, 12-item version

Benefits of the Checklist:

1. The checklist can help authors improve the transparency of their work before submission.
2. Disclosed checklist responses can help editors, reviewers, and readers gain insight into the transparency of the submitted studies.
3. Guidelines built on the checklist can be used for educational purposes and to raise the standards of social and behavioral sciences, as well as other scientific disciplines, regarding transparency and credibility.
4. Funding agencies can use a version of this checklist to improve the research culture and accelerate scientific progress.



Chapter 12

Seven Steps Toward More Transparency in Statistical Practice

Abstract

We argue that statistical practice in the social and behavioral sciences benefits from transparency, a fair acknowledgement of uncertainty, and openness to alternative interpretations. To promote such a practice, we recommend seven concrete statistical procedures: (1) visualizing data; (2) quantifying inferential uncertainty; (3) assessing data preprocessing choices; (4) reporting multiple models; (5) involving multiple analysts; (6) interpreting results modestly; and (7) sharing data and code. We discuss their benefits and limitations, and provide guidelines for adoption. Each of the seven procedures finds inspiration in Merton's ethos of science as reflected in the norms of communalism, universalism, disinterestedness, and organized skepticism. We believe that these ethical considerations –and their statistical consequences– establish common ground among data analysts, despite continuing disagreements about the foundations of statistical inference.

This chapter is published as Wagenmakers, E. J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š. ... & Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, *5*, 1473–1480.

12.1 Introduction

A superficial assessment of the published literature suggests that statisticians rarely agree on anything. Different schools –mostly frequentists, likelihoodists, and Bayesians– have fought one another tooth and nail for decades, debating the meaning of "probability", arguing about the role of prior knowledge, disputing the value of objective vs. subjective analyses, and disagreeing about the primary goal of inference itself: whether researchers should control error rates, update beliefs, or make coherent decisions. Fundamental disagreement exists not only between the different statistical schools, but is also present within the same school. For instance, within the frequentist school there is the perennial debate between those who seek to test hypotheses through p-values and those who emphasize estimation through confidence intervals; and within the Bayesian school, Jack Good's claim that there are $46,656$ varieties of Bayesians may prove an underestimate (Good, 1971; but see Aczel, Hoekstra, et al., 2020).

The disagreement also manifests itself in practical application, whenever multiple statisticians and practitioners of statistics find themselves independently analyzing the same data set. Specifically, recent "multiple-analyst" articles show that statisticians rarely used the same analysis, and often drew different conclusions, even for the exact same data set and research question (Bastiaansen et al., 2020; Botvinik–Nezer et al., 2020; Salganik, Lundberg, Kindel, Ahearn, Al-Ghoneim, et al., 2020; Silberzahn et al., 2018; van Dongen et al., 2019). Deep disagreement is also exhibited by contradictory guidelines on p-values (e.g., Amrhein, Greenland, & McShane, 2019; Benjamin et al., 2018; Harlow, Mulaik, & Steiger, 1997; McShane, Gal, Gelman, Robert, & Tackett, 2019; Wasserstein & Lazar, 2016; Wasserstein, Schirm, & Lazar, 2019). Should practitioners avoid the phrase "statistically significant"? Should they lower the p-value thresholds, or justify them, or abandon p-values altogether? And if p-values are abandoned, what should replace them? With statisticians fighting over these fundamental issues, users of applied statistics may be forgiven for adopting a wait-and-see attitude and carrying on as usual.

In this article, we claim that besides the numerous disputes and outstanding arguments, statisticians might agree on a set of scientific norms. We bring these norms to the fore, as we believe that they have considerable relevance for the practice of statistics in the social and behavioural sciences. The norms which we believe should guide statistical practice are communalism, universalism, disinterestedness, and organized skepticism, which are the four scientific norms proposed by Merton (1973)Merton, 1973 (originally published in 1942; see the Box 1 for a detailed overview of the Mertonian norms).

In general, when Mertonian norms are carried over to the field of statistics, general themes include the need to be transparent, to acknowledge uncertainty, and to be open to alternative interpretations. As such, the Mertonian norms, although proposed over half a century ago, embody the current aspirations to increase the transparency and reproducibility of science. Critically, the principles behind the Mertonian norms can be translated into concrete statistical practices. A non-exhaustive list of these practices include (1) visualizing data; (2) quantifying inferential uncertainty; (3) assessing data preprocessing choices; (4) reporting multiple models; (5) involving multiple analysts; (6) interpreting results modestly; (7) sharing data and code. We believe that most statisticians would generally endorse these practices (M. S. Anderson, Martinson, & De Vries, 2007), barring reasonable exceptions (e.g., privacy concerns, severe restrictions of time and money). In this article, we will explain these practices in more detail, including their benefits, limitations and guidelines.

Box 1. Merton's Ethos of Science

Merton Merton, 1973 proposed that scientific ethos is characterized by the following four norms:

1. Communalism. "The substantive findings of science are a product of social collaboration and are assigned to the community. (· · ·) Property rights in science are whittled down to a bare minimum by the rationale of the scientific ethic. (· · ·) The institutional conception of science as part of the public domain is linked with the imperative for communication of findings. Secrecy is the antithesis of this norm; full and open communication its enactment." (Merton, 1973, pp. 273–274)

2. Universalism. "truth-claims, whatever their source, are to be subjected to *preestablished impersonal criteria*: consonant with observation and with previously confirmed knowledge. The acceptance or rejection of claims entering the lists of science is not to depend on the personal or social attributes of their protagonist; his race, nationality, religion, class, and personal qualities are as such irrelevant." (Merton, 1973, p. 270; italics in original)

3. Disinterestedness."Science, as is the case with professions in general, includes disinterestedness as a basic institutional element. (· · ·) A passion for knowledge, idle curiosity, altruistic concern with the benefit to humanity (· · ·) have been attributed to the scientist." (Merton, 1973, pp. 275-276)

4. Organized Skepticism. This "involves a latent questioning of certain bases of established routine, authority, vested procedures and the realm of the "sacred" generally. (· · ·) Science which asks questions of fact concerning every phase of nature and society comes into psychological, not *logical*, conflict with other attitudes toward these same data which have been crystallized and frequently ritualized by other institutions. Most institutions demand unqualified faith; but the institution of science makes scepticism a virtue." (Merton, 1973, p. 264–265; italics in original)

12.2 Visualizing Data

12.2.1 Description

By visualizing data, researchers can graphically represent key aspects of the observed data as well as important properties of the statistical model applied.

12.2.2 Benefits and Examples

Data visualization is important in all phases of the statistical workflow. In exploratory data analysis, data visualization helps researchers formulate new theories and hypotheses (Tukey, 1977). In model assessment, data visualization supports the detection of model misfit and guides the development of appropriate statistical models (e.g., Gabry et al., 2019; Gelman, 2004; Heathcote, Brown, & Wagenmakers, 2015; Kerman, Gelman, Zheng, & Ding, 2008; Weissgerber, Milic, Winham, & Garovic, 2015). Finally, once the analysis is complete, visualization of data and model fit is arguably the most effective way to communicate the main findings to a scientific audience (Healy & Moody, 2014).

For an example of how data visualization facilitated the development of a new hypothesis, consider the famous "map of the distribution of deaths from cholera" created by London anaesthetist Dr. John Snow during the cholera outbreak in Soho, London in September 1854. In order to trace the source of the outbreak, Dr. Snow created a dot map that displayed the homes of the deceased as well as the water pumps in the neighborhood (Figure 12.1). The scatter of the data showed that the deaths clustered around a particular water pump in Broad Street, suggesting that the disease was waterborne instead of airborne (Gilbert, 1958). Upon Dr. Snow's request, the pump was disabled by removing its handle, which immediately ended the neighbourhood epidemic. It was discovered later that the well belonging to the pump was contaminated with sewage, which caused the outbreak in the neighborhood.

For an example of how data visualization can reveal model misspecification, consider Anscombe's quartet (Anscombe, 1973) shown in Figure Figure 12.2. The four scatter plots all have identical summary statistics (i.e., means, standard deviations, and Pearson correlation coefficient). By visually inspecting the panels, it becomes obvious that the bivariate relation is fundamentally different for each panel (see also Matejka & Fitzmaurice, 2017).

12.2.3 Current Status

Since William Playfair (1759–1823) invented the first statistical graphs –such as line graphs and bar charts (Playfair, 1786)– , data visualization has become an essential part of science. Today, graphs are part of most statistical software packages and have become an indispensable tool to perform certain analyses (e.g., principal component analysis, or prior and posterior predictive checks), or for handling big data sets (e.g., through cluster analysis Everitt, Landau, Leese, & Stahl, 2011). Technology now allows us to go beyond static visualizations and display the dynamic aspects of the data, for instance, by using the software packages R Shiny

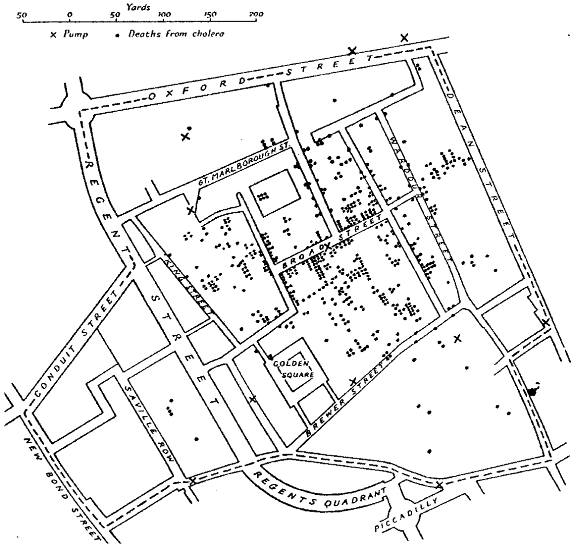


Figure 12.1: Recreation of Dr. Snow's map of the distribution of deaths from cholera. In this map, the points represent the homes of the deceased and the crosses represent the water pumps in the neighborhood. The contaminated water pump that triggered the cholera epidemic in the neighborhood is located on Broad Street. Reprinted with permission from *Pioneer maps of health and disease in England* (p. 174), by E. W. Gilbert, 1958, The Royal Geographical Society (with the Institute of British Geographers).

(W. Chang, Cheng, Allaire, Xie, & McPherson, 2020) or iNZight (iNZight Team, 2020).

12.2.4 Limitations

Despite the obvious benefits, data visualization also offers the opportunity to mislead, for instance, when displaying spurious patterns by either expanding the scale to minimize variation, or by minimizing the scale to accentuate differences (e.g., Cairo, 2019; Gelman, 2011; Wainer, 1984).

Furthermore, the informativeness of a graph often depends on the design capabilities of the researcher and how much thought they put into what information should be communicated. Scientists without programming experience often find themselves constrained by the options offered in standard graphics software. However, the example of Anscombe's quartet shows that even the simplest plots can be highly informative.

12.2.5 Guidelines

There are no uniform guidelines as to when and which graphical representations should be used. There is, however, a fundamental principle of good statistical graphics due to Tufte (Tufte, 1973, p.92): "Above all else show the data" (i.e.,

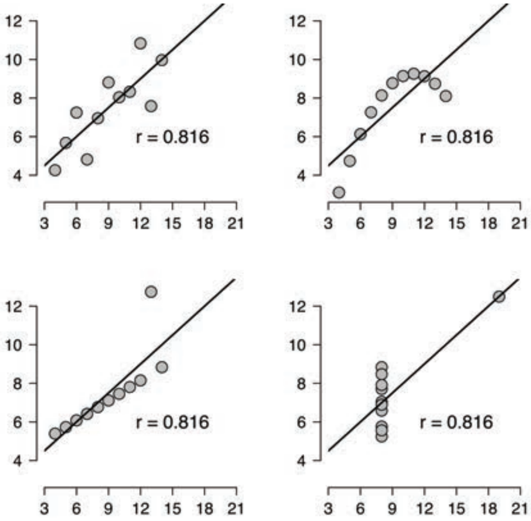


Figure 12.2: Anscombe's quartet emphasizes the importance of data visualization to detect model misspecification. Although the four data sets are equivalent in terms of their summary statistics, the Pearson correlation is only valid for the data set in the upper left panel. Figure is available at `https://tinyurl.com/y9je2mut` under CC license `https://creativecommons.org/licenses/by/2.0/`.

minimize non-data elements). In general, scientists should aim to create a graph that is as clean, informative, and as complete as possible. These characteristics are also emphasized in the ASA Ethical Guidelines (Committee on Professional Ethics of the American Statistical Association, 2018). The guidelines mention that to ensure the integrity of data and methods, the ethical statistician "[i]n publications and reports, conveys the findings in ways that are both honest and meaningful to the user/reader. This includes tables, models, and graphics" (p. 3).

Beyond that, guidelines depend on the individual aspects of the data (e.g., complexity of the data and experimental design) and context (cf. Diamond & Lerch, 1992); here we refer the interested reader to the numerous manuals describing good practices in graphical representation of statistical information (e.g., Chen, Härdle, & Unwin, 2008; Cleveland & McGill, 1984; Gelman, Pasarica, & Dodhia, 2002; Mazza, 2009; Tufte, 1973; Wilke, 2019; L. Wilkinson, 1999).

12.3 Quantifying Inferential Uncertainty

12.3.1 Description

By reporting the precision with which model parameters are estimated, the analyst communicates the inevitable uncertainty that accompanies any inference from a finite sample.

12.3.2 Benefits and Example

Only by assessing and reporting inferential uncertainty is it possible to make any claim about the degree to which results from the sample generalize to the population. For example, Strack, Martin, and Stepper (1988) studied whether participants rate cartoons to be funnier when they hold a pen with their teeth (which induces a smile) instead of holding it with their lips (which induces a pout). On a 10-point Likert scale, the authors observed a raw effect size of 0.82 units. For the interpretation of this result it is essential to know the associated inferential uncertainty. In this case, the 95% confidence interval ranges from −0.05 to 1.69, indicating that the data are not inconsistent with a large range of effect size estimates (including effect sizes that are negligible or negative).

Box 2. Seven Mertonian Statistical Procedures

This box outlines how each of the seven procedures discussed in the main manuscript fullfill the Mertonian norms. An overview is given in Table below.

	Communalism	Universalism	Disinterestedness	Organized Skepticism
1. Visualizing Data	Yes		Yes	Yes
2. Quantifying Inferential Uncertainty	Yes		Yes	Yes
3. Assessing Data Preprocessing Choices	Yes		Yes	Yes
4. Reporting Multiple Models	Yes		Yes	Yes
5. Involving Multiple Analysts		Yes	Yes	Yes
6. Interpreting Results Modestly			Yes	Yes
7. Sharing Data and Code	Yes	Yes	Yes	Yes

1. Visualizing Data

Well-designed visualizations show at a glance the key aspects of the data. Moreover, by giving the reader a more complete picture of the data and related statistics, visualizations can either support or weaken a conclusion drawn by the researcher, or help the reader find alternative ways of interpreting the results and analyzing the data.

2. Quantifying Inferential Uncertainty

Acknowledging inferential uncertainty (e.g., by presenting standard errors or confidence intervals) contributes to open communication. In addition, quantifying inferential uncertainty signals that researchers are openly acknowledging the extent to which their measurements are imprecise, especially when sample size is small. Finally, explicitly acknowledging inferential uncertainty may prompt readers to question how well the results from the sample generalize to the population.

3. Assessing Data Pre-processing Choices

When researchers share the results from only a single data pre-processing pipeline, they may unintentionally hide important information. If a result proves sensitive to particular pre-processing choices, this warrants skepticism and may initiate a debate on the importance and plausibility of relevant data pre-processing choices (cf. Leamer, 1985, p. 308).

4. Reporting Multiple Models

Similar to the previous section, reporting results from only a single model may unintentionally hide important information.

5. Involving Multiple Analysts

The multiple-analysts approach can reveal whether different (teams of) analysts reach converging or diverging conclusions from the same data set. By including other analysts with different backgrounds and interests, the potential impact of self-interest of any single analyst is counteracted. The multiple-analysts approach also stimulates skepticism by bringing to light alternative statistical perspectives on the data.

6. Interpreting Results Modestly

Disinterested analysts arguably have little need to exaggerate claims, impress reviewers, and downplay signs of model misfit. Analysts who facilitate organized skepticism do not attempt to suppress doubt — they are not defensive, and they do not wish to protect their work against good-faith scrutiny from their peers.

7. Sharing Data and Code

All secrecy about data is a limitation to knowledge accumulation and violates the ethos of science. All interested researchers should have access to relevant, properly anonymized data. Importantly, sharing data allows skeptical eyes to scrutinize the results, promoting quality control.

12.3.3 Current Status

In virtually all statistics courses, students are taught to provide not only the summary of statistical tests (such as F-, t-, p-values and associated degrees of freedom), but also parameter point-estimates (e.g., regression weights, effect sizes) and their associated uncertainty (e.g., standard error, posterior distribution, confidence intervals, credible intervals). Nevertheless, there exists a gap between what is taught and what is practiced. Studies of published articles in physiology (Weissgerber et al., 2015), the social sciences (Hoekstra, Finch, Kiers, & Johnson, 2006), and medicine (Cooper, Schriger, & Close, 2002; Schriger, Sinha, Schroter, Liu, & Altman, 2006) revealed that error bars, standard errors, or confidence intervals were not always presented. Also, popular metrics such as Cronbach's alpha (a measure of test score reliability) are virtually never presented with a measure of inferential uncertainty.

12.3.4 Limitations

Although not a limitation per se, it should be noted that inferential uncertainty always needs to be quantified relative to the inferential goal: does a researcher want to generalize across people, stimuli, time points, or another dimension? The proper way of computing standard errors depends on the researcher's purpose.

12.3.5 Guidelines

Various guidelines strongly recommend that effect size estimates are accompanied by measures of uncertainty in the form of standard errors or confidence intervals. For instance, the publication manual of the American Psychological Association (6th ed.) states: "When point estimates (e.g., sample means or regression coefficients) are provided, always include an associated measure of variability (precision), with an indication of the specific measure used (e.g., the standard error)," (p. 34). Also, the International Committee of Medical Journal Editors (International Committee of Medical Journal Editors, 2019) explicitly recommend to "[w]hen possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals)" (p. 17).

12.4 Assessing Data Preprocessing Choices

12.4.1 Description

By assessing the impact of plausible alternative data pre-processing choices (i.e., examining the "data multiverse", Steegen et al., 2016), the analyst determines the extent to which the finding under scrutiny is either fragile or sturdy.

12.4.2 Benefits and Example

A "data multiverse" analysis reveals the fragility or sturdiness of the finding under plausible alternative data pre-processing choices. This prevents researchers from falling prey to hindsight bias and motivated reasoning, which may lead them to unwittingly report only the pre-processing pipeline that yields the most compelling result (e.g., De Groot, 1956/2014; Simmons et al., 2011). But even a completely unbiased analysis will benefit from a “data multiverse” analysis, as it reveals uncertainty that would otherwise remain hidden.

For example, Steegen et al. (2016) reexamined the results of Durante, Rae, and Griskevicius (2013), who reported an interaction between relationship status (i.e., single or not) and menstrual cycle (i.e., fertile or not) on reported religiosity. After applying a series of 180 different data pre-processing procedures (e.g., five different ways to split women into high versus low fertility), the multiverse reanalysis showed that the resulting 180 p-values were distributed uniformly between 0 and 1, indicating that the reported interaction is highly fragile.

12.4.3 Current Status

The idea of assessing sensitivity to data-preprocessing choices dates back at least to (De Groot, 1956/2014, p. 190) and (Leamer, 1985, p. 308) and was revived by Simmons et al. (2011) and by Steegen et al. (2016). In the field of functional magnetic resonance imaging, both Carp (2012) and Poldrack et al. (2017) emphasized the hidden influence of different plausible pre-processing pipelines. In psychology, recent applications are Bastiaansen et al. (2020) and Wessel, Albers, Zandstra, and Heininga (2020). Nevertheless, the overwhelming majority of empirical articles does not report the results of a data multiverse analysis.

12.4.4 Limitations

A pragmatic limitation of the data multiverse lies in the extra work that it entails. Another limitation can be found in ambiguities surrounding the definition of the data multiverse. The analyst has to determine what constitutes a sufficiently representative set of pre-processing choices and whether all pre-processing choices are equally plausible, such that they should be given equal weight in the multiverse analysis. A final limitation is that it is not always clear how to interpret the results of a data multiverse analysis. Interpretation can be facilitated with certain graphical formats that cluster related pipelines (e.g., specification curves; Simonsohn, Nelson, & Simmons, 2020).

12.4.5 Guidelines

Some specific guidelines on assessing data pre-processing choices are offered by Simmons et al. (2011, see Requirements for Authors, numbers 5 and 6), but it is difficult to provide general guidelines as “$(\cdots)$ a multiverse analysis is highly context-specific and inherently subjective. Listing the alternative options for data construction requires judgment about which options can be considered reasonable and will typically depend on the experimental design, the research question, and the researchers performing the research” (Steegen et al., 2016, p. 709). More general guidelines that relate exclusively to the reporting of pre-processing choices are given in the ASA Ethical Guidelines (Committee on Professional Ethics of the American Statistical Association, 2018). These mention that to insure the integrity of data and methods, the ethical statistician “[w]hen reporting on the validity of data used, acknowledges data editing procedures, including any imputation and missing data mechanisms” (p. 2).

12.5 Reporting Multiple Models

12.5.1 Description

By assessing the impact of plausible alternative statistical models (i.e., examining the “model multiverse”), the analyst gauges the extent to which a statistical conclusion is either fragile or sturdy.

12.5.2 Benefits and Example

Similar to the “data multiverse” analysis discussion in the previous section, a model multiverse analysis examines the fragility or sturdiness of the finding under plausible alternative statistical modeling choices. Modeling choices comprise differences in estimators and fitting regimes, but also in model specification and variable selection. Reporting the outcomes of multiple plausible models reveals uncertainty that would remain hidden if only a single model were entertained. In addition, this practice protects analysts against hindsight bias and motivated reasoning, which may unwittingly lead them to select the single model that produces the most flattering conclusion. For example, Patel, Burford, and Ioannidis (2015) quantified the variability of results under different model specifications. They considered 13 clinical, environmental, and physiological variables as potential covariates for the association of 417 self-reported, clinical, and molecular phenotypes with all-cause mortality. Consequently, they computed p-values for $2^{13} = 8,192$ models and examined the instability of the inference, which they call the “vibration of effects”.

12.5.3 Current Status

Although the idea of the model multiverse dates back at least to De Groot (1956/2014) and Leamer (1985), most empirical researchers still base their conclusion on only a single analysis (but see Athey & Imbens, 2019; Levine & Renelt, 1992).

12.5.4 Limitations

As was the case for the construction of the data multiverse, a pragmatic limitation of the model multiverse lies in the extra work that it entails —for the analyst as well as the reader. Recent work suggests that the number of plausible models can be very large (i.e., Botvinik–Nezer et al., 2020; Silberzahn et al., 2018). Also, multiverses vary in their informativeness, and readers need to assess themselves whether a multiverse features notably distinct models or just runs the essentially same model multiple times. Model spaces can be overwhelming; any single analyst will naturally be drawn towards the subset of models that they are familiar with (or, unwittingly, the subset of models that yields the result that is most flattering or most in line with prior expectations). In addition, Del Giudice, Gangestad, and Steven (2021, p. 5) argue that "By inflating the size of the analysis space, the combinatorial explosion of unjustified specifications may, ironically, exaggerate the perceived exhaustiveness and authoritativeness of the multiverse while greatly reducing the informative fraction of the multiverse. At the same time, the size of the specification space can make it harder to inspect the results for potentially relevant findings. If unchecked, multiverse-style analyses can generate analytic "black holes": Massive analyses that swallow true effects of interest but, due to their perceived exhaustiveness and sheer size, trap whatever information is present in impenetrable displays and summaries."

12.5.5 Guidelines

Because the construction of the model multiverse depends on the knowledge and expertise of the analyst, it is challenging to provide general guidelines. For relatively simple regression models, however, clear guidelines do exist (e.g., Hoeting, Madigan, Raftery, & Volinsky, 1999; Patel et al., 2015). Furthermore, Simonsohn et al. (2020) suggested a specification curve analysis, and Dragicevic, Jansen, Sarma, Kay, and Chevalier (2019) suggest interactive ways of presenting the results. The ASA Ethical Guidelines (Committee on Professional Ethics of the American Statistical Association, 2018) mention that to meet the responsibilities towards funders and clients, the ethical statistician "[t]o the extent possible, presents a client or employer with choices among valid alternative statistical approaches that may vary in scope, cost, or precision" (p. 3). The ASA, however, does not mention that researchers share the same responsibility towards their scientific colleagues, although this may be implicit.

One general recommendation for constructing a comprehensive model multiverse is to collaborate with statisticians who have complementary expertise, bringing us to the next section.

12.6 Involving Multiple Analysts

12.6.1 Description

By having multiple analysts independently analyze the same data set, the researcher can decrease the impact of analyst-specific choices regarding data preprocessing and statistical modeling.

12.6.2 Benefits and Example

The multiple-analysts approach reveals the uncertainty that is due to the subjective choices of a single analyst and promotes the application of a wider range of statistical techniques. When the conclusions of the analysts converge, this bolsters one's confidence that the finding is robust; when the conclusions diverge, this undercuts that confidence and stimulates a closer look at the statistical reasons for the lack of consensus.

The multiple-analysts approach was used, for example, in a study by Silberzahn et al. (2018) where 29 teams of analysts examined, using the same dataset, whether the skin tone of soccer players influences their probability of getting a red card. While most of the analysis teams reported that players with a darker skin tone have a higher probability of getting a red card, some of the teams reported null results. The analysis approach used by the teams differed widely, both with respect to data pre-processing and statistical modeling (e.g., included covariates, link functions, assumption of hierarchical structure).

12.6.3 Current Status

A precursor to the multiple-analysts approach concerns the 1857 “Cuneiform competition”, where four scholars independently translated a previously unseen ancient Assyrian inscription (Rawlinson, Talbot, Hincks, & Oppert, 1857). The overlap between their translations –sent to the Royal Asian Society in sealed envelopes, and simultaneously opened and inspected by a separate committee of examiners– was striking and put to rest any doubts concerning the method used to decipher such inscriptions. The multiple-analysts approach never caught on in practice, although recent examples exist in psychology and neuroscience (Bastiaansen et al., 2020; Boehm, Hawkins, Brown, van Rijn, & Wagenmakers, 2016; Botvinik–Nezer et al., 2020; Dutilh, Annis, et al., 2019; Schweinsberg et al., 2020; Silberzahn et al., 2018; van Dongen et al., 2019).

12.6.4 Limitations

As was the case for the construction of the data multiverse and the model multiverse, a pragmatic limitation of the multiple analyst approach lies in the extra work that it entails, specifically with respect to (1) finding knowledgeable analysts who are interested in participating; (2) documenting the data set, describing the research question, and identifying the target of statistical inference; (3) collating the initial responses from each team, and potentially coordinating a review and feedback round. While differences in opinion should be respected, there need to be ways to filter out analysis approaches that involve clear mistakes. An additional limitation concerns possible homogeneity of the analysts. For instance, all analysts involved could be rigidly educated in the same school of thought, share cultural or social biases, or just make the same mistake. In such a case, the results may create an inflated sense of certainty in the conclusion that was reached. This potential limitations can be mitigated by selecting a diverse group of analysts and incorporating feedback and revision options in the process (Silberzahn et al., 2018), a round-table discussion (van Dongen et al., 2019) or, more systematically, a Delphi approach (Thangaratinam & Redman, 2005).

12.6.5 Guidelines

There are no explicit guidelines concerning the multiple-analysts approach. We propose that the optimal number of analysts to include depends on factors such as the complexity of the data, the importance of the research question (e.g., a clinical trial on the effectiveness of a new drug against COVID-19 warrants a relatively large number of analysts), and the probability that the analysts could reasonably reach a different conclusion (e.g., there may be multiple ways to interpret the research question, and there may be multiple dependent variables and predictor variables that could or could not be relevant).

When analysts are selected, care should be taken to ensure heterogeneity, diversity, and balance. Specifically, one should be mindful of the potential biasing effects of specific background knowledge, culture, education, and career stage of the analyst.

The ASA Guidelines emphasize the legitimacy and value in alternative analytic approaches, stating that "[t]he practice of statistics requires consideration of the entire range of possible explanations for observed phenomena, and distinct observers $(\cdots)$ can arrive at different and potentially diverging judgments about the plausibility of different explanations" (p. 5).

12.7 Interpreting Results Modestly

12.7.1 Description

By modestly interpreting the results, the analyst explicitly acknowledges any remaining doubts concerning the importance, replicability, and generalizability of the scientific claims at hand.

12.7.2 Benefits and Example

Modestly presented scientific claims enable the reader to evaluate the outcomes for what they usually are: not final, but tentative results pointing in a certain direction, with considerable uncertainty surrounding their generalizability and scope. Overselling results might lead to the misallocation of public resources towards approaches that are in fact not properly validated and not ready for application in practice. Also, researchers themselves risk losing long-term credibility for short-term gains of greater attention and higher citation counts. Moreover, after having publicly committed to a bold claim, it becomes difficult to admit that one's initial assessment was wrong; in other words, overconfidence is not conducive to scientific learning.

Scientists of true modesty remain doubtful even at moments of great success. For example, when James Chadwick found experimental proof of neutrons, the discovery that earned him the Nobel prize, he communicated it modestly under the title "Possible Existence of Neutron" (Chadwick, 1932).

12.7.3 Current Status

Tukey (1962) already remarked that "Laying aside unethical practices, one of the most dangerous $[(\cdots)$ practices of data analysis $(\cdots)]$ is the use of formal data-analytical procedures for sanctification, for the preservation of conclusions from all criticism, for the granting of an imprimatur." (p. 13). Almost 60 years later, an editorial in *Nature Human Behaviour* warns its readers about “conclusive narratives that leave no room for ambiguity or for conflicting or inconclusive results” (NHB Editorial, 2020, p. 1). Similarly, D. J. Simons, Shoda, and Lindsay (2017) suggested adding a mandatory Constraints on Generality statement in the discussion section of all primary research articles in the field of psychology to prevent authors from making wildly exaggerated claims of generality. This suggests that scientific modesty is rarer than we would expect if Mertonian norms were widely adopted. There are some clear indications of a lack of modesty. First of all, the frequency of stronger language (words like “amazing”, “groundbreaking”, “unprecedented”) seemed to have increased in the last few decades Vinkers, Tijdink, & Otte, 2015. Secondly, dichotomization of findings (i.e., ignoring the uncertainty inherent to statistical inference) is common practice (e.g., Hoekstra et al., 2006; also see paragraph 4.3). Thirdly, textbooks (which are typically a reflection of current practice) on how to write papers often explicitly encourage authors to overclaim (e.g., Bem, 1987; van Doorn et al., 2021).

12.7.4 Limitations

Publications and grants are important for scientific survival. Coupled with the fact that journals and funders often prefer groundbreaking and unequivocal outcomes, it may be detrimental to one’s success to modestly interpret the results. The encouragement of this Mertonian practice may require change at an institutional level, although some have argued that scientists should not hide behind the system when defending their behavior (Yarkoni, 2018).

12.7.5 Guidelines

There are several ways we can contribute to increasing intellectual modesty. First of all, we could encourage intellectual modesty in others’ work when we act as reviewers of papers and grant proposals (Hoekstra & Vazire, 2020). Since a reviewer’s career is independent of how they evaluate a paper, they can make a positive review conditional on a more modest presentation of outcomes. Hoekstra and Vazire (2020) present a list of suggestions for increasing modesty in the traditional sections of an empirical article, which can be used by authors as well. One example (p. 16) includes “Titles should not state or imply stronger claims than are justified (e.g., causal claims without strong evidence)”.

Also, the ASA Guidelines state: “[t]he ethical statistician is candid about any known or suspected limitations, defects, or biases in the data that may affect the integrity or reliability of the statistical analysis” (p. 2).

12.8 Sharing Data and Code

12.8.1 Description

By sharing data and analysis code, researchers provide the basis for their scientific claims. Ideally, data and code should be shared publicly, freely, and in a manner that facilitates reuse.

12.8.2 Benefits and Example

Since there are many different ways of processing and analyzing data (Silberzahn et al., 2018; Steegen et al., 2016), sharing code promotes reproducibility and encourages sensitivity analyses. Sharing data and code also allows other researchers to establish the validity of the original analyses, it can facilitate collaboration, but it can also serve as protection against data loss. When publishing his theory on “general intelligence”, Spearman (1904) shared his data as an appendix to the article. A century later, this act of foresight enabled scientists to use this data set for both research and education. Because Spearman made his data publicly available, other researchers could establish the reproducibility and generalizability of the findings.

12.8.3 Current Status

Data sharing has never been easier. Public repositories offer free storage space for research materials, data (e.g., the Open Science Framework), and code (e.g., Github). While data sharing is not yet a general practice in most scientific fields, several recent initiatives (e.g., Open Data/Code/Materials badges, Kidwell et al., 2016), standards (TOP Guidelines, Nosek et al., 2015), journals (e.g., Scientific Data) and checklists (e.g., Transparency Checklist, Aczel, Szaszi, et al., 2020) are helping to promote this research practice. When sharing raw data is unfeasible, researchers can make aggregated data summaries available, for example, the data used to generate certain plots or covariance matrices of involved variables.

12.8.4 Limitations

Restrictions imposed by funders, ethics review boards in universities and other institutions, collaborators, and legal contracts may limit the extent to which data can be publicly shared. There may also be practical considerations (e.g., sharing big data), data use agreements, privacy rights, and institutional policies that can curtail sharing intentions. What remains central is to inform the readers about the accessibility of the data of the analysis. It should be noted that these limitations should not apply to the analysis code as long as code is solely reflective of the researcher’s analysis actions and is free of any data privacy issues.

12.8.5 Guidelines

An important principle of sharing data is that they should be Findable, Accessible, Interoperable, and Reusable (FAIR, M. D. Wilkinson et al., 2016). Several guides are available discussing the practical (e.g., O. Klein et al., 2018) and ethical (e.g., Alter & Gonzalez, 2018) aspects of data sharing. Researchers should follow the data sharing procedures and requirements of their fields (e.g., Taichman et al., 2017; Wagenmakers, Kucharsky, & the JASP Team, 2020) and indicate the accessibility of the data in the research report (Aalbersberg et al., 2018; Nosek et al., 2015). The ASA Ethical GuidelinesCommittee on Professional Ethics of the American Statistical Association, 2018 for Statistical Practice state that the ethical statistician “[p]romotes sharing of data and methods as much as possible”, and "[m]akes documentation suitable for replicate analyses, metadata studies, and other research by qualified investigators." (p. 5).

12.9 Concluding Comments

If the statistical literature is any guide, one may conclude that statisticians rarely agree with one another. For instance, the 2019 special issue in *The American Statistician* featured 43 articles on p-values, and in their editorial Wasserstein et al. (2019) stated that "the voices in the 43 papers in this issue do not sing as one". However, despite the continuing disagreements about the foundations of statistical inference, we believe there is nevertheless much common ground among statisticians, specifically with respect to the ethical aspects of their profession. To explore this ethical dimension more systematically, we started by considering the Mertonian norms that characterize the ethos of science and outlined a non-exhaustive list of seven concrete, teachable, and implementable practices that we believe need wider propagation.

In essence, these practices are about promoting transparency and the open acknowledgement of uncertainty. With agreement on such practices explicitly acknowledged, we believe that commonly discussed contentious issues (e.g., p-values) may become less crucial. Indeed, in a letter to his frequentist nemesis Sir Ronald Fisher, the arch-Bayesian Sir Harold Jeffreys wrote "Your letter confirms my previous impression that it would only be once in a blue moon that we would disagree about the inference to be drawn in any particular case, and that in the exceptional cases we would both be a bit doubtful" (J. H. Bennett, 1990, p. 162).

We hope that the proposed statistical practices will improve the quality of data analysis across the board, especially in applied disciplines that are perhaps unfamiliar with the ethical aspects of statistics, aspects that a statistician may take for granted. Also, instead of counting on them to be absorbed through osmosis, we believe it is important to include these ethical considerations –and their statistical consequences– explicitly in the statistics curricula. Statistical techniques other than those discussed here may also further the Mertonian ideals. We hope that this contribution provides the impetus for a deeper exploration of how data analysis in applied fields can become more transparent, more informative, and more open about the uncertainties that inevitably arise in any statistical data analysis problem.

Chapter 13

Teaching Good Research Practices: Protocol of a Research Master Course

Abstract

The current crisis of confidence in psychological science has spurred on field-wide reforms to enhance transparency, reproducibility, and replicability. To solidify these reforms within the scientific community, student courses on open science practices are essential. Here we describe the content of our Research Master course "Good Research Practices" which we have designed and taught at the University of Amsterdam. Supported by Chambers' recent book *The 7 Deadly Sins of Psychology*, the course covered topics such as QRPs, the importance of direct and conceptual replication studies, preregistration, and the public sharing of data, code, and analysis plans. We adopted a pedagogical approach that (1) reduced teacher-centered lectures to a minimum; (2) emphasized practical training on open science practices; (3) encouraged students to engage in the ongoing discussions in the open science community on social media platforms.

This chapter is published as Sarafoglou, A., Hoogeveen, S., Matzke, D., & Wagenmakers, E. J. (2020). Teaching good research practices: Protocol of a research master course. *Psychology Learning & Teaching*, 19, 46–59.

13.1 Introduction

Over the last eight years, psychological research has been in the midst of a "crisis of confidence" (e.g., Pashler & Wagenmakers, 2012; Simmons et al., 2011). Central to the crisis is the increasing realization that common research practices may in fact be deeply problematic. Examples include poor study design (i.e., low statistical power; Button et al., 2013; Ioannidis, 2005), the field's reluctance to conduct direct replication studies (Pashler & Harris, 2012; Schmidt, 2009), and a bias to selectively report positive results (Francis, 2013; Scargle, 1999). Moreover, many researchers self-admit to the use of so-called questionable research practices (QRPs; John et al., 2012). Hidden from the reader, QRPs exploit researchers' degrees of freedom in study design and analysis in order to produce significant findings. For instance, researchers may decide to stop data collection when the result reaches significance, exclude data points based on their impact on the results, or report unexpected findings as having been predicted from the start (Kerr, 1998; Simmons et al., 2011). The detrimental effect of these practices is evident from recent surveys and large-scale replication projects. For instance, a survey among over 1,500 scientists revealed that 90% believe there is indeed a crisis, with 52% observing a "significant crisis" (Baker, 2016). These perceptions are substantiated by large-scale replication efforts, which demonstrated replication rates from 36% to 77% (Camerer et al., 2018; R. Klein, Ratliff, Vianello, Adams, et al., 2014; R. Klein et al., 2018; Open Science Collaboration, 2015).

To combat the crisis of confidence, the scientific community has begun to adopt research standards that reduce cherry-picking and significance chasing. For instance, an effective practice that has quickly gained popularity is *preregistration.* When preregistering their studies, researchers outline their analysis plan before the data are collected. Because the analysis pipeline cannot be tailored to the data, researchers protect themselves against hindsight bias and other QRPs that may unwittingly contaminate the results. Researchers can choose to preregister their study either independently or integrate preregistration with the peer-review process (i.e., in the form of a Registered Report; Chambers, 2013; Nosek & Lakens, 2014). In addition, the scientific community has launched various initiatives to increase transparency. For instance, to encourage data sharing, Morey et al. (2016) started the peer reviewers openness (PRO) initiative. PRO signatories agree to provide a full review only for articles that share data and materials in a public repository, or provide reasons why this is not possible. Journals have also promoted transparency standards, for instance by signing on to the Transparency and Openness Promotion guidelines (TOP; Nosek et al., 2015), or by providing open science badges for preregistration and sharing of data and materials (Kidwell et al., 2016). Open science advocates have argued that the methodological reforms within the scientific community have been so substantial that warrant descriptions such as "Revolution 2.0" (Spellman, 2015) or "Credibility Revolution" (Vazire, 2018).

In addition to the reforms within the research community itself, researchers have emphasized the need to overhaul methodological education. For instance, in the survey by Baker (2016), three of the five factors considered most promising for increasing the reproducibility in science were directly related to improvements in scientific training (i.e.,"better statistical understanding", "better mentoring/-supervision", and "better teaching"). Central among the proposed changes are offering lectures on the crisis of confidence and open scientific practices (Chopik, Bremner, Defever, & Keller, 2018; Funder et al., 2014; Munafò et al., 2017).

We believe that a course on good research practices deserves a place in the standard psychology curriculum, and that open scientific practices should be an inherent part of the methodological training of students for several reasons. First, without the proper education, students' opinions on the crisis of confidence tend to be "quite radical, superficial, or even emotional" (Chopik et al., 2018, p. 159). Educating students about the ongoing methodological changes allows them to develop informed opinions on these topics. Second, when students –the next generation of scientists– understand open science practices, they can confidently introduce them in their future labs. Third, students who pursue an academic career, will ultimately be evaluated on whether they adhere to these practices. As journals and university policies are making increasing demands on transparency criteria, educating students about these practices seems advisable if not imperative. Lastly, regardless of students' future career plans, advancing the methodological curriculum also benefits the students' development at a more general level. By following a course on good scientific practices, students learn to recognize scientific studies that meet certain quality standards reflected by, for instance, being preregistered, having open materials and data, being published as a Registered Report, including a power analysis, or reporting effect sizes. As such, a course on open science enhances students' skills to critically evaluate research, be it from the published literature or conducted by themselves, for instance as part of a thesis requirement.

Since 2015 we offer the open science course *Good Research Practices* at the University of Amsterdam. The course covers the current crisis of confidence in psychological science and outlines attempts by the scientific community to increase the reliability and transparency in the field. *Good Research Practices* is a Research Master course; students generally know basic statistics and have had practical experience with the empirical cycle. This background makes it easier to understand the challenges and advantages of implementing open science practices. Nevertheless, the course is not technical in nature and mostly demands common sense –hence, the material may also be useful for a course for undergraduate students.

In this chapter, we aim to provide an overview of our *Good Research Practices* in order to assist lecturers who intend to develop a similar course. Below we discuss the course objectives, describe our pedagogical approach, and illustrate the contents of two classes in more detail. Furthermore, we will list the lecture topics together with suggested literature for students. Readers interested in the full course catalogue and materials can access it in our online appendix (accessible via `https://osf.io/v3z7q/`).[1]

[1] Interested readers may also find the collection of teaching materials for university courses by Davis et al. (2017) helpful which can be accessed via `https://osf.io/f82ej/`, or the Massive Open Online Courses compiled by the Open Science MOOC Team (2019) which are available at `https://opensciencemooc.eu`. These teaching materials are specifically designed to educate psychology students on open science, reproducibility, and replicability.

13.2 General Information

Good Research Practices is designed as a seven-week course including a total of 14 two-hour classes. A total of 43 Research Master Psychology students at the University of Amsterdam participated in the course last year, for which they were awarded six ECTS credits after completion (equivalent to 180 hours of work). Grading was based on a combination of bi-weekly quizzes about the background literature, and on the quality of their short presentations and in-class assignments.

13.3 Course Objectives

In general, a course on good research practices should teach students how to critically review the scientific literature and how to conduct open, transparent, and reliable research. In addition, we wanted to immerse students in current debates and recent developments in the open science community. Specifically, our course had four objectives, discussed below.

Our first objective was for students to reflect on various types of questionable research practices. In particular, we emphasize that researchers are not immune to biases (e.g., hindsight bias and confirmation bias) that cause them to selectively report analyses that yield publishable findings. To protect themselves against their own biases, researchers must rely on scientific practices that minimize hidden degrees of freedom (Wagenmakers et al., 2012). As primary course literature we used the recent book "The 7 Deadly Sins of Psychology" by Chambers (2017), which presents a clearly written, authoritative, and comprehensive account of the causes and proposed solutions for the current crisis in psychological science.

Our second objective was to engage students in current debates and recent developments in the open science community. Social media platforms constitute a prominent stage for science communication and debates on research-methods reforms. These platforms include Twitter, scientific blogs, and podcasts. As part of the curriculum, we encouraged students to stay informed about ongoing discussions and new developments within the open science movement, and educate their peers in weekly "Newsflashes" on interesting debates, articles, or events. A list with Twitter handles, Podcasts, and scientific blogs we recommended is available via `https://osf.io/mcqa5/`.[2]

Our third objective was to let our students contribute to the curriculum themselves. We believe that students learn more when they are stimulated to actively participate in the course (e.g., Jang, Reeve, & Halusic, 2016; Reeve, 2016). Therefore, we adopted a flipped-classroom setting to reduce teacher-centered classes to a minimum. In this setting, students give short lectures, design in-class assignments, and lead group discussions.

Our fourth objective was to provide multiple perspectives on the open science movement. Therefore, we invited a series of guest speakers to present their most recent research projects, their perspectives on the developments within the scien-

[2] It should be noted that during this exercise most students took part as passive observers, that is, they were instructed to follow the discussions but were not urged to participate in the debates.

tific community, and their opinions on possible ways to resolve the crisis. Since the course was designed to illustrate the necessity and benefits of open science, we exclusively invited proponents of the open science movement. At the same time, we tried to select speakers who differ in their level of seniority, and who approach methodological reforms from different angles. In the current installment of the course, the guest speakers included a former student from the Research Master program (Bobby Houtkoop), a science journalist (Hans van Maanen), metascience researchers (Balazs Aczel, Nick Brown, and Olmo van den Akker), Chris Chambers, who is the chair of the Registered Reports committee at the Center for Open Science (`https://cos.io/`) and leading force within the open science movement.

13.4 Pedagogical Approach

In line with our course objectives, we alternated regular classes with classes organized by students. Lectures in regular classes were given either by us or one of our guest speakers, and focused on the substantial impact that QRPs may have on the reliability of research findings. In particular, we explained why certain research practices can be considered "bad science" (Goldacre, 2009), and how such practices can be detected, and –importantly– avoided. The classes also featured specific in-class assignments and group discussions to deepen students' understanding. In addition, the regular classes covered recent developments and debates within the open science movement; specifically, we reserved the last twenty minutes of each regular class for a "Newsflash" item, where students gave lighting presentations about relevant events, discussions, or articles they encountered on social media platforms that week. It should be noted that discussions following the lightning presentations were led by one of the lecturers who could provide context and insight about the presented topics. These guided discussions are recommended, since students might not be aware that they are exposed to only a selective group of people who typically dominate these debates, and who may not be representative of the entire scientific community.

Classes given by students were structurally similar to regular classes. However, at about 10 minutes each, the student lectures were much shorter than regular lectures, leaving considerable time for active learning during the in-class assignments. Shorter student lectures also allowed us to have multiple groups present each week.

To encourage creativity and originality, students were instructed to base their lectures on relevant topics that had not already been elaborately discussed in their assigned readings. With respect to the in-class assignments, we emphasized that the exercises should have practical value for their peers, that is, the exercises should be training material for open science practices.[3] Examples of this year's in-class assignments are: tutorials on how to preregister a study or share data on the Open Science Framework (`https://osf.io`), trying out software tools that examine possible anomalies in individual articles (e.g., `statcheck`; Epskamp & Nuijten, 2016, or `SPRITE`; Heathers, Anaya, van der Zee, & Brown, 2018), or

[3] Without this instruction it is our experience that students tend to create in-class assignments that simply demonstrate a QRP (e.g., a frequently suggested exercise is to let students p-hack a data set to obtain a significant result).

detecting hidden analytic flexibility in entire research fields (e.g., with a p-curve analysis as proposed by Simonsohn, Nelson, & Simmons, 2014). To illustrate our pedagogical approach, the next two sections describe a regular class and a student-organized class.

13.4.1 Example of a Regular Class: The Sin of Data Hoarding

The fifth week of the course focused on "The Sin of Data Hoarding" Chambers (2017), that is, the chapter on data sharing (for a recent special issue see D. Simons, 2018). As an expert on this topic we invited Bobby Lee Houtkoop, a former student from the same program. Houtkoop recently conducted and published a survey study to reveal reasons why researchers are reluctant to share their data, and what can be done to overcome this reluctance (Houtkoop et al., 2018). In her lecture, Houtkoop discussed the dominant scientific culture in which data sharing is not the norm, even though data sharing offers unequivocal advantages for both the author and the scientific community. In cancer research, for instance, it was found that studies for which data were publicly shared received higher citation rates compared to studies for which data were not available (Piwowar, Day, & Fridsma, 2007). In addition, data sharing may improve the reputation or perceived integrity of the researcher. The scientific community benefits from data sharing since (1) it increases the longevity of the data, (2) data can be reanalyzed and reused efficiently (e.g., for meta–analyses), and (3) statistical or reporting errors are more likely to be found (Vanpaemel, Vermorgen, Deriemaecker, & Storms, 2015; J. Wicherts, Borsboom, Kats, & Molenaar, 2006). Houtkoop then presented the methods and results of the survey study. The survey results demonstrated that data are shared only infrequently. Most respondents acknowledged the benefits and importance of data sharing in general; however, they perceived data sharing as less beneficial for their own research projects. Among the perceived barriers to data sharing are the respondents' belief that data sharing is not a common practice in their fields, their preference to share data only upon request, their perception that data sharing requires additional work, and their perceived lack of training in data sharing. Houtkoops study sparked a lively discussion among the students about future research, about initiatives that encourage data sharing, but also about limitations of the study. In particular, the students were critical about potential biases in the results due the low response rate of the survey (i.e., a response rate of only about 5% which, however, translated into a sample of 600 respondents) and the self-selection of the respondents.

The end of the class featured a "Newsflash". In that particular week, the science community was excitedly debating the results of the "Many Labs 2" project (R. Klein et al., 2018) which had just been published. In this project, the participating research teams conducted high powered preregistered replications of 28 classic and contemporary findings across many samples and settings. The replication efforts showed that only 54% (i.e., 15 studies) could be replicated. In the newsflash, students discussed the article by R. Klein et al. (2018), the related news article published in *The Atlantic* titled "Psychology's Replication Crisis Is Running Out Of Excuses" (Yong, 2018), and the BBC radio episode on the replication crisis (BBC Radio 4, 2018).

13.4.2 Example of a Student Class: The Sin of Data Hoarding

The student lecture continued where Houtkoop's study left off. The student presenters emphasized the benefits of data sharing and created a tutorial for their peers on how to archive and share data of simple empirical studies on the Open Science Framework (see also Soderberg, 2018). The objective of this lecture was to encourage their peers to ask their future thesis supervisors permission to share the collected data in a public repository. The in-class assignment revolved around the Peer Reviewer's Openness initiative (PRO; Morey et al., 2016) mentioned in the introduction. Specifically, the students let their peers create a set of questions for the signatories of the PRO initiative, inquiring about signatories' post-PRO experiences with journals and editors, their attitude towards data sharing in their own research, and whether and how the signatories would improve the initiative. Students were divided into small groups and were instructed to read the article by Morey et al. (2016) on the PRO initiative. Then, each group had to propose concrete questions for the PRO signatories. In a plenary discussion, the students reviewed the questions, selected the ones they found most relevant, and created a survey. Since this exercise generated items that seemed informative and useful, the students who prepared the class decided to continue and execute the survey as a separate research project. Currently, the PRO initiative survey has elicited responses from over 120 of the current 340 signatories for whom Email information could be retrieved (i.e., 37.4%).

13.5 Topics Covered

Table 13.1 lists the topics covered in the lectures, including the guest lectures. The table also contains pointers for students to the relevant literature. Most topics follow the chapters of Chambers (2017); however, we added topics that we deemed relevant in the current research debate. For instance, we dedicated one lecture to the recently published and much debated article by Benjamin et al. (2018) who proposed a more stringent significance threshold for new scientific discoveries. We also discussed analysis blinding–a promising and underused method that allows researchers maximum flexibility while preserving the confirmatory status of the analyses.

Table 13.1: Topics Covered and Suggested Literature for the Course "Good Research Practices".

Topic	**Description**
The crisis of confidence	In two classes we covered some of the main events that led to the crisis of confidence: multiple instances of scientific fraud, the wide acceptance of QRPs among researchers, and the preference of journals to publish novel and positive findings. *Suggested Literature:* Pashler & Harris, 2012; Spellman et al., 2018.

Topic	**Description**
Biases in scientific research	This class covered cognitive biases, such as confirmation bias and hindsight bias, that lead researchers to unwittingly present unexpected findings in their data as if they were hypothesized from the beginning. *Suggested Literature:* Chambers, 2017, Chapter 1.
Lack of transparency (with Balazs Aczel)	In this class we argued that whenever part of the scientific process remains hidden from view, the trustworthiness of the associated conclusions is eroded, since QRPs cannot be detected. To combat this issue, researchers will be able to use a transparency checklist (which is nearing completion) that facilitates the disclosure of the transparency and openness-related factors of their study. This lecture was given by Balazs Aczel, the leading researcher of this project. *Suggested Literature:* De Groot, 1956/2014; Simmons et al., 2011
Hidden flexibility in data analysis	In this class we stressed the point that the reliability of research findings is ensured only when researchers adhere to the empirical cycle. Specifically, we argued that if researchers do not strictly separate between the stage of hypothesis generation and the stage of hypothesis testing, the predictive interpretation is lost. *Suggested Literature:* Chambers, 2017, Chapter 2.
Blinded analyses	In this class we discussed analysis blinding as a valuable addition to study preregistration to avoid hidden flexibility in data analysis. Analysis blinding, just as preregistration, prevents implicit or explicit forms of significance-chasing, but it retains the possibility for the data analyst to account for unexpected features of the data. *Suggested Literature:* Dutilh, Sarafoglou, & Wagenmakers, 2019; MacCoun & Perlmutter, 2015; MacCoun & Perlmutter, 2018.
Unreliability of scientific findings	Science depends on direct replications of scientific studies to determine the validity of alleged effects. In this lecture we discussed recent large-scale replication efforts and the impact they had on psychological science. *Suggested Literature:* Chambers, 2017, Chapter 3.
Data hoarding (with Bobby Lee Houtkoop)	This class covered the importance of data sharing and discussed reasons why researchers are still reluctant to share their data. This lecture was given by Bobby Lee Houtkoop, the leading researcher of a recently published survey study that identified these perceived barriers and possible remedial action. *Suggested Literature:* Chambers, 2017, Chapter 4; Houtkoop et al., 2018.

Topic	Description
Scientific fraud (with Nick Brown)	In this class we discussed how to detect anomalies in research articles, for instance, by reconstructing plausible samples from descriptive statistics. This lecture was given by Nick Brown, who was involved in the development of these techniques. *Suggested Literature:* Brown & Heathers, 2017; Heathers et al., 2018; Levelt, Drenth, & Noort, 2012; Stapel, 2014.
Overselling scientific findings (with Hans van Maanen)	In this class we discussed how the over-generalization or exaggeration of study conclusions in abstracts and press releases distort the representation of scientific findings in the media. As an expert on this topic we invited science journalist Hans van Maanen, who is known for his columns in the Dutch newspaper *De Volkskrant* in which he eviscerates published research. *Suggested Literature:* Chambers, 2017, Chapter 5.
Redefining statistical significance	In this class we discussed the recently published paper by Benjamin et al. (2018) in which the authors propose to lower the α-levels for claims of new discoveries from 0.05 to 0.005. *Suggested Literature:* Benjamin et al., 2018 *Blogpost articles:* Wagenmakers, 2019, Redefine Statistical Significance (Parts I– XVII) *YouTube debate on Benjamin et al. (2017):* BITSS, 2017.
Statistical errors (with Olmo van den Akker)	Statistical reporting errors can lead to erroneous substantive conclusions. In this class we discussed how researchers can minimize the chance of statistical reporting errors by using software that automatically detects inconsistencies. This lecture was given by Olmo van den Akker, who is part of the Meta-Research Center at Tilburg University that is specialized in scientific misconduct and reproducibiity. *Suggested Literature:* Chambers, 2017, Chapter 6; Epskamp & Nuijten, 2016; Greenland et al., 2016; Nuijten et al., 2016.
Registered Reports (with Chris Chambers)	Apart from publishing the course textbook “The 7 Deadly Sins of Psychology”, Chambers has participated in drafting the TOP guidelines and is the chair of the Registered Reports committee supported by the Center for Open Science. In his class, Chris Chambers shared his experiences of how he first proposed the Registered Report format to the *Cortex* editorial board, how the initiative was implemented in the journal, and how Registered Reports are having a growing influence on the scientific community. *Suggested Literature:* Chambers, 2017, Chapter 8; Chambers, 2013.

13.6 Student Evaluation and Recommendations for Future Courses

Student feedback was highly positive. Students particularly appreciated (1) the guest lectures; (2) the group discussions about ongoing debates and recent articles; (3) the assigned literature (i.e., the course book and the additional articles), which was perceived as relevant and enjoyable; and (4) the teaching of important practical skills. The perceived work load was deemed appropriate, and students liked the fact that the course was designed to encourage regular work through quizzes and assignments.

Students were most critical about our emphasis on negative facets during regular classes, that is, QRPs and the crisis of confidence. Some students stated that discussing these aspects so frequently made them pessimistic about the current state of science. Furthermore, the students felt the 2-hour classes were too short. In particular, students were disappointed that often only one group rather than two groups (as anticipated) could present during the student classes. This lack of time also repeatedly forced us to skip the weekly "Newsflashes".

We believe the student feedback is constructive and helpful. We agree with the students that scheduling an additional hour for each class will reduce the time pressure. With regard to the focus on negative facets, we believe that the recognition of QRPs and "bad science" (Goldacre, 2009) is essential to motivate the methodological reorientation towards more transparency and rigor; on the other hand, our main objective was to inspire students to embrace open research practices, not to instill a sense of despair. As nicely put by Michèle Nuijten (2019), we want to "turn students into skeptics not cynics". Therefore, the next installment of our course will devote a larger proportion of time to the positive changes within the scientific community. For instance, we suggest to reconstruct the lecture "Unreliability of Scientific findings". During this lecture, we focused mainly on the importance of conducting direct replications to determine the validity of alleged effects, and emphasized the lack thereof in the scientific literature. However, this lecture offers the opportunity to highlight recent large-scale replication efforts and multi-lab collaborations, such as the Open Science Collaboration (2015), the Many Labs projects (R. Klein, Ratliff, Vianello, Adams, et al., 2014; R. Klein et al., 2018), the ManyBabies project (Frank et al., 2017), and the Psychological Science Accelerator (Moshontz et al., 2018). In addition to a lecture which gives students a general overview on these collaborative efforts, it would be particularly interesting to invite a guest speaker who participated in one of these collaborations to share his or her experiences in working and publishing in such an environment.

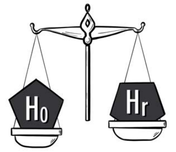
Additionally, we would like to replace the lecture "Scientific Fraud" by a lecture on "Open Science within the University of Amsterdam" to educate our students on the concrete steps our university has taken to improve reproducibility, transparency, and openness. For instance, the ethical committee of the psychology department demands a detailed methods and analysis plan as precondition to grant ethical approval for any research project; similarly, students are requested to write their introduction, methods, and analysis plan of their internship and thesis projects before data collection. Additionally, we would like to highlight the methodological and statistical consulting which is offered to both researchers and students, as well as several open science initiatives that were launched recently.[4]

13.7 Concluding Remarks

Across 14 lectures, the course *Good Research Practices* taught psychology students about the causes of the crisis of confidence and about recent attempts by the scientific community to increase transparency, reproducibility, and replicability. In addition, students acquired practical skills on how to conduct research that is open, transparent, and reliable. We believe that this learning success was primarily due to the active role we gave students in our course. By being instructed to create lectures and in-class assignments that go beyond the assigned literature, students were able to choose articles covering topics that they consider most relevant for their future research projects. Furthermore, the students developed a sense of ownership for the lectures and in-class assignments, which facilitated ambitious student projects such as the PRO initiative survey.

As the scientific culture changes, practical knowledge on open scientific practices is becoming an increasingly important scientific skill. A course on this topic helps students not only to develop critical thinking, but also to get excited about conducting research that distinguishes sharply between its exploratory and confirmatory components. We hope that courses on open science practices inspire the future generation of psychological researchers to deliver psychology from the deadly sins that have so stained it in the past.

[4] This lecture can be adapted to the specific situation of the university in which this course is offered. For instance, if open science policies are still absent in the university, lecturers can highlight promising initiatives in other universities, recently enacted journal policies (i.e., TOP guidelines), or open science policies that are advanced on country level (e.g., the National Institutes of Health Public Access Plan in the United States; `https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf`, or the open-access science publishing initiative Plan S in the European Union; `https://www.coalition-s.org/`).



Chapter 14

Combine Statistical Thinking with Open Scientific Practice: A Protocol of a Bayesian Research Project

Abstract

Current developments in the statistics community suggest that modern statistics education should be structured holistically, that is, by allowing students to work with real data and to answer concrete statistical questions, but also by educating them about alternative frameworks, such as Bayesian inference. In this chapter, we describe how we incorporated such a holistic structure in a Bayesian research project on ordered binomial probabilities. The project was conducted with a group of three undergraduate psychology students who had basic knowledge of Bayesian statistics and programming, but lacked formal mathematical training. The research project aimed to (1) convey the basic mathematical concepts of Bayesian inference; (2) have students experience the entire empirical cycle including collection, analysis, and interpretation of data and (3) teach students open science practices.

This chapter is published as Sarafoglou, A., van der Heijden, A., Draws, T., Cornelisse, J., Wagenmakers, E. J., & Marsman, M. (2022). Combine Statistical Thinking With Open Scientific Practice: A Protocol of a Bayesian Research Project. *Psychology Learning & Teaching*, 147572572210773.

14.1 Introduction

The curriculum guidelines of the American Statistical Association (ASA) argue that statistics education in undergraduate programs should not be primarily focused on teaching statistical methods and mathematical foundations, but also emphasize scientific practice, that is, study design, data collection, programming skills, and data analysis (American Statistical Association, 2014; Horton & Hardin, 2015; Wasserstein & Lazar, 2016). In general, students should learn to "think with and about data" (Cobb, 2015, p. 267) and thus develop a holistic understanding of statistics (Horton & Hardin, 2015).

This holistic understanding of statistics also includes learning and understanding alternatives to classical inference based on *p*-values. Bayesian inference is becoming increasingly popular and its adoption has been advocated for both scientific practice (Wasserstein & Lazar, 2016) and statistics education (Cobb, 2015). Recent examples for undergraduate courses and in-class demonstrations on Bayesian methods that require only little or no mathematical or statistical training are described in Witmer (2017; on teaching Markov chain Monte Carlo methods), Rouder and Morey (2018; on teaching Bayes' rule), and van Doorn, Matzke and Wagenmakers (2020; on teaching the key concepts of Bayesian inference). However, little attention has been paid to the design and structure of Bayesian research projects that can be conducted with a small group of students, for instance, in the context of a thesis or internship project, or a seminar. These formats, as opposed to standard courses, allow for more extensive research projects, since supervisors can offer individual support for students, and can dedicate more time to the execution of the project.

We believe that a research project on Bayesian inference should take advantage of the rather long project duration and the small group size by introducing students in detail to the theoretical and practical aspects of Bayesian inference. Theoretical aspects of Bayesian inference entail that by the end of the project students should feel comfortable with the standard terminology, be able to understand how to assign a prior distribution, specify a likelihood function, derive a posterior distribution, and compute a marginal likelihood. The practical aspects entail that students should be able to apply their theoretical knowledge to address a concrete research question, and experience the entire empirical cycle, including study planning, preregistration, data collection and analysis, and interpretation of the results. These teaching goals resulted in three guiding principles for structuring the project, listed below.

The first principle is to introduce students to the mathematics underlying Bayesian statistics. In our own teaching of Bayesian methods in undergraduate psychology courses, we usually hide the mathematics and instead aim to provide students with an intuition about how Bayesians use distributions to quantify uncertainty about model parameters and hypotheses. This approach helps students interpret posterior distributions, credible intervals, and Bayes factors (for a gentle technical introduction to Bayesian inference without mathematical derivations see Etz & Vandekerckhove, 2018). However, for students who want to specialize in research methods and statistics it is important to go beyond an intuitive understanding and be introduced to the mathematics behind these key concepts.

Without the mathematical foundations, students will find the statistical literature difficult to understand.

The second principle is to let students experience scientific practice. In line with the ASA guidelines on statistics education (American Statistical Association, 2014), we believe that students learn most when they are given the opportunity to gain hands-on experience on how to apply the methods taught to a real data example. We therefore set up a Bayesian replication study that demonstrates a series of Bayesian benefits. For instance, in contrast to frequentist analyses, the Bayesian framework allows students (1) to discriminate between "absence of evidence" and "evidence of absence" of the effect in the replication study (Dienes, 2014; Keysers, Gazzola, & Wagenmakers, 2020; Verhagen & Wagenmakers, 2014); (2) to experience Bayesian learning by incorporating prior knowledge –such as data from previous experiments– to construct a more informative test (Verhagen & Wagenmakers, 2014); (3) to monitor evidence as the data accumulates (Rouder, 2014b). In addition, it allows students to learn how conclusions from significant p-values differ from conclusions drawn from Bayes factors by conducting a Bayesian reanalysis of the results of the original experiment.[1]

The third and final principle is to convey open science practices. Reproducibility and replicability are core scientific values, but yet psychological science is currently facing a crisis of confidence as a disappointing proportion of key findings appear to be reproducible (Baker, 2016; Camerer et al., 2018; R. Klein, Ratliff, Vianello, Adams, et al., 2014; R. Klein et al., 2018; Nature Publishing Group, 2016; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). To a large extent, the low rate of reproducible findings can be attributed to the great flexibility in data analysis in combination with selective reporting of significant results (Simmons et al., 2011), the high prevalence of questionable research practices (John et al., 2012), the reluctance to conduct direct replication studies (Pashler & Harris, 2012; Schmidt, 2009), and the poor availability of research data (Houtkoop et al., 2018; J. Wicherts et al., 2006; for a special issue on data sharing see D. Simons, 2018). To address these problems, psychological science today relies on numerous open scientific practices, such as preregistration and Registered Reports, large-scale collaborations, and sharing of data, materials, and code (e.g., Chambers, 2013; Chambers & Tzavella, 2021; Kidwell et al., 2016; Morey et al., 2016; Moshontz et al., 2018; Nosek et al., 2015). However, to truly integrate these practices into the research culture, it is necessary to introduce the principles of open science to students at an early stage (Chopik et al., 2018; Funder et al., 2014; Morling & Calin-Jageman, 2020; Munafò et al., 2017; Sarafoglou, Hoogeveen, Matzke, & Wagenmakers, 2020). Since thesis projects often require detailed design and analysis plans we view them as a good opportunity for supervisors to teach them both the philosophy behind open science and the practical skills needed to apply open science practices. Therefore, we set up a preregistered replication study, and have students publish the analysis code, and share the data and materials on the Open Science Framework (OSF; Center for Open Science, 2021).

The purpose of this chapter is to share our experiences on designing and super-

[1] We refer the interested reader to Wagenmakers (2007); Wagenmakers, Marsman, et al. (2018) for a more detailed discussion on the benefits of Bayesian inference.

vising a Bayesian thesis project for undergraduate psychology students. Lecturers who intend to offer a Bayesian research project for a small group of students which emphasizes mathematical training as well as practical experience with real data might find helpful advice on what focal points to set when planning their project. In addition, the described project can serve as illustrative example in a classroom setting, to teach students Bayesian learning, and a simple method to evaluate ordinal expectations. In the following, we will describe the course structure, the theoretical and the practical part of the project in more detail.

14.2 Supplemental Material

Interested readers can visit our OSF project folder (https://osf.io/zfhbc/) to access the following information: the study preregistration, the analysis code, all data and materials, and the student evaluations. Furthermore, it contains the results of the Bayesian reanalysis of the original studies and the formal description of the mathematical model for multiple independent binomial probabilities.

14.3 Project Overview

Here we describe the thesis project titled "A Bayesian View on 'Science versus the Stars': Bayes factor analysis for ordered binomial probabilities" at the University of Amsterdam. The topic of the thesis project was the Bayesian analysis of ordinal expectations of multiple binomial probabilities. We chose this topic due to both its relevance in the psychological literature and the simplicity of the statistical model. Ordinal expectations of binomial probabilities are common in the area of psychometrics and theories on rational decision making (see e.g., Cavagnaro & Davis-Stober, 2014; Davis-Stober, 2009; Guo & Regenwetter, 2014; Haaf et al., 2020; Heck & Davis-Stober, 2019; J. Myung et al., 2005; Regenwetter et al., 2018, 2011; Tijmstra et al., 2015). For instance, a psychometrician who evaluates whether a test for cognitive performance can be measured on an interval scale needs to test the assumption that the probability to solve a given item is non-decreasing for the ability of a person. One argument to use Bayesian methods for these problems is that we can easily incorporate ordinal expectations of the binomial probabilities in the respective prior distributions (Klugkist et al., 2010). This makes the corresponding statistical model particularly simple and enables students to derive the method even without formal mathematical training.

During the theoretical part of the project students familiarized themselves with the computation of Bayes factors for ordered binomial probabilities using the encompassing prior method (Klugkist et al., 2005). During the practical part of the project the students applied the methods in practice by conducting a preregistered reanalysis and replication study.

14.3.1 Course Structure

The full thesis project–starting from the first introductory lesson to submission of the research report–took 16 weeks. A weekly overview of the research project is provided in Table 14.1. Our students had to hand-in two writing assignments, create the preregistration of the empirical study, and write the final report. On average the students worked 22−23 hours per week on the project for which they were rewarded with 12 ECTS credits. The following section describes these components in more detail.

Table 14.1: *A week-by-week overview of our project "A Bayesian view on science versus the stars: Bayes factor analysis for ordered binomial probabilites".*

Week	**Goal**	**Activities**
1	Reiterating knowledge	Bayesian parameter estimation and hypothesis testing for the beta-binomial model Write methods section of research report
2	Establishing knowledge	Generalize concepts to multiple binomials Write methods section of research report
3	Establishing knowledge	Derive and apply Savage-Dickey density ratio Write methods section of research report
4	Establishing knowledge	Derive and apply encompassing prior approach Bayesian reanalysis of Carlson (1985) and Wyman and Vyse (2008) Write introduction of research report
5	Writing	Finalize the methods section of the research report; Write introduction of research report
6–7	Preregister study	Plan replication study Create preregistration document
8	Preregister study	Print all necessary documents, prepare data collection (e.g, book lab) Finalize preregistration
9–10	Data collection 1	Participants fill out NEO-FFI and report date and place of birth
11	Create study materials	Generate personality descriptions Prepare follow-up data collection
12–13	Data collection 2	Participants perform choice task
14	Analyzing data	Analyze data and upload the dataset to the OSF Write results section of research report
15–16	Finalizing project	Finalize research report Prepare 20-minute presentation

14.3.2 Supervision

During the theoretical part of the project we supervised the students intensively; we had weekly group meetings that were structured as lectures, we gave students two writing assignments, and we reviewed and discussed these assignments with each student individually. During the practical part, the students then primarily worked independently with little need for supervision. The weekly group meetings were replaced by individual contact hours that gave students the opportunity to discuss details of their report.

14.3.3 Writing Assignments

We dedicated the first four sessions at the beginning of the project to the theoretical concepts of Bayesian inference. During these four weeks, students wrote two short 1−2 page reports explaining Bayesian parameter estimation and hypothesis testing. The first report concerned Bayesian inference for one binomial probability. In the second report, they had to generalize these concepts to multiple binomial probabilities. Students could incorporate these reports as part of the methods section in their final report.

14.3.4 Preregistration

Our students had three weeks to create the preregistration document. Since our students answered the same research question, we let them create the preregistration together. The preregistration featured the following components: the study design; the sampling plan, sampling plan rationale, and stopping rule for data collection; exclusion criteria; the description of the materials and procedure; the research question and hypotheses (including the expected direction of the effect); details on the statistical model and analysis plan, including specifications for prior distributions, number of samples drawn, inference criteria, and handling of missing data.

14.3.5 Grading Criteria

For the most part, grading was based on the individual research reports. We assessed whether students were able to (1) justify the proposed research question and methods; (2) describe the Bayesian concepts accurately by using the specific terminology; (3) discuss and interpret the results correctly; and (4) adopt a scientific writing style. In addition, students could receive a pass or fail both on their final presentation and on their learning progress. The writing assignments and the preregistration were not graded.

14.4 The Theoretical Part: Bayesian Parameter Estimation and Hypothesis Testing for Multiple Binomial Probabilities

The goal for the theoretical part of the project was to teach students when and how the encompassing prior approach is used, and how it is derived. To ease the students into this topic, we asked them to reiterate the basic mathematical concepts in Bayesian inference by means of one binomial success probability, that is, Bayesian parameter estimation (including Bayes' rule, the prior distribution, the likelihood function, marginal likelihood, and the posterior distribution) and

Bayesian hypothesis testing (including prior model odds, the Bayes factor, and posterior model odds). Subsequently, students had to generalize these concepts to multiple binomial success probabilities.

14.5 The Practical Part: Reanalysis and Replication of Wyman and Vyse

We searched for empirical studies which involved hypotheses about the ordering of multiple binomial probabilities. The study by Wyman and Vyse (2008) is a suitable candidate for a replication study, for several reasons. First, the study had an engaging research question, that is, whether the accuracy of psychological personality descriptions is similar to the accuracy of astrological natal charts. Second, the dependent variables in Wyman and Vyses' study allowed for the formulation of an ordinal expectation. Third, replicating the study did not require knowledge about sophisticated concepts such as item response theory (Birnbaum, 1968; Rasch, 1960). Fourth, the experimental setup for the study was straightforward which made the planning and execution of a preregistered replication study feasible for our time frame. Note that the study by Wyman and Vyse is itself a conceptual replication of a study conducted by Carlson (1985). We chose to replicate the study by Wyman and Vyse (2008), however, since the authors had a clearer setup and material that was easier to reproduce. We aimed to replicate the study by Wyman and Vyse (2008) as closely as possible which meant that we adapted the original research design with only a few practical changes.

14.5.1 Methods

14.5.1.1 Sampling plan

We preregistered to collect data from 50 participants or stop data collection by June 1st, 2018. The target sample size was based on the number of participants in the original studies which was 56 participants in Carlson (1985) and 52 participants in Wyman and Vyse (2008). Unfortunately, we were not able to reach the preregistered target sample before our testing period ended. We were only able to recruit 31 participants. Of those, 2 participants did not attend the second meeting, leaving us with a final sample of 29 participants.

14.5.1.2 Materials

In their study, Wyman and Vyse used the NEO Five Factor inventory (NEO-FFI, Costa & McCrae, 1985, 1992) to create psychological personality descriptions and the software Astrolabe (Astrolabe Inc, 2018) to create astrological natal charts for each participant. Then, an experimenter gave each participant their own psychological personality description and a psychological personality description belonging to another participant. The participant was then asked to decide which of the two personality descriptions was their own. This procedure was then repeated for the astrological personality description.

14.5.1.3 Procedure

The research design required two testing periods that were one week long and approximately two weeks apart. During the first testing period, the students assessed participants with the NEO-FFI personality inventory and collected information, that is, date and place of birth, that allowed them to create astrological natal charts for each participant with the free version of the software used by the original authors. In the second testing period the participants had to perform a simple choice task; they were asked to identify both their own psychological personality description and their astrological natal chart out of two descriptions each (i.e., a chance level of 50%).

14.5.1.4 Hypotheses

For the replication, our students took into account the direction of the original results and thus tested the ordinal hypothesis $\mathcal{H}_r$ that the success probability for psychological personality descriptions is *higher* than that for astrological personality descriptions. This hypothesis was then tested against a point-null hypothesis $\mathcal{H}_0$ that both success probabilities are equal to chance. Furthermore, as will be explained in the next section, the calculation of the Bayes factor required another hypothesis –referred to as the encompassing hypothesis $\mathcal{H}_e$– that both success probabilities can vary freely.

14.5.1.5 Analysis Plan

The students assigned a beta prior distribution to the model parameters and used the data from Wyman and Vyse (2008) to inform their prior beliefs. Specifically, based on Wyman and Vyse's data, the students assigned a $\text{Beta}(42, 12)$ prior distribution to the probability of correctly identifying one's own psychological personality description and a $\text{Beta}(25, 29)$ prior distribution to the success probability of correctly identifying one's own astrological personality description. That is, the prior for psychological personality descriptions favors success probabilities well above chance level while for astrological personality descriptions success probabilities at chance level are favored, with medians and 95% credible intervals of $0.77\,[0.65, 0.87]$ and $0.46\,[0.34, 0.59]$, respectively.

To compare $\mathcal{H}_0$ versus $\mathcal{H}_r$, the students first had to take a two-step approach. First, they needed to compute the Bayes factor between $\mathcal{H}_0$ and $\mathcal{H}_e$, denoted as BF_{0e}, using the Savage-Dickey density ratio (Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010) and the Bayes factor between $\mathcal{H}_r$ and $\mathcal{H}_e$, denoted as BF_{re}, using the encompassing prior approach (Klugkist et al., 2005). The students then obtained BF_{r0} through transitivity, that is: $\text{BF}_{r0} = \text{BF}_{re} \times \text{BF}_{e0}$. A detailed description of the statistical model is available in the online appendix.

14.5.2 Results of the replication study

In our replication study, out of 29 participants, 25 correctly identified their own psychological personality description and 18 participants correctly identified their own astrological personality description (see Table 14.2). Given our data and the prior knowledge provided by the Wyman and Vyse study, the result suggests extreme evidence (i.e., $\text{BF}_{r0} = 1884$) in favor of the hypothesis that people recognize their psychological personality description more reliably than their astrological personality description.

Table 14.2: Data from the current research project, as well as from Wyman and Vyse (2008) and Carlson (1985), where x_{psy} and x_{astro} denote the number of participants who correctly identified their psychological personality description and their astrological personality description, respectively, n_{psy} and n_{astro} denote the respective total number of observations, and $\hat{\theta}_{\text{psy}}$ and $\hat{\theta}_{\text{astro}}$ denote the sample proportions of correctly identifying one's own personality description.

	Data						
Study	x_{psy}	n_{psy}	$\hat{\theta}_{\text{psy}}$	x_{astro}	n_{astro}	$\hat{\theta}_{\text{astro}}$	Chance level
Current project	25	29	0.86	18	29	0.62	0.50
Wyman and Vyse (2008)	41	52	0.79	24	52	0.46	0.50
Carlson (1985)	25	56	0.45	28	83	0.38	0.33

14.6 Considerations for Lecturers

Our experience with this project suggests that three considerations warrant special attention. First, lectures should be aware of the prior knowledge of their students. The described project was designed for students who have some knowledge of Bayesian statistics, but also a basic background in the programming language R (R Core Team, 2021). However, despite their familiarity with the key concepts of Bayesian inference, our students found the mathematical parts of the project particularly challenging. Therefore, we recommend lectures to allow enough time to reiterate necessary mathematical components.

Second, when students are required to independently draft the preregistration we recommend the use of preregistration templates. For instance, the OSF offers preregistration templates for standard empirical research, but also replication studies (see `https://osf.io/zab38/wiki/home/` for an overview of all preregistration forms). The transparency checklist by Aczel, Szaszi, et al. (2020) is another highly accessible tool which covers the most important aspects for achieving transparency and openness in preregistrations and manuscripts.

Finally, in the described project, we based our target sample size on the number of participants in the original studies. Alternatively, lecturers could based their target sample size on a Bayesian design analysis (Stefan et al., 2019). A Bayesian design analysis is considered the Bayesian version of a frequentist power analysis and allows researchers to determine the minimum number of participants needed to achieve compelling evidence either in favor or against the hypothesis. Lecturers could also choose to do sequential testing, that is, monitor the evidence as the data accumulates and stop data collection as soon as the evidence is sufficiently compelling (e.g., Rouder, 2014b).

14.7 Summary

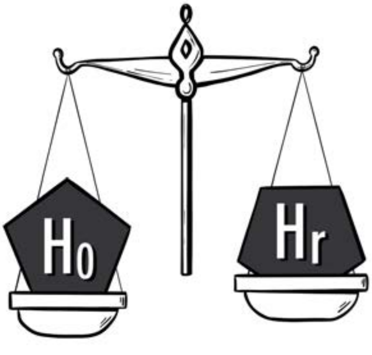
The discussed research project allowed students to learn a relevant Bayesian method to compute Bayes factors for ordinal expectations (i.e., the encompassing prior approach), and increase their understanding of the underlying mathematical concepts of Bayesian inference. We believe that this learning success was primarily due to the simplicity of the discussed statistical model which enabled the students to formulate the likelihood function, assign a prior distribution, derive the posterior distribution, and understand the encompassing prior approach even without strong mathematical background.

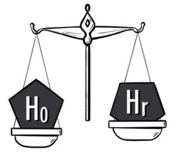
In addition, students gained practical experience through designing and conducting a reanalysis and replication study. Through this experience the students learned the advantages of Bayesian statistics in the context of replication research, for instance, by being able to quantify evidence for the absence of the predicted effect, but also by incorporating prior knowledge into their analyses and hence draw more informed decisions. In addition, the project gave students the opportunity to practice open research practices by letting them preregister their study, that is, create an analysis plan prior to data collection, and share their data, materials, and code. The confrontation with real data challenged the students to think in broader terms, that is, by discovering how different methods (i.e., the Savage-Dickey density ratio and the encompassing prior approach) can be utilized to answer specific research questions.

We believe that a research project is an ideal opportunity to integrate the theory and mathematics of Bayesian inference with hands-on experience, and confront students with all aspects of the empirical cycle. This experience gives students valuable insights into scientific practice, and equips them with problem solving skills that are necessary when they pursue their careers as psychological researchers and methodologists.

Part IV

Conclusion





Chapter 15

Discussion and Future Directions

15.1 Summary

In this dissertation, entitled “Good Research Practices”, I examined research practices and reform ideas aiming to combat the crisis of confidence (Pashler & Wagenmakers, 2012) in psychology. I did so through theoretical contributions and empirical work, developed practical guidelines for researchers, and demonstrated how principles of good research can be conveyed to students. In particular, the research methods and statistical practices presented in this dissertation facilitate the adherence to the following three principles: (1) respect the empirical cycle; (2) acknowledge uncertainty; and (3) enrich statistical models with theoretical knowledge. In the following subsections, I will discuss these principles in turn, put forward some ideas about statistical meta-science tools, highlight other facets of good research practices, and suggest how good research practices might be encouraged. I will end with some concluding remarks.

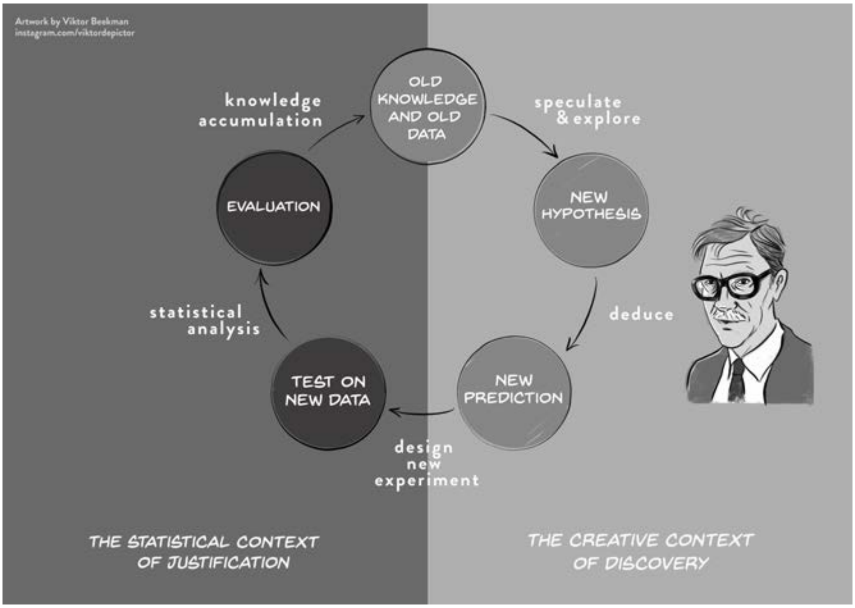
15.1.1 The First Principle: Respect The Empirical Cycle

The first principle is to respect the empirical cycle depicted in Figure 15.1 (De Groot, 1956/2014). Preregistration and analysis blinding allow researchers to make a sharp distinction between hypothesis-generation and hypothesis-testing, which prevents conscious and subconscious “fishing expeditions” for statistically significant effects. In psychology, a popular and widely used method to accomplish this is preregistration, in which researchers describe their hypotheses and complete analysis plan before coming into contact with the data on which they are tested.

Preregistration is an effective means of distinguishing exploratory from confirmatory results. In Chapter 5, we explored subjective experiences and expectations related to preregistration. Interestingly, researchers reported benefits of preregistration that extend beyond safeguarding the confirmatory status of the analysis, including the overall improvement of the quality of the projects. Criticism on preregistration concerned the increase of the overall project duration and work-related stress. This chapter establishes that the benefits outweigh the challenges, but only for researchers who already have experience with preregistration.

It has been recognized that preregistration goes along with effort and time investment (e.g., Nosek & Lindsay, 2018; van ’t Veer & Giner-Sorolla, 2016), but also that many preregistered analyses do not survive contact with the data, as unexpected peculiarities (e.g., outliers or violated assumptions) often demand that statistical models are adjusted after the fact (Claesen et al., 2021; Heirene et al., 2021; Nosek et al., 2019). An alternative to preregistration might be analysis blinding, a practice which promises more flexibility in data analysis. With analysis blinding, the statistical model is developed in interaction with the observed data; however, the data are altered in order to allow for an unbiased evaluation of the hypotheses (as illustrated in Figure 15.2). Chapter 6 advocated this practice and made concrete suggestions about how analysis blinding may be implemented for different research designs in experimental psychology.

We suspected that blinded analyses might have several advantages over preregistration. By providing greater flexibility in data analysis, we hypothesized that researchers would deviate less often from their analysis plan. In addition, since



Artwork by Viktor Beekman
instagram.com/viktordepictor

Figure 15.1: Empirical cycle and the distinction between the context of discovery and the context of justification. Figure available at `https://www.bayesianspectacles.org/library/` under CC license `https://creativecommons.org/licenses/by/2.0/`.

analysis blinding does not require the development of a preregistration protocol we suspected that it might additionally save researchers time and effort to complete a research project. Whether analysis blinding indeed had benefits over preregistration was assessed in Chapter 7. In this chapter, we compared the subjective time and effort required to develop an analysis plan and deviations from it among researchers who either preregistered their analysis or did analysis blinding in the Many-Analysts Religion Project. The hypothesis that analysis blinding takes less time and is less effortful than preregistration was not confirmed. However, researchers who developed their analysis plan based on a blinded version of the data were found to deviate less often from their analysis plans than researchers who preregistered their analysis, suggesting that analysis blinding achieves at least one important improvement.

Although this dissertation contrasted preregistration and analysis blinding, researchers may also use these practices in combination. Personally I would recommend that researchers preregister the study but finalize the statistical analysis on a blinded version of the data–in fact this was the procedure we used ourselves in Chapter 7.

Together with the transparent reporting of research methods, preregistration and analysis blinding were one of the topics covered in the Transparency Checklist presented in Chapter 11. The Transparency Checklist is a reporting checklist for the social sciences that lists key steps for achieving transparent research. The checklist was developed to be endorsed and implemented by journals as part of

Figure 15.2: Lady Justice evaluating two competing hypotheses as a symbol of unbiased research. Figure available at `https://www.bayesianspectacles.org/library/` under CC license `https://creativecommons.org/licenses/by/2.0/`.

the submission process as journals can have a profound impact in initiating research reforms (see e.g., Chambers, 2013; Kidwell et al., 2016; Nosek et al., 2015). Therefore, we developed the checklist not only in collaboration with open science advocates, but also with journal editors. Finally, we demonstrated in Chapter 13, how to teach students at the Master's level the importance of following the empirical cycle, as well as necessary skills to apply recent methodological reforms. Specifically, we described the teaching concept of the course "Good Research Practices", which is offered to students following the Research Master program at the University of Amsterdam.

15.1.2 The Second Principle: Acknowledge Uncertainty

The second principle is to acknowledge uncertainty. The most complete picture of a scientific phenomenon can be painted only if uncertainties and inconsistencies in statistical results and conclusions are explored and communicated transparently. This principle benefits the empirical cycle, as statistical analyses concern the step from 'test on new data' to 'evaluation'.

Preregistration and analysis blinding are usually seen in the context of a single research team executing a single analysis, implicitly assuming that there is a single, uniquely appropriate analysis procedure to answer a specific research question. Chapter 12 challenged this idea. In addition to good statistical practices, which improve a single analysis (e.g., data visualization and quantifying inferential uncertainty), the chapter introduced several ways to acknowledge uncertainty across multiple plausible analyses. Suggestions include assessing the impact of data pre-processing choices, reporting multiple models, and conducting many-analysts projects.

Chapter 3 and Chapter 4 demonstrated how the many-analyst approach can be applied in practice. The chapters described the Many-Analysts Religion Project in which 120 teams were invited to answer research questions on a much debated topic in the psychology of religion, that is, whether (1) religious people self-report higher well-being, and whether (2) the relation between religiosity and self-reported well-being depends on perceived cultural norms of religion. Although analytic procedures vastly differed from each other and there was a considerable variability in effect sizes, the conclusions regarding the first research question were robust against alternative statistical decisions, that is, most teams concluded that religious people do indeed self-report higher levels of well-being. With regard to the second research question, the conclusions differed across teams. The Many-Analysts Religion Project highlighted two benefits of involving multiple analysts. That is, the many-analysts approach allowed for the assessment of the robustness of the effect, but also yielded interesting additional insights (both thematic and methodological) through the contributions by the participating teams.

15.1.3 The Third Principle: Enrich Statistical Models With Theoretical Knowledge

The last principle is to enrich statistical models with theoretical knowledge. Incorporating theory-based knowledge into statistical models, for instance, through ordinal hypotheses and theory-informed parameter priors, allows psychological theories to be tested more comprehensively and effectively, both in replication research and original work. This principle benefits the empirical cycle in two ways. First, adequately quantifying predictions improves the step from ‘new hypothesis’ to ‘new prediction’. Second, the corresponding statistical procedure for testing these predictions improves the step from ‘test on new data’ to ‘evaluation’.

Chapter 8 proposed a methodology to evaluate specific ordinal hypotheses for categorical data. The chapter introduced an efficient statistical technique based on a bridge sampling routine to evaluate ordinal hypotheses and mixtures of ordinal hypotheses and equality hypotheses for binomial and multinomial variables. The associated software package `multibridge` was introduced in Chapter 9. So far, we could apply the proposed bridge sampling routine only to simple multinomial and binomial models. However, I also see great value in further developing the method and generalizing it to problems of increasing sophistication. For instance, in the field of cognitive psychology, the bridge sampling routine has already been successfully applied to multinomial processing tree models (Gronau et al., 2019) but not yet to test ordinal hypotheses. In Chapter 10, we discussed the advantages of testing ordinal hypotheses in MPT research and demonstrated its usefulness with two case studies. However, the methods we used in this chapter did not rely on the bridge sampling routine, but on the less efficient unconditional encompassing approach, which, although intuitive and simple, can become unreliable under certain conditions. Beyond demonstrating the usefulness of ordinal hypotheses, Chapter 10 also highlighted other aspects of theory-driven inference, such as the importance of theory-informed parameter priors.

We can inform model comparison by incorporating theoretical knowledge into the statistical model. Another way we aimed to incorporate knowledge in statistical inference is by quantifying the plausibility of research hypotheses. Chapter 2 established that there seems to be an intuitive plausibility inherent to psychological studies that can be picked up by laypeople. That is, intuitive plausibility of a research hypothesis seems to be indicative of replication success. Within the Bayesian statistical framework, these predictions may be used to inform the prior odds of two competing hypotheses. Moreover, laypeople's replication predictions could prove useful for selecting potential candidates for replication research.

Chapter 14 demonstrated how thesis projects on the undergraduate level can be structured so that students can experience the entire empirical cycle and learn to incorporate their theoretical knowledge into their statistical inference. The project described in the chapter outlined the teaching principles we applied to a Bachelor thesis project, in which students to conducted a preregistered replication study, informed parameter priors from results of the original studies, and tested an ordinal hypothesis.

15.2 Statistical Tools for Meta-Science

In this dissertation, I explored reform ideas, developed new statistical techniques, and analyzed the acceptance of existing methods among researchers. Thus, my work is situated at the intersection between statistical research methods and meta-science. In psychology, meta-science is still in its infancy. Therefore, it is not surprising that its methodology is not yet fully refined and many conclusions are based on descriptive statistics. I therefore see great research potential in further integrating my research interests, that is, working on the development of statistical methods for the evaluation of meta-scientific studies. Specifically, I would like to work on problems that tie in with two of my empirical papers, namely Chapter 2, in which we measured the subjective plausibility of social science studies, and Chapters 3 and 4, in which we conducted a many-analysts study. In the following, I will highlight how the methods in these chapters could be improved.

In Chapter 2 laypeople indicated for a particular social science study whether they thought it would replicate and how confident they were in their decision. We then plotted the prediction pattern for each study on a -100 to 100 scale, with 100 representing extreme confidence that the effect would replicate and -100 representing extreme confidence that the effect would fail to replicate. Inspecting the distributions of these prediction patterns led us to an interesting discovery: for studies where laypeople were in near agreement, predictions were highly accurate. In particular, when laypeople showed relatively high agreement that a study could not be replicated, it indeed failed to replicate. In this study, we assessed agreement on the replicability of a given study by visual inspection of the prediction patterns. That is, we roughly divided the prediction patterns into those for which laypeople as a collective were divided (as indicated by a bimodal distribution) and those for which laypeople agreed (as indicated by a unimodal distribution).

Rather than classifying prediction patterns based on visual inspection, future studies should apply statistical methods to examine whether prediction patterns are bimodal or unimodal. To classify prediction patterns, one could test, for instance, whether the data stem from a single data generating processes or a mixture of two processes (e.g., Schlattmann, 2009). Furthermore, to inform priors in the Bayesian statistical framework with laypeople's predictions, it would be useful to establish appropriate methods for summarizing prediction patterns into a common prior probability of the hypothesis. This quantity could form the basis for informed prior odds–that is, the ratio of the prior probabilities of two competing hypotheses– which in combination with the Bayes factor determine the posterior odds for the competing hypotheses (Kass & Raftery, 1995).

In Chapter 3 and 4, we described the Many-Analysts Religion Project (MARP), in which we examined the relationship between religiosity and well-being. For this project, we adopted a many-analysts approach to assess the robustness of the results and to describe the variability of the findings based on theory-driven analytic choices and plausible statistical models. The assumption we made regarding the many-analysts approach was that the more consistent the results are across analysis teams, the more confident we could be in the teams' conclusions.

To determine the extent to which results were consistent across analysis teams, we again exclusively relied on descriptive statistics. To summarize the effect size across analysis teams, we calculated the median effect size, reported the proportion of teams that found a positive effect, and reported the proportion of teams that concluded that there was evidence for the effect. While descriptive statistics provide valuable insight into whether results depend on the choice of an analysis method (e.g., are there teams that find inconclusive results while others find the expected effect?), interesting research questions remained unanswered. When can we assume that there is an effect, that is, do all teams have to agree on their results? What exactly does it mean when a certain proportion of teams find no effect? At what point can one assume low or high variability between teams?

For instance, concerning our second research question "Does the relation between religiosity and self-reported well-being depend on perceived cultural norms of religion?" there was some variability in the conclusions: 66 out of 101 teams (65%) the confidence/credible interval excluded zero and 54% of the teams concluded that there is good evidence for an effect. In this case, it is difficult to make a statement about the research question except to note that there seems to be considerable between-team variability. As many-analysts projects and multiverse analyses become more popular in psychology, it is necessary to develop more principled statistical techniques to adequately summarize the outcomes. Inspired by meta-analysis, we need "meta-analysts" techniques.

It seems an exciting challenge to me to develop statistical methods to address research questions such as *"Does every research team/analytic pipeline yield an effect in the same predicted direction?"*, *"How large is the between-analysis variability?"*, *What is the consensus-based effect size?*, and *"Can we classify analysis teams into groups?"*. In some ways many-analysts project might benefit from meta-analysis methods. That is, in meta-analysis, a common effect size is determined across the included studies, which traditionally denotes the average effect size across all studies weighted by the respective sample size. In addition, recently Rouder, Haaf, Davis-Stober, and Hilgard (2019) and Haaf and Rouder (2021) proposed the *Does-every-study meta-analysis* which answers the question "Does every study show the predicted effect?". A comparable measure would also be valuable in the case of many-analysts studies. What appeals to me about the development of such a 'meta-analysts" technique is that we can use ordinal hypotheses to test interesting research questions (e.g., "Does every research team find a positive effect?"). The studying this approach would combine two of the three principles of good research practices, that is, enrich statistical models with theoretical knowledge (by formulating ordinal hypotheses) and acknowledging uncertainty (through involving multiple analysts).

However, the methods of meta-analyses cannot simply be applied to many-analysts projects. The challenge of summarizing results from these projects is the large dependency structure of the results as teams address the research question based on the same data (in meta-analysis results are based on independent studies). This dependency must be adequately captured in the statistical model. If meta-analytic effect sizes are summarized as a weighted average, criteria need to be defined by which the results are weighted. Instead of weighing the results based on sample size, many-analysts results may be weighed according to their quality, so that higher quality analyses have a greater impact on the final results. One problem with this approach is the subjectivity that is introduced: as is evident in Chapter 3 and commentaries submitted by the teams (summarized in Chapter 4), analysts have strong and sometimes conflicting opinions about which analysis method is best to answer the research questions.

To me, the question also remains whether effect sizes are in fact the best way to summarize the results of many-analysts projects. In MARP, we based our conclusions mostly on effect sizes, but also assessed the teams' subjective beliefs. In particular, we asked analysis teams whether the data provide evidence in favor for the hypothesis and how likely they thought the hypothesis was (before and after they saw the data). Extending this set of questions further, it would be possible to extract the teams' shared knowledge about the hypotheses using cultural consensus theory for ordinal data (Anders & Batchelder, 2015; W. H. Batchelder & Romney, 1988; Oravecz, Vandekerckhove, & Batchelder, 2014; van den Bergh et al., 2020). Cultural consensus theory is applied when the true answer to a particular problem is not known a priori (e.g., whether or not an effect is present) and assumes that the agreement of members within a culture (or in this case analysis teams) contains information about a shared cultural truth. This theoretical framework is promising, since it gives an estimate about the consensus answer for each question, the teams' competencies in answering the questions, and the difficulty for each question.

Finally, we could apply Bayesian nonparametric models (Fraley & Raftery, 1998; Quintana, 2006) in particular cluster analyses, to investigate whether analysis teams can be classified into groups. With this approach, we can estimate the variance between clusters and try to identify commonalities of analysis teams within a group. As such, this approach could possibly shed light on which characteristics in the analysis pipeline lead to different effect sizes (e.g., number and type of covariates, a priori beliefs, final sample size).

15.3 Other Facets of Good Research Practices

My thesis primarily focused on the way in which good research practices can improve how theories are tested – in terms of the empirical cycle shown in Figure 15.1, my thesis emphasised the role of the statistical context of justification. However, good research practices also benefit the complementary part of the empirical cycle: the creative context of discovery.

Firstly, methodologists have argued that it is crucial to improve the development of psychological theories themselves (Borsboom, van der Maas, Dalege, Kievit, & Haig, 2021; Oberauer & Lewandowsky, 2019; Proulx & Morey, 2021). Theory development happens in the first step of the empirical cycle. 'Old knowledge and old data' typically constitutes either an existing theoretical framework the current study builds upon or is the basis of forming a new theory. The researcher then derives their new hypotheses from this theory.

Formalizing theory and hypotheses are good research practices that are often neglected. Oberauer and Lewandowsky (2019) argue that most psychological theories are not formulated strongly enough to be tested properly. A researcher who finds a certain effect may consider this evidence for the theory. However, the opposite often does not apply: the failure to find an effect could indicate that the theory is wrong, but it could also indicate that the construct of interest was not operationalized correctly. For instance, a researcher might decide to test priming theory, which states that activating a particular concept in a participant's mind influences their behavior. One hypothesis that can be derived from this theory is that when analytical thinking is active in people's minds, it triggers their religious disbelief (Gervais & Norenzayan, 2012). If the researcher finds the desired effect, he might conclude that this is evidence for the priming theory. If the researcher does not find an effect, he might conclude that, among other things, the concept was not activated strongly enough or that priming theory, although valid, does not apply to religiosity.

Many psychological theories are not sufficiently specific, and consequently they do not lend themselves to conclusive tests. This concern was echoed by Borsboom et al. (2021). In addition, Borsboom et al. (2021) claimed that while researchers in psychology are well trained in testing predictions they often lack skills that are conducive to constructing theories, such as theoretical modeling using mathematical tools. An example in psychology where theories have been quantified successfully as mathematical models are the multinomial processing tree (MPT) models introduced in Chapter 10. Through model equations built as a tree-like architecture, MPT models can capture the interplay of different psychological mechanisms, allowing the derivation of specific predictions.

Secondly, in order to move successfully from theory development to theory testing, researchers must ensure that the theoretical constructs have been properly measured. In the empirical cycle, this step falls under 'designing new experiment'. The way researchers in psychology handle measurement has been increasingly scrutinized; one point of critique concerns the fact that researchers pay little attention to psychometric properties, which may jeopardize the validity of the conclusions (Flake & Fried, 2020; Lilienfeld & Strother, 2020). This dissertation also shows some weaknesses regarding measurement practices. For instance, contrary to the suggestions in Flake & Fried, 2020, Chapter 7 did not report psychometric properties of the measurements used. In MARP, some participating teams pointed out that the assumption of measurement invariance, an important precondition for cross-cultural comparisons, was violated in the religiosity construct (Ross et al., 2022; Schreiner et al., 2022) potentially invalidating the statistical analyses.

There are many areas for improvement in psychology, and even a 300+ page dissertation titled "Good Research Practices" cannot do justice to all facets of good research. The ever-evolving abundance of good research practices can feel overwhelming and cause inertia: rather than implementing the suggested practices, researchers may decide not to implement any at all. I therefore take a pragmatic stance: instead of expecting psychologists to perfectly adhere to each facet of good research, I instead expect them to be open to scientific debates, to rethink their current research practices, and to be willing to learn new skills. At the same time, I expect my colleagues in the field of meta-science to offer constructive criticism rather than attacks, and to understand that other researchers may not see their specific meta-science agenda as a top priority.

15.4 Moving Forward: Encouraging Good Research Practices

Researchers may recognize the value of certain good research practices, but may not adopt them due to various subjective barriers and the assumed additional time required to implement the practices (see Houtkoop et al., 2018 for subjective barriers regarding data sharing). For instance, with regard to preregistration, a respondent in the survey reported in Chapter 5 wrote: "I understand the importance of [preregistration], but the amount of time and effort needed to preregister is probably the biggest reason I have avoided it in the past".

The ultimate goal in meta-science should be that good research practices are applied in practice. To achieve this, it is necessary to create incentives for researchers–simply pointing out their benefits in scientific publications may not be enough to change an entire research culture. Instead, the entire academic system must pull together: journals, funding agencies, and institutions.

In recent years, we have seen many positive developments in this regard. An increasing number of journals are promoting transparency standards, for instance by signing on to the Transparency and Openness Promotion guidelines (Nosek et al., 2015), or by providing open science badges for preregistration and sharing of data and materials (Kidwell et al., 2016). In collaboration with journal editors we developed the Transparency Checklist (Chapter 11) so that participating journals may make the Transparency Checklist part of their submission process in the future.

Institutions and funding agencies have a profound impact on the research culture because they are involved in the hiring process and fund research projects. University hiring processes may focus more on the quality of applicants' published work (e.g., whether their studies were preregistered, whether they assessed the robustness of results, whether their shared their data, materials, and code) rather than the sheer number of publications. This would provide an incentive especially for those researchers who currently shy away from implementing practices such as preregistration for fear of being less productive. In the Netherlands, research practices such as open access publishing and data sharing are particularly encouraged by government institutions. They form two of the three pillars of the National Plan for Open Science (NPOS; National Program Open Science, 2022), whose steering committee consists of the Association of Dutch Universities, the Dutch Federation of University Medical Centers, and the Dutch Research Council. The latter also requires that all projects it funds be published in open access and that data be shared whenever possible (Dutch Research Council, 2022).

While I strongly support these measures, I would like to see some of the research practices discussed in this thesis stronger promoted, such as preregistration, analysis blinding, or many-analysts/multiverse projects. It is undoubtedly challenging to combine the demands of different disciplines regarding these practices – what works in psychology does not necessarily work in other research fields. For instance, preregistration might be not feasible in disciplines such as physics, as the analysis is often too complex to anticipate. But researchers in these fields might find it more feasible to do analysis blinding (e.g., MacCoun & Perlmutter, 2015). To discuss these challenges and achieve nationwide change, we are currently in the process of establishing an interdisciplinary ‘Reproducibility Network’ in the Netherlands, involving both researchers and other stakeholders.

Individual departments may also enforce good research practices more strongly. A simple measure would be to integrate good research practices within ethical reviews. In the department of psychology at the University of Amsterdam, both students and researchers must include in their ethics application a statement about whether their study–if it is confirmatory research–has been preregistered. If the study is preregistered they must provide the link of the preregistration, if they do not preregister the study they must provide a justification why this is not the case. In addition, our department has made good research practices an integral part within the curricula of students. The course “Good Research Practices” (presented in Chapter 13) is mandatory for all students following the Research Master program. In the future, it would also be worth considering to offer a similar course at the undergraduate level, so that students specializing for industry or clinical work can also benefit.

Finally, individual researchers can promote good research practices within their scientific networks. For instance, researchers can launch bottom-up networks at their institutions so-called Open Science Communities (e.g., Armeni et al., 2021). Open Science Communities (OCSs) aim to exchange knowledge about good research practices across faculty and disciplines, to learn together, to identify barriers that hinder peers in opening up their workflows, and to provide the support needed to foster cultural change at the institutional level (Armeni et al., 2021; Nosek, 2019). In the Netherlands, OSCs are currently represented at all major Dutch universities (see Figure 15.3). At the beginning of my doctoral studies, I founded the OSC Amsterdam together with my colleague Suzanne Hoogeveen. Since then, the community has grown to be a joint initiative of four institutions:

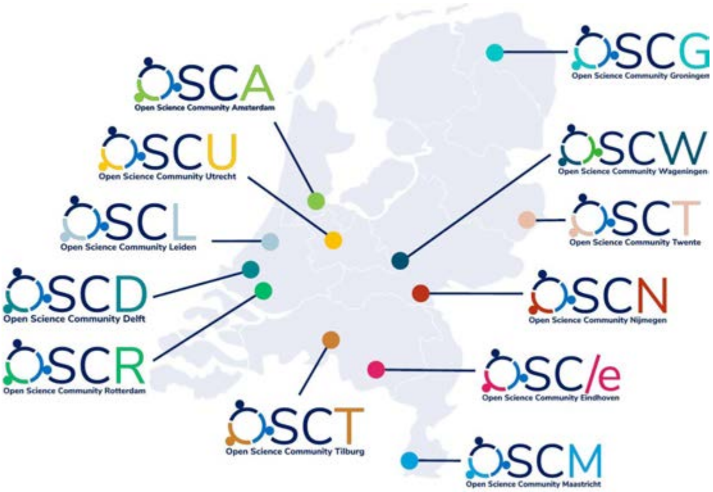

OSCA
Open Science Community Amsterdam
OSCU
Open Science Community Utrecht
OSCL
Open Science Community Leiden
OSCD
Open Science Community Delft
OSCR
Open Science Community Rotterdam
OSCT
Open Science Community Tilburg
OSCG
Open Science Community Groningen
OSCW
Open Science Community Wageningen
OSCT
Open Science Community Twente
OSCN
Open Science Community Nijmegen
OSC/e
Open Science Community Eindhoven
OSCM
Open Science Community Maastricht

Figure 15.3: Open Science Communities are currently represented in twelve cities in the Netherlands. Figure available at `https://www.osc-nl.com`.

the University of Amsterdam, the Free University Amsterdam, the University of Applied Sciences, and the Amsterdam Medical Centrum with currently 225 scholars subscribed to the initiative (see `https://openscience-amsterdam.com`).

Some universities (such as the University of Utrecht) explicitly encourage OSCs by providing financial support and creating positions for open science coordinators to build and strengthen the communities and organize events. In Amsterdam, of the four organizations, only the Free University of Amsterdam is currently receiving funding; the University of Amsterdam is not. One of our main goals is to ensure the sustainability of the communities in all four institutions.

15.5 Concluding Remarks

The goal of my dissertation was to study promising reform ideas to combat the crisis of confidence in psychology and to develop statistical methods that support researchers in their work. I hope that I have been able to make a contribution to combating the crisis, and that researchers will find the insights presented here valuable for their own work and that of their students.

References

Aalbersberg, I. J., Appleyard, T., Brookhart, S., Carpenter, T., Clarke, M., Curry, S., . . . Vazire, S. (2018). *Making science transparent by default; Introducing the TOP statement.* OSF Preprints. Retrieved from `https://osf.io/sm78t`
232, 252

Abdel-Khalek, A. M. (2006). Measuring happiness with a single-item scale. *Social Behavior and Personality: An International Journal*, *34*, 139–150. doi: 10.2224/sbp.2006.34.2.139
38

Abu-Raiya, H. (2013). On the links between religion, mental health and inter-religious conflict: A brief summary of empirical research. *The Israel Journal of Psychiatry and Related Sciences*, *50*, 130–139.
39

Aczel, B., Hoekstra, R., Gelman, A., Wagenmakers, E., Klugkist, I. G., Rouder, J. N., . . . van Ravenzwaaij, D. (2020). Discussion points for Bayesian inference. *Nature Human Behaviour*, *4*, 561–566. Retrieved from `https://doi.org/10.1038/s41562-019-0807-z`
238

Aczel, B., Szaszi, B., Nilsonne, G., Van den Akker, O., Albers, C. J., van Assen, M. A. L. M., . . . Wagenmakers, E.-J. (2021). Science forum: Consensus-based guidance for conducting and reporting multi-analyst studies. *Elife*, *10*, e72185.
40, 41, 70, 83

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, v., Benjamin, D., . . . Wagenmakers, E.-J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, *4*, 4–6.
82, 252, 275

Aird, F., Kandela, I., Mantis, C., & Reproducibility Project: Cancer Biology. (2017). Replication study: BET bromodomain inhibition as a therapeutic strategy to target c-myc. *eLife*, e21253.
91

Alister, M., Vickers-Jones, R., Sewell, D. K., & Ballard, T. (2021). How do

we choose our giants? Perceptions of replicability in psychological science. *Advances in Methods and Practices in Psychological Science*, *4*, 1–21.
73

Alter, G., & Gonzalez, R. (2018). Responsible practices for data sharing. *American Psychologist*, *73*, 146–156.
252

American Statistical Association. (2014). *Curriculum guidelines for undergraduate programs in statistical science.* Published by the American Statistical Association at `http://www.amstat.org/education/curriculumguidelines.cfm`.
268, 269

Amrhein, V., Greenland, S., & McShane, B. B. (2019). Retire statistical significance. *Nature*, *567*, 305–307.
238

Anders, R., & Batchelder, W. H. (2015). Cultural Consensus Theory for the ordinal data case. *Psychometrika*, *80*, 151–181.
286

Anderson, C., Kraus, M. W., Galinsky, A. D., & Keltner, D. (2012). The local-ladder effect: Social status and subjective well-being. *Psychological Science*, *23*, 764–771. doi: 10.1177/0956797611434537
20

Anderson, M. S., Martinson, B. C., & De Vries, R. (2007). Normative dissonance in science: Results from a national survey of US scientists. *Journal of Empirical Research on Human Research Ethics*, *2*, 3–14.
239

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*, 17–21.
240

Armeni, K., Brinkman, L., Carlsson, R., Eerland, A., Fijten, R., Fondberg, R., ... Zurita-Milla, R. (2021). Towards wide-scale adoption of open science practices: The role of open science communities. *Science and Public Policy*, *48*, 605–611.
289

Arnold, N. R., Heck, D. W., Bröder, A., Meiser, T., & Boywitt, C. D. (2019). Testing hypotheses about binding in context memory with a hierarchical multinomial modeling approach. *Experimental Psychology*, *66*, 239–251.
189

Astrolabe Inc. (2018). *Astrolabe [Computer software].* `https://alabe.com/freechart/`.
273

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725.
247

Atkinson, Q. D., Claessens, S., Fischer, K., Forsyth, G. L., Kyritsis, T., Wiebels, K., & Moreau, D. (2022). Being specific about generalisability. *Religion, Brain, & Behaviour*.
64, 65, 66

Bains, S. (2011, January). Questioning the integrity of the John Templeton Foundation. *Evolutionary Psychology*, *9*, 92–115.
41

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*, 452.
14, 232, 256, 269

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.
93

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666–678.
93

Balkaya-Ince, M., & Schnitker, S. (2022). Advantages of using multilevel modeling approaches for the Many Analysts Religion Project. *Religion, Brain, & Behaviour*.
68

Barber, T. X. (1976). *Pitfalls in human research: Ten pivotal points* (Vol. 67). Elsevier.
93

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*, 1281–1311.
203, 227

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., . . . Bringmann, L. F. (2020). Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, *137*, 110211.
39, 40, 238, 246, 249

Batchelder, W., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*, 548–564.
196

Batchelder, W., & Riefer, D. M. (1991). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86.
160

Batchelder, W., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Reviewy*, *6*, 57–86.
196

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*, 71–92.
286

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 197–215.
197, 198

BBC Radio 4. (2018). The replication crisis [Radio Broadcast]. *BBC*. Retrieved

from https://www.bbc.co.uk/programmes/m00013p9
260

Bell, R., & Buchner, A. (2010). Valence modulates source memory for faces. *Memory & Cognition*, *38*, 29–41.
209

Bell, R., Buchner, A., Kroneisen, M., & Giang, T. (2012). On the flexibility of social source memory: A test of the emotional incongruity hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1512–1529.
209

Bell, R., Mieth, L., & Buchner, A. (2015). Appearance-based first impressions and person memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 456–472.
195, 198, 205, 206, 207, 209, 212

Bello, S., Krogsbøll, L. T., Gruber, J., Zhao, Z. J., Fischer, D., & Hróbjartsson, A. (2014). Lack of blinding of outcome assessors in animal model experiments implies risk of bias. *Journal of clinical epidemiology*, *67*, 973–983.
110

Bem, D. J. (1987). Writing the empirical journal. In M. R. Zanna & J. M. Darley (Eds.), *The compleat academic: A practical guide for the beginning social scientist* (pp. 171–201). Mahwah, NJ: Lawrence Erlbaum Associates.
251

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 551–572.
138, 170, 178, 179

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10.
7, 14, 15, 36, 106, 238, 261, 263

Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, *22*, 245–268.
149, 171

Bennett, J. H. (Ed.). (1990). *Statistical inference and analysis: Selected correspondence of R. A. Fisher.* Oxford: Clarendon Press.
253

Berger, J. O., & Molina, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica*, *59*, 3–15.
144, 182

Bergin, A. E. (1983). Religiosity and mental health: A critical reevaluation and meta-analysis. *Professional Psychology: Research and Practice*, *14*, 170–184.
38

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading: Addison-Wesley.
273

BITSS. (2017). 2017 BITSS Annual Meeting – Day 1 [YouTube movie]. *Berkeley*

Initiative for Transparency in the Social Sciences (BITSS). Retrieved from `https://www.youtube.com/watch?v=4IgSVxkXMaM`
263

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., ... Wagenmakers, E.-J. (2018, December). Estimating across-trial variability parameters of the Diffusion Decision Model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *87*, 46–75.
39, 40

Boehm, U., Hawkins, G. E., Brown, S. D., van Rijn, H., & Wagenmakers, E.-J. (2016). Of monkeys and men: Impatience in perceptual decision–making. *Psychonomic Bulletin & Review*, *23*, 738–749.
249

Bohannon, J. (2015). *I fooled millions into thinking chocolate helps weight loss. Here's how.* (Blog No. May 27). http://io9.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800.
93

Böing-Messing, F., & Mulder, J. (2016). Automatic Bayes factors for testing variances of two independent normal distributions. *Journal of Mathematical Psychology*, *72*, 158–170.
159

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*, 756–766.
287

Botvinik–Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., ... Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*, 84–88. Retrieved from `https://doi.org/10.1038/s41586-020-2314-9`
39, 40, 238, 247, 249

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.
26

Brown, N., & Heathers, J. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8*, 363–369.
263

Browne, M. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*, 108?-132.
95

Bulbulia, J. A. (2022). Causal models are needed to infer how religion affects mental health. *Religion, Brain, & Behaviour*.
65

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 1–12.
14, 256

Cairo, A. (2019). *How charts lie: Getting smarter about visual information.* New York: WW Norton & Company.
241

Camerer, C. F., Dreber, A., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., ... Razen, M. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*, 1433–1436. doi: 10.1126/science.aaf0918
35

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2018). Evaluating replicability of social science experiments in *Nature* and *Science. Nature Human Behaviour*, *2*, 637–644.
6, 14, 15, 17, 19, 28, 31, 34, 35, 90, 256, 269

Campbell, P. (2013). Announcement: Reducing our irreproducibility. *Nature*, *496*, 398.
232

Captari, L. E., Hook, J. N., Hoyt, W., Davis, D. E., McElroy-Heltzel, S. E., & Worthington Jr., E. L. (2018). Integrating clients' religion and spirituality within psychotherapy: A comprehensive meta-analysis. *Journal of Clinical Psychology*, *74*, 1938–1951.
41

Carlson, S. (1985). A double-blind test of astrology. *Nature*, *318*, 419–425.
271, 273

Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, *6*, 1–13.
101, 246

Cavagnaro, D. R., & Davis-Stober, C. P. (2014). Transitive in our preferences, but transitive in different ways: An analysis of choice variability. *Decision*, *1*, 102–122.
138, 270

Center for Open Science. (2021). *Open Science Framework* [Website]. `https://osf.io/`.
117, 269

Chadwick, J. (1932). Possible existence of a neutron. *Nature*, *129*, 312.
250

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at *Cortex. Cortex*, *49*, 609–610.
6, 15, 72, 91, 232, 256, 263, 269, 282

Chambers, C. D. (2015). Ten reasons why journals must review manuscripts before results are known. *Addiction*, *110*, 10–11.
91

Chambers, C. D. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice.* Princeton: Princeton University Press.
6, 90, 116, 258, 260, 261, 262, 263

Chambers, C. D., & Tzavella, L. (2021). The past, present and future of Registered Reports. *Nature Human Behaviour*, 1–14.
269

Chang, M.-C., Chen, P.-F., Lee, T.-H., Lin, C.-C., Chiang, K.-T., Tsai, M.-F., ... Lung, F.-W. (2021, March). The effect of religion on psychological resilience in healthcare workers during the coronavirus disease 2019 pandemic. *Frontiers in Psychology*, *12*, 628894.
41

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). *shiny: Web application framework for R [Computer software].* `http://CRAN.R-project.org/package=shiny`.
241

Chechile, R. A. (2009). Pooling data versus averaging model fits for some prototypical multinomial processing tree models. *Journal of Mathematical Psychology*, *53*, 562–576.
202

Chen, C., Härdle, W., & Unwin, A. (Eds.). (2008). *Handbook of data visualization.* Berlin: Springer Verlag.
242

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*, 1313–1321.
143

Chiu, C.-Y., Gelfand, M. J., Yamagishi, T., Shteynberg, G., & Wan, C. (2010). Intersubjective culture: The role of intersubjective perceptions in cross-cultural research. *Perspectives on Psychological Science*, *5*, 482–493.
39

Chopik, W., Bremner, R., Defever, A., & Keller, V. (2018). How (and whether) to teach undergraduates about the replication crisis in psychological science. *Teaching of Psychology*, *45*, 158–163.
257, 269

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*, 211037.
116, 122, 132, 280

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*, 531–554.
242

Cobb, G. (2015). Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. *The American Statistician*, *69*, 266–282.
268

Committee on Professional Ethics of the American Statistical Association. (2018). *Ethical guidelines for statistical practice.* Retrieved from `https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx`
242, 246, 248, 252

Conley, A., Goldhaber, G., Wang, L., Aldering, G., Amanullah, R., Commins, E. D., ... The Supernova Cosmology Project (2006). Measurement of ωm, $\omega\lambda$ from a blind analysis of type Ia supernovae with CMAGIC: Using color

information to verify the acceleration of the Universe. *The Astrophysical Journal*, *644*, 1–20.
92, 94

Cooper, R. J., Schriger, D. L., & Close, R. J. (2002). Graphical literacy: The quality of graphs in a large-circulation journal. *Annals of Emergency Medicine*, *40*, 317–322.
245

Costa, P. T., & McCrae, R. R. (1985). *The NEO personality inventory.* Odessa, FL: Psychological Assessment Resources.
273

Costa, P. T., & McCrae, R. R. (1992). *NEO-PI-R and NEO-FFI professional manual* (Vol. 38). Odessa, FL.
273

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., ... Wagenmakers, E.-J. (2016). Hidden multiplicity in multiway ANOVA: Prevalence, consequences, and remedies. *Psychonomic Bulletin & Review*, *23*, 640–647.
104

Crüwell, S., & Evans, N. J. (2019). Preregistration in complex contexts: A preregistration template for the application of cognitive models. *Manuscript submitted for publication.* Retrieved from `https://psyarxiv.com/2hykx/`
82

Cumming, G. (2008). Replication and p intervals: p-values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286–300. doi: 10.1111/j.1745-6924.2008.00079.x
34

Damien, P., & Walker, S. G. (2001). Sampling truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics*, *10*, 206–215.
145, 153, 176

Davis, W. E., Sellers, P. D., Miller, E., Machluf, K., Peters, P., Scott J, Salvatore, J., ... Steltenpohl, C. (2017, Sep). *Integrating Open Science in the undergraduate psychology curriculum.* OSF. Retrieved from `osf.io/f82ej`
257

Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology*, *53*, 1–13.
138, 270

De Groot, A. D. (1956/2014). The meaning of "significance" for different types of research. Translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas. *Acta Psychologica*, *148*, 188–194.
5, 90, 245, 246, 247, 262, 280

De Groot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioral sciences.* The Hague: Mouton.
90

de Molière, L., & Harris, A. J. L. (2016). Conceptual and direct replications

fail to support the stake–likelihood hypothesis as an explanation for the interdependence of utility and likelihood judgments. *Journal of Experimental Psychology: General*, *145*, e13.
90

de Vrieze, J. (2018). The metawars. *Science*, *361*, 1184–1188.
41

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta–analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*, 627–668.
103

Del Giudice, M., Gangestad, S. W., & Steven, W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, *4*, 2515245920954925.
248

DellaVigna, S., & Vivalt, E. (2019). *Social Science Prediction Platform* [Website]. `https://socialscienceprediction.org/`.
36

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020). The case for formal methodology in scientific reform. *bioRxiv*. Retrieved from `https://www.biorxiv.org/content/early/2020/11/22/2020.04.26.048306`
6

Devroye, L. (1986). Sample-based non-uniform random variate generation. In *Proceedings of the 18th Conference on Winter Simulation* (pp. 260–265).
145

Diamond, L., & Lerch, F. J. (1992). Fading frames: Data presentation and framing effects. *Decision Sciences*, *23*, 1050–1071.
242

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*, 204–223.
143, 205

Dickey, J. M., & Lientz, B. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.
143, 205, 274

Diener, E., Tay, L., & Myers, D. G. (2011). The religion paradox: If religion makes people happy, why are so many dropping out? *Journal of Personality and Social Psychology*, *101*, 1278–1290.
39

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.
197

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, *5*, 781.
269

Dovidio, J. F. (2016). Commentary: A big problem requires a foundational change.

Journal of Experimental Social Psychology, *66*, 159–165. doi: 10.1016/j.jesp.2016.01.008
15

Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., & Chevalier, F. (2019). Increasing the transparency of research papers with explorable multiverse analyses. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15.
248

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., ... Johannesson, M. (2015). Using Prediction Markets to Estimate the Reproducibility of Scientific Research. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 15343–15347. doi: 10.1073/pnas.1516179112
14, 19, 34, 35, 36

Dressler, W. W., Balieiro, M. C., Ribeiro, R. P., & Santos, J. E. D. (2007). Cultural consonance and psychological distress: Examining the associations in multiple cultural domains. *Culture, Medicine and Psychiatry*, *31*, 195–224.
39

Durante, K. M., Rae, A., & Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, *24*, 1007–1016.
246

Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting*, *5*, 17–34.
179

Dutch Research Council. (2022). *Open science.* `https://www.nwo.nl/en/open-science`.
289

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., ... Donkin, C. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, *26*, 1051–1069.
40, 249

Dutilh, G., Sarafoglou, A., & Wagenmakers, E.-J. (2019). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, *198*, S5745–S5772.
7, 83, 116, 262

Dutilh, G., Vandekerckhove, J., Ly, A., Matzke, D., Pedroni, A., Frey, R., ... Wagenmakers, E.-J. (2017a). A test of the diffusion model explanation for the worst performance rule using preregistration and blinding. *Attention, Perception, & Psychophysics*, *79*, 713–725.
93, 99, 101

Dutilh, G., Vandekerckhove, J., Ly, A., Matzke, D., Pedroni, A., Frey, R., ... Wagenmakers, E.-J. (2017b). A test of the diffusion model explanation for the worst performance rule using preregistration and blinding. *Attention, Perception, & Psychophysics*, *79*, 713–725.

116

Ebert, T., Gebauer, J. E., Talman, J. R., & Rentfrow, P. J. (2020). Religious people only live longer in religious cultural contexts: A gravestone analysis. *Journal of Personality and Social Psychology*, *119*, 1–6.
38, 39

Edelsbrunner, P. A., Sebben, S., Frisch, L. K., Schüttengruber, V., Protzko, J., & Thurn, C. M. (2022). How to understand a research question – A challenging first step in setting up a statistical model. *Religion, Brain, & Behaviour*.
64

Eerland, A., Sherrill, A. M., Magliano, J. P., & Zwaan, R. A. (2016). Registered replication report: Hart & albarracín (2011). *Perspectives on Psychological Science*, *11*(1), 158–171.
90

Epskamp, S., & Nuijten, M. B. (2016). statcheck: Extract statistics from articles and recompute *p* values (R package version 1.3.0) [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=statcheck` (R package version 1.3.0)
185, 259, 263

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: Areview of the literature. *Zeitschrift für Psychologie/Journal of Psychology*, *217*, 108–124.
160, 196

Erdfelder, E., Hu, X., Rouder, J. N., & Wagenmakers, E.-J. (2020). Cognitive psychometrics: The scientific legacy of William H. Batchelder (1940–2018) [editorial]. *Journal of Mathematical Psychology*, *99*, 102468.
160

Etz, A., Bartlema, A., Vanpaemel, W., Wagenmakers, E.-J., & Morey, R. D. (2019). An exploratory survey of student and researcher intuitions about statistical evidence. Poster presented at the 31st annual convention of the Association of Psychological Science (APS), Washington, USA.
18

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. *PLoS ONE*, *11*, e0149794.
90

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34.
268

European Commision. (2004). *Report by Eurostat on the revision of the Greek government deficit and debt figures* [Eurostat Report]. `https://ec.europa.eu/eurostat/web/products-eurostat-news/-/GREECE`.
181

European Commision. (2010). *Report on Greek government deficit and debt statistics* [Eurostat Report]. `https://ec.europa.eu/eurostat/web/products-eurostat-news/-/COM_2010_REPORT_GREEK`.
181

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis.* Chichester: John Wiley & Sons.

240

Feynman, R. (1998). *The meaning of it all: Thoughts of a citizen–scientist.* Reading, MA: Perseus Books.
90

Fiedler, K. (2018). The creative cycle and the growth of psychological science. *Perspectives on Psychological Science*, *13*, 433–438.
6, 72, 73

Field, S. M., Wagenmakers, E.-J., Kiers, H. A., Hoekstra, R., Ernst, A. F., & van Ravenzwaaij, D. (2020). The effect of preregistration on trust in empirical research findings: Results of a Registered Report. *Royal Society Open Science*, *7*, 181351.
73

Fillard, P., Descoteaux, M., Goh, A., Gouttard, S., Jeurissen, B., Malcolm, J., . . . Poupon, C. (2011, May). Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. *NeuroImage*, *56*, 220–234.
40

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*, 456–465.
288

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., . . . Dreber, A. (2018). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*. doi: 10.1016/j.joep.2018.10.009
14, 15, 19, 31, 34, 35

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre–SMA facilitate decision–making under time pressure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 17538–17542.
101

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, *41*, 578–588.
286

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*, 153–169.
14, 256

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., . . . Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*, 421–435.
264

Frankenhuis, W. E., & Nettle, D. (2018). Open science is liberating and can foster creativity. *Perspectives on Psychological Science*, *13*, 439–447.
82

Frigyik, B. A., Kapila, A., & Gupta, M. R. (2010). *Introduction to the Dirichlet distribution and related processes* (Tech. Rep.). Washington: Department of Electrical Engineering, University of Washington.

151, 190

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, *7*, 143–167.
151

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, *18*, 3–12.
257, 269

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*, 389–402.
203, 240

Garssen, B., Visser, A., & Pool, G. (2020, February). Does spirituality or religion positively affect mental health? Meta-analysis of longitudinal studies. *The International Journal for the Psychology of Religion*, *31*, 4–20.
38

Gebauer, J. E., Sedikides, C., Schönbrodt, F. D., Bleidorn, W., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2017). The religiosity as social value hypothesis: A multi-method replication and extension across 65 countries and three levels of spatial aggregation. *Journal of Personality and Social Psychology*, *113*, e18–e39.
38, 39

Gelfand, A. E., Smith, A. F., & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association*, *87*, 523–532.
123, 147, 148, 205, 223

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, *13*, 755-779. doi: 10.1198/106186004X11435
240

Gelman, A. (2011). Why tables are really much better than graphs. *Journal of Computational and Graphical Statistics*, *20*, 3–7.
241

Gelman, A. (2020). *Why we sleep – A tale of institutional failure* [Scientific Blog]. `https://statmodeling.stat.columbia.edu/2020/03/24/why-we-sleep-a-tale-of-institutional-failure/`.
4

Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis–a "garden of forking paths"–explains why many statistically significant comparisons don't hold up. *American scientist*, *102*(6), 460.
90, 93

Gelman, A., Pasarica, C., & Dodhia, R. (2002). Let's practice what we preach: Turning tables into graphs. *The American Statistician*, *56*, 121–130.
242

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipland, F., & Hothorn,

T. (2020). mvtnorm: Multivariate normal and t distributions [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=mvtnorm` (R package version 1.1-1.)
176

George, L. K., Ellison, C. G., & Larson, D. B. (2002, July). Explaining the relationships between religious involvement and health. *Psychological Inquiry*, *13*, 190–200. doi: 10.1207/S15327965PLI1303_04
38

Gernsbacher, M. A. (2018). Writing empirical articles: Transparency, reproducibility, clarity, and memorability. *Advances in methods and practices in psychological science*, *1*, 403–414.
232

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, *336*, 493–496.
287

Gilbert, E. W. (1958). Pioneer maps of health and disease in England. *The Geographical Journal*, *124*, 172–183.
240

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*, 562–571. doi: 10.1177/1745691612457576
14, 15

Gneezy, U., Keenan, E. A., & Gneezy, A. (2014). Avoiding Overhead Aversion in Charity. *Science*, *346*, 632–635. doi: 10.1126/science.1253932
20, 24

Goldacre, B. (2009). *Bad science.* London: Fourth Estate.
90, 259, 264

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90.
138

Good, I. J. (1967). A Bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *29*, 399–431.
143

Good, I. J. (1971). 46656 varieties of Bayesians. *The American Statistician*, *25*, 62–63.
238

Gopalakrishna, G., Wicherts, J., Vink, G., Stoop, I., van den Akker, O., ter Riet, G., & Bouter, L. (2021). Prevalence of responsible research practices and their potential explanatory factors: A survey among academic researchers in The Netherlands. *Manuscript submitted for publication.* Retrieved from `https://osf.io/preprints/metaarxiv/xsn94/`
73

Gøtzsche, P. C. (1996). Blinding during data analysis and writing of manuscripts. *Controlled Clinical Trials*, *17*, 285–290.
93

Grayson, D. (1988). Two-group classification in latent trait theory: Scores with

monotone likelihood ratio. *Psychometrika*, *53*, 383–392.
138

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.
29

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*, 337–350.
263

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ... Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97.
123, 140, 149, 151, 176, 177

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software, Articles*, *92*(10), 1–29.
123, 176, 224

Gronau, Q. F., Wagenmakers, E.-J., Heck, D. W., & Matzke, D. (2019). A simple method for comparing complex models: Bayesian model comparison for hierarchical multinomial processing tree models using Warp-III bridge sampling. *Psychometrika*, *84*, 261–284.
198, 203, 224, 283

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020, September). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, *15*, 1243–1255.
41

Gu, X., Hoijtink, H., Mulder, J., & Rosseel, Y. (2019). Bain: a program for bayesian testing of order constrained hypotheses in structural equation models. *Journal of Statistical Computation and Simulation*, *89*, 1526-1553.
139, 140, 157, 159, 171

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*, 511-527.
139, 171

Guo, Y., & Regenwetter, M. (2014). Quantitative tests of the perceived relative argument model: Comment on Loomes (2010). *Psychological Review*, *121*, 696–705.
138, 270

Guzey, A. (2019). *Matthew Walker's "why we sleep" is riddled with scientific and factual errors* [Scientific Blog]. `https://guzey.com/books/why-we-sleep/#appendix-what-do-you-do-when-a-part-of-the-graph-contradicts-your-argument-you-cut-it-out-of-course`.
4

Haaf, J. M., Klaassen, F., & Rouder, J. (2019). Capturing ordinal theoretical constraint in psychological science. *Manuscript submitted for publication*. Retrieved from `https://psyarxiv.com/a4xu9/`
8, 170

Haaf, J. M., Merkle, E. C., & Rouder, J. N. (2020). Do items order? the psychology in IRT models. *Journal of Mathematical Psychology*, *98*. doi: https://doi.org/10.1016/j.jmp.2020.102398
139, 148, 159, 270

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*, 779–798.
139, 148, 214, 225

Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, *26*, 772–789.
212, 214, 225

Haaf, J. M., & Rouder, J. N. (2021). Does every study? Implementing ordinal constraint in meta-analysis. *Psychological Methods*.
139, 148, 286

Haberman, S. J. (1978). *Analysis of qualitative data: Introductory topics* (Vol. 1). Academic Press.
138, 139, 143

Hackney, C. H., & Sanders, G. S. (2003). Religiosity and mental health: A meta–analysis of recent studies. *Journal for the Scientific Study of Religion*, *42*, 43–55.
38, 41

Hanel, P. H. P., & Zarzeczna, N. (2022). From multiverse analysis to multiverse operationalisations: 262,143 ways of measuring well-being. *Religion, Brain, & Behaviour*.
66

Hardwicke, T. E., Tessler, M. H., Peloquin, B., & Frank, M. C. (2018). A Bayesian decision-making framework for replication. *Behavioral and Brain Sciences*, *41*, e132. doi: 10.1017/S0140525X18000675
15

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah (NJ): Lawrence Erlbaum.
238

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS ONE*, *8*, e72467.
90

Harris, J. I., Usset, T., Voecks, C., Thuras, P., Currier, J., & Erbes, C. (2018, September). Spiritually integrated care for PTSD: A randomized controlled trial of "Building Spiritual Strength". *Psychiatry Research*, *267*, 420–428.
41

Hart, S. G. (2021). Nasa-Task Load Index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*, 904–908.
122

Haven, T. L., Errington, T. M., Gleditsch, K. S., van Grootel, L., Jacobs, A. M., Kern, F. G., … Mokkink, L. B. (2020). Preregistering qualitative research: A Delphi study. *International Journal of Qualitative Methods*, *19*, 1609406920976417.

82

Haven, T. L., & van Grootel, L. (2019). Preregistering qualitative research. *Accountability in Research*, *26*(3), 229–244.
82

Hayward, R. D., & Elliott, M. (2014, March). Cross-national analysis of the influence of cultural norms and government restrictions on the relationship between religion and well-being. *Review of Religious Research*, *56*, 23–43.
39

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p–hacking in science. *PLoS Biol*, *13*, e1002106.
93

Healy, K., & Moody, J. (2014). Data visualization in sociology. *Annual Review of Sociology*, *40*, 105–128.
240

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In *An introduction to model-based cognitive neuroscience* (pp. 25–48). Springer Verlag.
240

Heathers, J., Anaya, J., van der Zee, T., & Brown, N. J. (2018). Recovering data from summary statistics: Sample Parameter Reconstruction via Iterative TEchniques (SPRITE). *Manuscript submitted for publication.* Retrieved from `https://peerj.com/preprints/26968.pdf`
259, 263

Heck, D. W. (2019). A caveat on the Savage–Dickey density ratio: The case of computing Bayes factors for regression parameters. *British Journal of Mathematical and Statistical Psychology*, *72*, 316–333.
206

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). Treebugs: An R package for hierarchical multinomial-processing-tree modeling. *Behavior research methods*, *50*, 264–284.
200, 201, 222, 226

Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of Mathematical Psychology*, *91*, 70–87.
139, 148, 159, 170, 224, 270

Heck, D. W., & Wagenmakers, E.-J. (2016). Adjusted priors for Bayes factors involving reparameterized order constraints. *Journal of Mathematical Psychology*, *73*, 110–116.
205

Heinrich, J. (2003). Benefits of blind analysis techniques. *University of Pennsylvania CDF/MEMO/STATISTICS/PUBLIC/6576 Version*, *1*.
90, 93, 96, 97

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. *Manuscript submitted for publication.* Retrieved from `https://psyarxiv.com/nj4es`

116, 132, 280

Hill, T. P. (1995). A statistical derivation of the significant-digit law. *Statistical science*, 354–363.
179, 181

Himawan, K. K., Martoyo, I., Himawan, E. M., Aditya, Y., & Suwartono, C. (2022). Religion and well-being in Indonesia: Exploring the role of religion in a society where being atheist is not an option. *Religion, Brain, & Behaviour*.
67

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*, 1033–1037.
245, 251

Hoekstra, R., & Vazire, S. (2020). Intellectual humility is central to science: Some practices to aspire to. *PsyArXiv*. `https://psyarxiv.com/edh2s/`.
251

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 382–401.
248

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists.* Boca Raton, FL: Chapman & Hall/CRC.
123, 139, 148, 159, 171

Hoijtink, H., Klugkist, I., & Boelen, P. (Eds.). (2008). *Bayesian evaluation of informative hypotheses.* New York: Springer Verlag.
139, 147, 170, 171

Holman, L., Head, M. L., Lanfear, R., & Jennions, M. D. (2015). Evidence of experimental bias in the life sciences: We need blind data recording. *PLoS Biol*, *13*, e1002190.
110

Hoogeveen, S., Haaf, J. M., Bulbulia, J. A., Ross, R. M., McKay, R., Altay, S., ... van Elk, M. (2021). The Einstein effect: Global evidence for scientific source credibility effects and the influence of religiosity. *PsyArXiv*. doi: 10.31234/osf.io/sf8ez
42, 120

Hoogeveen, S., Sarafoglou, A., van Elk, M., & Wagenmakers, E.-J. (in preparation). *A many-analysts approach to the relation between religiosity and well-being: Reflection and conclusion.*
38, 62

Hoogeveen, S., & van Elk, M. (2018). Advancing the Cognitive Science of Religion through Replication and Open Science. *Journal for the Cognitive Science of Religion*, *6*, 158–190. doi: 10.1558/jcsr.39039
42, 120

Horrigan, S. K., Courville, P., Sampey, D., Zhou, F., Cai, S., & Reproducibility Project: Cancer Biology. (2017). Replication study: Melanoma genome sequencing reveals frequent prex2 mutations. *Elife*, e21634.
91

Horton, N. J., & Hardin, J. S. (2015). Teaching the next generation of statistics students to "Think with Data": Special issue on statistics and the under-

graduate curriculum. *The American Statistician*, *69*, 259–265.
268

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, *1*, 70–85.
260, 262, 269, 288

Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., … Brorson, S. (2012). Observer bias in randomised clinical trials with binary outcomes: Systematic review of trials with both blinded and non-blinded outcome assessors. *Bmj*, *344*, e1119.
92

Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Rasmussen, J. V., Hilden, J., … Brorson, S. (2014). Observer bias in randomized clinical trials with time–to–event outcomes: systematic review of trials with both blinded and non–blinded outcome assessors. *International journal of epidemiology*, 937–948.
110

Huijts, T., & Kraaykamp, G. (2011, March). Religious involvement, religious context, and self-assessed health in Europe. *Journal of Health and Social Behavior*, *52*, 91–106.
39

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., … Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, *59*, 944–960.
40

International Committee of Medical Journal Editors. (2019). *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals.* http://www.icmje.org/icmje-recommendations.pdf.
245

iNZight Team. (2020). *iNZight (Version 4.0.2.) [Computer software].* https://inzight.nz.
241

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701.
14, 90, 256

Islam, C.-G., & Lorenz, J. (2022). How to increase the robustness of survey studies. *Religion, Brain, & Behaviour*.
67

Iverson, G. J. (2006). An essay on inequalities and order-restricted inference. *Journal of Mathematical Psychology*, *50*, 215–219.
138

Jang, H., Reeve, J., & Halusic, M. (2016). A new autonomy–supportive way of teaching that increases conceptual learning: Teaching in students' preferred ways. *The Journal of Experimental Education*, *84*, 686–701.
258

JASP Team. (2021). *JASP (Version 0.16.0.0) [Computer software].* https://

jasp-stats.org/.
17, 27, 123, 159

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. In (Vol. 31, pp. 203–222).
139, 170, 205

Jeffreys, H. (1939). *Theory of Probability* (1st ed.). Oxford, UK: Oxford University Press.
18, 35

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.
15, 18

Jevons, W. S. (1874/1913). *The principles of science: A treatise on logic and scientific method.* London: MacMillan.
90

John, L., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532.
14, 109, 256, 269

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*, 3–28.
196, 215

Johnson, T. R. (2007). Discrete choice models for ordinal response variables: A generalization of the stereotype model. *Psychometrika*, *72*, 489–504.
160

Jones, R. (1973). The theory of practical joking – its relevance to physics. In E. Mendoza (Ed.), *A random walk in science: An anthology compiled by the late R L Weber (1914 – 1997)* (p. 14). Bristol: Institute of Physics Publishing.
72

Kambouris, S., Singleton Thorn, F., Van den Akker, O., De Jonge, M., Rüffer, F., Head, A., & Fidler, F. (2020). *Database of articles with Open Science badges: 2020-02-21 snapshot.* Retrieved from https://osf.io/q46r5
75

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, *2*, 389–423.
159

Karabatsos, G., & Sheu, C.-F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement*, *28*, 110–125.
159

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.
15, 139, 170, 197, 205, 224, 285

Kelley, J., & de Graaf, N. D. (1997). National context, parental socialization, and religious belief: Results from 15 nations. *American Sociological Review*, *62*, 639–659. doi: 10.2307/2657431

39

Kerman, J., Gelman, A., Zheng, T., & Ding, Y. (2008). Visualization in Bayesian data analysis. In C. Chen, W. Härdle, & A. Unwin (Eds.), *Handbook of data visualization* (pp. 709 – 724). Berlin: Springer Verlag.
240

Kerr, N. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.
4, 256

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*, 788–799.
269

Kidd, D. C., & Castano, E. (2013). Reading Literary Fiction Improves Theory of Mind. *Science*, *342*, 377–380. doi: 10.1126/science.1239918
14

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., … Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low cost, effective method for increasing transparency. *PLOS Biology*, *14*, e1002456.
6, 72, 232, 252, 256, 269, 282, 288

Kirtley, O. J., Lafit, G., Achterhof, R., Hiekkaranta, A. P., & Myin-Germeys, I. (2021). Making the black box transparent: A template and tutorial for registration of studies using experience-sampling methods. *Advances in Methods and Practices in Psychological Science*, *4*, 1–16.
82

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*, 70–98.
196, 197, 200, 201, 222

Klauer, K. C., Singmann, H., & Kellen, D. (2015). Parametric order constraints in multinomial processing tree models: An extension of Knapp and Batchelder (2004). *Journal of Mathematical Psychology*, *64*, 215–229.
198, 204

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Hofelich Mohr, A., … Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*, 1–15.
252

Klein, R., Ratliff, K., Vianello, M., Adams, R. B., Jr., Bahník, v., Bernstein, M., … Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*, 142–152.
6, 14, 256, 264, 269

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Theory building through replication: Response to commentaries on the "Many Labs" replication project. *Social Psychology*, *45*, 299–311.
15

Klein, R., Vianello, M., Hasselman, F., Adams, B., Adams, R., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability

across sample and setting. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490.
6, 14, 15, 17, 19, 28, 34, 35, 90, 256, 260, 264, 269

Klugkist, I. (2008). Encompassing prior based model selection for inequality constrained analysis of variance. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 53–83). New York: Springer Verlag.
123, 148

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69.
26, 123, 139, 147, 159, 171, 175, 205, 223, 270, 274

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological Methods*, *15*, 281–299.
145, 270

Knapp, B. R., & Batchelder, W. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, *48*, 215–229.
198, 204, 223

Koenig, H. G. (2009). Research on religion, spirituality, and mental health: A review. *The Canadian Journal of Psychiatry*, *54*, 283–291. doi: 10.1177/070674370905400502
38

Koenig, H. G., Al-Zaben, F., & VanderWeele, T. J. (2020, September). Religion and psychiatry: Recent developments in research. *BJPsych Advances*, *26*, 262–272.
41

Koenig, H. G., Hill, T. D., Pirutinsky, S., & Rosmarin, D. H. (2021, January). Commentary on "Does spirituality or religion positively affect mental health?". *The International Journal for the Psychology of Religion*, *31*, 27–44.
41

Koenig, H. G., & Larson, D. B. (2001, January). Religion and mental health: Evidence for an association. *International Review of Psychiatry*, *13*, 67–78. doi: 10.1080/09540260124661
38

Kooperberg, C. (2020). logspline: Routines for logspline density estimation [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=logspline` (R package version 2.1.16)
206

Krypotos, A.-M., Klein, R., & Jong, J. (2022). Resolving religious debates through a multiverse approach. *Religion, Brain, & Behaviour*.
64, 66, 67

Kuhlmann, B. G., Erdfelder, E., & Moshagen, M. (2019). Testing interactions in multinomial processing tree models. *Frontiers in psychology*, *10*, 2364.
160, 204

Kvarven, A., Strømland, E., & Johannesson, M. (2020, April). Comparing meta-

analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434.
41

Ladd, K. L., & Messick, K. J. (2016). A brief history of the psychological study of the role(s) of religion. In *Psychological specialties in historical context: Enriching the classroom experience for teachers and students.* (pp. 204–216).
41

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*, 1–14.
123

Laudy, O. (2006). *Bayesian inequality constrained models for categorical data* (Unpublished doctoral dissertation). Utrecht University.
139, 148, 171

Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review*, *75*, 308–313.
244, 246, 247

Lee, M. D. (2018). Bayesian methods in cognitive modeling. In J. T. Wixted & E.-J. Wagenmakers (Eds.), *The Stevens' handbook of experimental psychology and cognitive neuroscience: Vol. 5 Methodology* (4th ed., pp. 37–84). Oxford: Wiley.
197, 202

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., ... Vandekerckhove, J. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*, 141–153.
159, 208

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*, 114–127.
197, 203, 225

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course.* Cambridge University Press.
18, 30

Levelt, W. J. M., Drenth, P. J. D., & Noort, E. (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel.* `https://pure.mpg.de/rest/items/item_1569964/component/file_1569967/content`.
263

Levine, R., & Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *The American Economic Review*, 942–963.
247

Lewis, C. A., & Cruise, S. M. (2006, June). Religion and happiness: Consensus, contradictions, comments and concerns. *Mental Health, Religion & Culture*, *9*(3), 213–225. doi: 10.1080/13694670600615276
38

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology/Psychologie Canadienne*, *61*, 281–288.
288

Lim, C., & Putnam, R. D. (2010, December). Religion, social networks, and life

satisfaction. *American Sociological Review*, *75*(6), 914–933.
38, 39

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*, 1827–1832.
91

Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016). Research preregistration 101. *APS Observer*, *30*.
91

Lodder, P. (2022). Why researchers should not ignore measurement error and skewness in questionnaire item scores. *Religion, Brain, & Behaviour*.
68

Luo, W., & Chen, F. (2021, December). The salience of religion under an atheist state: Implications for subjective well-being in contemporary China. *Social Forces*, *100*, 852–878.
41

MacCoun, R. (2020). Blinding to remove biases in science and society. In R. Hertwig & C. Engel (Eds.), *Deliberate ignorance: Choosing not to know* (pp. 51–64). Cambridge: MIT Press.
7, 83, 116

MacCoun, R., & Perlmutter, S. (2015). Hide results to seek the truth: More fields should, like particle physics, adopt blind analysis to thwart bias. *Nature*, *526*, 187–190.
6, 7, 83, 89, 90, 92, 116, 262, 289

MacCoun, R., & Perlmutter, S. (2018). Blind analysis as a correction for confirmatory bias in physics and in psychology. In S. O. Lilienfeld & I. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 297–322). John Wiley and Sons.
7, 83, 90, 100, 116, 262

MacKay, D. J. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*, 590–604. doi: 10.1162/neco.1992.4.4.590
15

Maier-Hein, K. H., Neher, P. F., Houde, J.-C., Côté, M.-A., Garyfallidis, E., Zhong, J., . . . Descoteaux, M. (2017, November). The challenge of mapping the human connectome based on diffusion tractography. *Nature Communications*, *8*, 1349.
40

Marsman, M., Schönbrodt, F., Morey, R. D., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A Bayesian bird's eye view of ”Replications of Important Results in Social Psychology”. *Royal Society Open Science*, *4*, 160426.
90

Matejka, J., & Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1290–1294.
240

Matzke, D., Dolan, C. V., Batchelder, W., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in par-

ticipants and items. *Psychometrika*, *80*, 205–235.
196, 197, 199, 200, 201, 222, 226

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, *144*, e1–e15.
90

May, M., & Smilde, D. (2016). Minority participation and well-being in majority Catholic nations: What does it mean to be a religious minority? *Journal of Religion and Health*, *55*(3), 874–894.
39

Mazza, R. (2009). *Introduction to information visualization.* London: Springer Science & Business Media.
242

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan.* Boca Raton: Chapman & Hall/CRC Press.
123

McKenna, H. P. (1994). The Delphi technique: A worthwhile research approach for nursing? *Journal of advanced nursing*, *19*, 1221–1225.
233

McNamara, A. A. (2022). The impact (or lack thereof) of analysis choice on conclusions with Likert data from the Many Analysts Religion Project. *Religion, Brain, & Behaviour*.
68

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*, 235–245.
238

Meng, X.-L., & Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, *91*, 1254–1267.
159

Meng, X.-L., & Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, *11*, 552–586.
158

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*, 831–860.
140, 149, 150, 153, 171, 176, 177

Mertens, G., & Krypotos, A.-M. (2019). Preregistration of analyses of preexisting data. *Psychologica Belgica*, *59*, 338–352.
82

Merton, R. K. (1973). The normative structure of science (1942). In R. K. Merton (Ed.), *The sociology of science: Theoretical and empirical investigations* (pp. 267–278). Chicago, IL: University of Chicago Press.
232, 238, 239

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., … Schuldt, J. P. (2015). Disfluent fonts don't help people

solve math problems. *Journal of Experimental Psychology: General*, *144*, e16–e30.
90

Milewski, M. D., Skaggs, D. L., Bishop, G. A., Pace, J. L., Ibrahim, D. A., Wren, T. A., & Barzdukas, A. (2014). Chronic lack of sleep is associated with increased sports injuries in adolescent athletes. *Journal of Pediatric Orthopaedics*, *34*(2), 129-133.
3, 4, 8

Milkman, K. L., & Berger, J. (2014). The science of sharing and the sharing of science. *Proceedings of the National Academy of Sciences*, *111*, 13642–13649. doi: 10.1073/pnas.1317511111
14

Miller, L. E., & Stewart, M. E. (2011). The blind leading the blind: Use and misuse of blinding in randomized controlled trials. *Contemporary Clinical Trials*, *32*, 240–243.
92

Miller, W., J. amd Schwarz. (2018). Implications of individual differences in on-average null effects. *Journal of Experimental Psychology: General*, *147*, 377–397.
212

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P., . . . Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *Journal of Clinical epidemiology*, *63*, e1–e37.
93, 107

Moher, J., Lakshmanan, B. M., Egeth, H. E., & Ewen, J. B. (2014). Inhibition drives early feature-based attention. *Psychological science*, *25*, 315–324.
93

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., . . . Zwaan, R. A. (2016). The peer reviewers' openness initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, *3*, 150547.
256, 261, 269

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419.
159

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*, 368–378.
206

Morling, B., & Calin-Jageman, R. J. (2020). What psychology teachers should know about open science and the new statistics. *Teaching of Psychology*, *47*, 169–179.
269

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behaviour Research Methods*, *42*, 42–54.
199, 204

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., ... Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*, 501–515. (Publisher: SAGE Publications Sage CA: Los Angeles, CA)
36, 264, 269

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in) equality constrained hypotheses. *Computational Statistics & Data Analysis*, *71*, 448–463.
139, 159, 171

Mulder, J. (2016). Bayes factors for testing order–constrained hypotheses on correlations. *Journal of Mathematical Psychology*, *72*, 104–115.
139, 171

Mulder, J., Gu, X., Olsson-Collentine, A., Tomarken, A., Böing-Messing, F., Hoijtink, H., ... van Lissa, C. (2021). Bfpack: Flexible Bayes factor testing of scientific theories in R. *Journal of Statistical Software*, *100*, 1–63.
139, 159, 171

Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). BIEMS: A Fortran 90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software*, *46*, 1–39.
139, 159, 171

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*, 887–906.
159

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *53*, 530–546.
139, 140, 147, 148, 171, 224

Mulder, J., Wagenmakers, E.-J., & Marsman, M. (in press). A generalization of the Savage-Dickey density ratio for testing equality and order constrained hypotheses. *The American Statistician*.
156

Munafò, M., Nosek, B. A., Bishop, D., Button, K., Chambers, C., Du Sert, N., ... Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.
6, 72, 90, 91, 116, 232, 257, 269

Murphy, J., & Martinez, N. (2022). Quantifying religiosity: A comparison of approaches based on categorical self-identification and multidimensional measures of religious activity. *Religion, Brain, & Behaviour*.
64, 65, 66

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*, 221–229.
6, 72

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.
224

Myung, J., Karabatsos, G., & Iverson, G. (2008). A statistician's view on Bayesian evaluation of informative hypotheses. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 310 – 329). Berlin: Springer Verlag.
148

Myung, J., Karabatsos, G., & Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, *49*, 205–225.
138, 159, 270

National Program Open Science. (2022). *National program open science.* `https://www.openscience.nl/en/national-programme-open-sciencee`.
289

Nature Publishing Group. (2016). Reality check on reproducibility [editorial]. *Nature*, *533*, 437.
269

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, *4*, 39–40.
170, 178

NHB Editorial. (2020). Tell it like it is. *Nature Human Behaviour*, *4*, 1.
251

Nigrini, M. (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud detection* (Vol. 586). Hoboken, New Jersey: John Wiley & Sons.
138, 179

Nigrini, M. J. (2019). The patterns of the numbers used in occupational fraud schemes. *Managerial Auditing Journal*, *34*, 602–622.
184

Nigrini, M. J., & Mittermaier, L. J. (1997). The use of benford's law as an aid in analytical procedures. *Auditing*, *16*, 52.
179

Nosek, B. A. (2019). *Strategy for culture change* [Scientific Blog]. `https://www.cos.io/blog/strategy-for-culture-change`.
289

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425.
6, 91, 232, 252, 256, 269, 282, 288

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*, 815–818.
73, 132, 133, 280

Nosek, B. A., & Lakens, D. (2014). Registered Reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141.
72, 90, 256

Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer*, *31*, 19–21.
6, 72, 73, 82, 84, 116, 280

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*, 615–631. doi: 10.1177/1745691612459058
15

Nuijten, M. B. (2019). *Teaching Open Science. Turn students into skeptics, not cynics.* Conference Talk at ICPS in Paris. Retrieved from `https://static1.squarespace.com/static/5b03f9039f87701b3fd495cf/t/5c87bff753450a6e2f7502a2/1552400378346/Nuijten_TeachingOpenScience.pdf`
264

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.
138, 170, 185, 186, 189, 263

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*, 1596–1618.
6, 72, 73, 287

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.).* London: Arnold.
142, 143

Okulicz-Kozaryn, A. (2010, March). Religiosity and life satisfaction across nations. *Mental Health, Religion & Culture*, *13*, 155–169.
39

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.
6, 14, 33, 35, 90, 110, 256, 264, 269

Open Science MOOC Team. (2019). *Open Science MOOC* [Website]. `https://opensciencemooc.eu`.
257

Oravecz, Z., Vandekerckhove, J., & Batchelder, W. H. (2014). Bayesian Cultural Consensus Theory. *Field Methods*, *26*, 207–222.
286

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776–783.
92

Overstall, A. M., & Forster, J. J. (2010). Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, *54*, 3269–3288.
149, 151, 176

Paloutzian, R. F. (2017). *Invitation to the psychology of religion* (3rd ed.). New York: Guilford Press.
66

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536.
256, 261, 269

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*, 528–530.
6, 116, 256, 269, 280, 337, 341

Patel, C. J., Burford, B., & Ioannidis, J. P. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*, 1046–1058.
247, 248

Pearson, H. I., Lo, R. F., & Sasaki, J. Y. (2022). How do culture and religion interact worldwide? A cultural match approach to understanding religiosity and well-being in the Many Analysts Religion Project Hannah I. Pearson1, *Ronda F. Lo2, Joni Y. Sasaki. *Religion, Brain, & Behaviour*.
64, 65

Peirce, C. S. (1878a). Deduction, induction, and hypothesis. *Popular Science Monthly*, *13*, 470–482.
90

Peirce, C. S. (1878b). The probability of induction. *Popular Science Monthly*, *12*, 705–718.
72

Peirce, C. S. (1883). A theory of probable inference. In C. S. Peirce (Ed.), *Studies in logic* (pp. 126–181). Boston: Little & Brown.
90

Pericchi Guerra, L., Liu, G., & Torres, D. (2008). Objective Bayes factors for informative hypotheses: “completing” the informative hypothesis and “splitting” the Bayes factors. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 131 – 154). Berlin: Springer Verlag.
153

Piwowar, H., Day, R., & Fridsma, D. (2007). Sharing detailed research data is associated with increased citation rate. *Plos ONE*, *2*, e308.
260

Plante, T. G., & Sherman, A. C. (2001). *Faith and health: Psychological perspectives.* Guilford Press.
38

Playfair, W. (1786). *Commercial and political atlas: Representing, by copperplate charts, the progress of the commerce, revenues, expenditure, and debts of england, during the whole of the eighteenth century.* London: Corry. (Republished in Wainer, H. and Spence, I. (eds.), The Commercial and Political Atlas and Statistical Breviary, 2005, Cambridge University Press)
240

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., … Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*, 115.
101, 246

Poloma, M. M., & Pendleton, B. F. (1989). Exploring types of prayer and quality of life: A research note. *Review of Religious Research*, *31*, 46–53. doi: 10.2307/3511023
38, 41

Proulx, T., & Morey, R. D. (2021). Beyond statistical ritual: Rheory in psychological science. *Perspectives on Psychological Science*, *16*, 671–681.
287

Qualtrics. (2021). *Online survey sofware Qualtrics.* Retrieved from `https://www.qualtrics.com`
18, 75

Quintana, F. A. (2006). A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference*, *136*, 2407–2429.
286

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`
124, 275

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute for Educational Research.
273

Rauch, B., Göttsche, M., Brähler, G., & Engel, S. (2011). Fact and fiction in EU-governmental economic data. *German Economic Review*, *12*, 243–255.
138, 181

Rawlinson, H., Talbot, F., Hincks, E., & Oppert, J. (1857). *Inscription of Tiglath Pileser I., king of Assyria, B.C. 1150, as translated by Sir Henry Rawlinson, Fox Talbot, Esq., Dr. Hincks, and Dr. Oppert.* London, UK: J. W. Parker and Son.
249

Reeve, J. (2016). Autonomy–supportive teaching: What it is, how to do it. In W. Liu, J. Wang, & R. Ryan (Eds.), *Building autonomous learners: Perspectives from research and practice using self–determination theory* (pp. 129–152). Singapore: Springer Verlag.
258

Regenwetter, M., Cavagnaro, D. R., Popova, A., Guo, Y., Zwilling, C., Lim, S. H., & Stevens, J. R. (2018). Heterogeneity and parsimony in intertemporal choice. *Decision*, *5*, 63–94.
138, 270

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*, 42–56.
138, 170, 270

Regenwetter, M., & Davis-Stober, C. P. (2012). Behavioral variability of choices versus structural inconsistency of preferences. *Psychological review*, *119*, 408–416.
170

Reichenbach, H. (1938). *Experience and prediction: An analysis of the foundations and the structure of knowledge.* Chicago: The University of Chicago Press.
5

Retraction Watch. (2022). *Retraction Watch Database* [Website]. `http://retractiondatabase.org/RetractionSearch.aspx#?auth%3dWansink%252c%2bBrian`.
5

Riefer, D. M., & Batchelder, W. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318–339.
196

Rijkeboer, M., & van den Hout, M. (2008). A psychologists' view on Bayesian evaluation of informative hypotheses. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 299 – 309). Berlin: Springer Verlag.
170

Rogan, J. (2018). *Joe Rogan Experience #1109 - Matthew Walker* [Podcast]. `https://www.youtube.com/watch?v=pwaWilO_Pig`.
3

Rohrer, J. M. (2018, March). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*, 27–42.
41

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* Appleton-Century-Crofts.
92

Rosmarin, D. H., Pargament, K. I., Pirutinsky, S., & Mahoney, A. (2010, October). A randomized controlled evaluation of a spiritually integrated treatment for subclinical anxiety in the Jewish community, delivered via the Internet. *Journal of Anxiety Disorders*, *24*, 799–808.
41

Ross, R. M., Sulik, J., Buczny, J., & Schivinski, B. (2022). Many analysts and few incentives. *Religion, Brain, & Behaviour*.
68, 69, 288

Rouder, J. N. (2014a). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308.
93

Rouder, J. N. (2014b). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308.
269, 276

Rouder, J. N., Haaf, J. M., Davis-Stober, C. P., & Hilgard, J. (2019). Beyond overall effects: A Bayesian approach to finding constraints in meta-analysis. *Psychological methods*, *24*, 606–621.
286

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.
196, 199

Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process-dissociation model. *Journal of Experimental Psychology: General*, *137*(2), 370–389.
199

Rouder, J. N., & Morey, R. D. (2018). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, 1–5.

Rouder, J. N., Morey, R. D., & Pratte, M. S. (2017). Bayesian hierarchical models of cognition. In W. Batchelder, H. Colonius, E. N. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology: Foundations and methodology* (pp. 504–551). Cambridge University Press.
196

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*, 520–547.
7

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, *2*, 6.
197

Ruiter, S., & van Tubergen, F. (2009). Religious attendance in cross-national perspective: A multilevel analysis of 60 countries. *American Journal of Sociology*, *115*, 863–895. doi: 10.1086/603536
39

Sainz, A., Bigelow, N., & Barwise, C. (1957). On a methodology for the clinical evaluation of phrenopraxic drugs. *The Psychiatric Quarterly*, *31*, 10–16.
93

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., … McLanahan, S. (2020, April). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*, 8398–8403.
42

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., … McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*, 8398–8403. doi: 10.1073/pnas.1915006117
40, 238

Sarafoglou, A., Aust, F., Marsman, M., Wagenmakers, E.-J., & Haaf, J. M. (2021). multibridge: An R package to evaluate informed hypotheses in binomial and multinomial models. *Manuscript submitted for publication*. Retrieved from `https://psyarxiv.com/qk4cy`
224

Sarafoglou, A., Haaf, J. M., Ly, A., Gronau, Q. F., Wagenmakers, E.-J., & Marsman, M. (2021). Evaluating multinomial order restrictions with bridge sampling. *Psychological Methods*.
171, 174, 175, 190, 224

Sarafoglou, A., Hoogeveen, S., Matzke, D., & Wagenmakers, E.-J. (2020). Teaching good research practices: Protocol of a research master course. *Psychology Learning & Teaching*, *19*, 46–59.
269

Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E.-J. (2022). Comparing analysis blinding with preregistration in the many-analysts religion project. *Manuscript submitted for publication*. Retrieved from `https://psyarxiv.com/6dn8f` doi: 10.31234/osf.io/6dn8f
42, 44, 45, 69

Sarafoglou, A., Kovacs, M., Bakos, B. E., Wagenmakers, E.-J., & Aczel, B. (2021). A survey on how preregistration affects the research workflow: Better science but more work 252. *Manuscript submitted for publication.* doi: 10.31234/osf.io/6k5gr
116, 134

Scargle, J. (1999). Publication bias (the "file–drawer problem") in scientific inference. *arXiv*.
256

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological methods*, *26*(1), 103–126.
203

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, 813.
75

Schlattmann, P. (2009). *Medical applications of finite mixture models.* Berlin: Springer.
285

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*, 90–100.
256, 269

Schnuerch, M., Haaf, J. M., Sarafoglou, A., & Rouder, J. (2021). Meaningful comparisons with ordinal-scale items. *Manuscript submitted for publication.* Retrieved from `https://psyarxiv.com/nkydg`
159

Schnuerch, M., Nadarevic, L., & Rouder, J. N. (2020). The truth revisited: Bayesian analysis of individual differences in the truth effect. *Psychonomic Bulletin & Review*, 1–16.
139

Schreiner, M. R., Mercier, B., Frick, S., Wiwad, D., Schmitt, M. C., Kelly, J. M., & Quevedo Pütter, J. (2022). Measurement issues in the Many Analysts Religion Project. *Religion, Brain, & Behaviour*.
68, 288

Schriger, D. L., Sinha, R., Schroter, S., Liu, P. Y., & Altman, D. G. (2006). From submission to publication: A retrospective review of the tables and figures in a cohort of randomized controlled trials submitted to the British Medical Journal. *Annals of Emergency Medicine*, *48*, 750–756.
245

Schulz, K. F., & Grimes, D. A. (2002). Blinding in randomised trials: hiding who got what. *The Lancet*, *359*(9307).
92

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C., van Assen, M. A., ... Luis Uhlmann, E. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, *165*, 228–249.

40

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C. M., & Uhlmann, E. L. (2020). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, *165*, 228–249.

249

Sedransk, J., Monahan, J., & Chiu, H. (1985). Bayesian estimation of finite population parameters in categorical data models incorporating order restrictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *47*, 519–527.

123, 138, 140, 147, 148, 171, 205, 223

Selker, R., van den Bergh, D., Criss, A. H., & Wagenmakers, E.-J. (2019). Parsimonious estimation of signal detection models from confidence ratings. *Behavior Research Methods*. doi: 10.3758/s13428-019-01231-3

32

Seybold, K. S., & Hill, P. C. (2001, February). The role of religion and spirituality in mental and physical health. *Current Directions in Psychological Science*, *10*(1), 21–24. doi: 10.1111/1467-8721.00106

38

Shah, A. K., Mullainathan, S., & Shafir, E. (2012). Some consequences of having too little. *Science*, *338*, 682–685. doi: 10.1126/science.1222426

20, 24

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., . . . Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE*, *8*, e56515.

90

Shariff, A. F., & Norenzayan, A. (2007). God is watching you priming God concepts increases prosocial behavior in an anonymous economic game. *Psychological science*, *18*, 803–809.

103

Silberzahn, R., & Uhlmann, E. L. (2015). Many hands make tight work. *Nature*, *526*, 189.

40, 83, 119

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., . . . Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337–356.

39, 40, 45, 58, 69, 238, 247, 249, 252

Simkin, H. (2020, September). The centrality of events, religion, spirituality, and subjective well-being in Latin American Jewish immigrants in Israel. *Frontiers in Psychology*, *11*, 576402.

41

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

6, 14, 93, 245, 246, 256, 262, 269

Simons, D. (Ed.). (2018). Challenges in making data available [Invited Forum]. *Advances in Methods and Practices in Psychological Science*, *1*.
260, 269

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128.
251

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569. doi: 10.1177/0956797614567341
35

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.
260

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*, 1208–1214.
246, 248

Singmann, H. (2019). *Bayesian cognitive modeling: MPT case studies.* `https://github.com/stan-dev/example-models/blob/master/Bayesian_Cognitive_Modeling/CaseStudies/MPT/MPT_5_Stan.R`. GitHub.
227

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, *56*, 196–201.
224

Smith, E. (2022). Individual-level versus country-level moderation. *Religion, Brain, & Behaviour*.
64, 65

Smith, J. B., & Batchelder, W. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*, 713–731.
199

Smith, J. B., & Batchelder, W. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183.
200

Smith, T. B., McCullough, M. E., & Poll, J. (2003). Religiousness and depression: Evidence for a main effect and the moderating influence of stressful life events. *Psychological Bulletin*, *129*, 614–636. doi: 10.1037/0033-2909.129.4.614
38

Soderberg, C. (2018). Using OSF to share data: A step-by-step guide. *Advances in Methods and Practices in Psychological Science*, *1*, 115–120.
261

Song, Y., Zhao, H., & Wang, T. (2020). An adaptive independence test for microbiome community data. *Biometrics*, *76*, 414–426.
138, 160

Spearman, C. (1904). General intelligence, objectively determined and measured.

American Journal of Psychology, *15*, 201–293.
252

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*, 886–899.
6, 72, 256

Spellman, B. A., Gilbert, E. A., & Corker, K. S. (2018). Open science. In J. Wixted & E.-J. Wagenmakers (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience (4th ed.), Volume 5: Methodology* (pp. 297–322). New York: Wiley.
72, 261

Stan Development Team. (2021). Stan user's guide, version 2.28.0 [Computer software manual]. Retrieved from `http://mc-stan.org/`
151, 190

Stan Development Team. (2022). Multivariate priors for hierarchical models. In *Stan user's guide, version 2.29.0* (pp. 35–43). Retrieved from `http://mc-stan.org/`
227

Stapel, D. (2014). *Faking science: A true story of academic fraud. Translated by Nicholas JL Brown.*

$$jan 14, 2014$$

. `http://beinspired.no/wp-content/uploads/2016/03/FakingScience-20141214.pdf`.
263

Starns, J. J., Cataldo, A. M., Rotello, C. M., Annis, J., Aschenbrenner, A., Bröder, A., . . . Wilson, J. (2019, December). Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Advances in Methods and Practices in Psychological Science*, *2*, 335–349.
40

Stavrova, O. (2015). Religion, self-rated health, and mortality: Whether religiosity delays death depends on the cultural context. *Social Psychological and Personality Science*, *6*, 911–922. doi: 10.1177/1948550615593149
38, 39

Stavrova, O., Fetchenhauer, D., & Schlösser, T. (2013). Why are religious people happy? The effect of the social norm of religiosity across countries. *Social Science Research*, *42*, 90–105. doi: 10.1016/j.ssresearch.2012.07.002
39

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712.
40, 83, 245, 246, 252

Stefan, A. M., Evans, N. J., & Wagenmakers, E.-J. (2020). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*.
197

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis using informed priors. *Behavior Research Methods*, *51*, 1042–1058.

95, 197, 275

Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics*, *25*, 1371–1470.
206

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*, 768–777.
243

Swigart, K. L., Anantharaman, A., Williamson, J. A., & Grandey, A. A. (2020, July). Working while liberal/conservative: A review of political ideology in organizations. *Journal of Management*, *46*, 1063–1091.
41

Symeonidou, N., & Kuhlmann, B. G. (2021). A novel paradigm to assess storage of sources in memory: The source recognition test with reinstatement. *Memory*, 1–17.
195, 198, 204, 205, 207, 212, 215

Szollosi, A., Kellen, D., Navarro, D., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences*, *24*, 94–95.
6, 72

Taichman, D. B., Sahni, P., Pinborg, A., Peiperl, L., Laine, C., James, A., ... Flor, F. (2017). Data sharing statements for clinical trials: A requirement of the International Committee of Medical Journal Editors. *JAMA*, *317*, 2491–2492.
252

Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*, 401–409. doi: 10.1037/h0058700
29

TechCrunch. (2019). *Online translation sofware DeepL.* Retrieved from `https://www.deepl.com/en/translator`
18

Thangaratinam, S., & Redman, C. W. (2005). The delphi technique. *The Obstetrician & Gynaecologist*, *7*, 120–125.
249

The MARP Team. (2022). A many-analysts approach to the relation between religiosity and well-being. *Manuscript submitted for publication.* doi: 10.31234/osf.io/pbfye
64, 117, 118

Thiele, J. E., Haaf, J. M., & Rouder, J. N. (2017). Is there variation across individuals in processing? Bayesian analysis for systems factorial technology. *Journal of Mathematical Psychology*, *81*, 40–54.
214

Thoresen, C. E. (1999, May). Spirituality and health: Is there a relationship? *Journal of Health Psychology*, *4*, 291–300. doi: 10.1177/135910539900400314
38

Tijmstra, J., & Bolsinova, M. (2019). Bayes factors for evaluating latent monotonicity in polytomous item response theory models. *Psychometrika*, *84*, 846–869.
159

Tijmstra, J., Hoijtink, H., & Sijtsma, K. (2015). Evaluating manifest monotonicity using Bayes factors. *Psychometrika*, *80*, 880–896.
138, 159, 270

Tufte, E. R. (1973). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.
241, 242

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*, 1–67.
250

Tukey, J. W. (1977). *Explanatory data analysis.* Reading, MA: Addison–Wesley.
240

Tversky, A., & Gati, I. (1978). Studies of similarity. *Cognition and Categorization*, *1*, 79–98.
20, 24

Uhlenhuth, E. H., Lipman, R. S., Balter, M. B., & Stern, M. (1974). Symptom intensity and life stress in the city. *Archives of General Psychiatry*, *31*, 759–764.
138, 139, 156

Unsworth, N., Redick, T. S., McMillan, B. D., Hambrick, D. Z., Kane, M. J., & Engle, R. W. (2015). Is playing video games related to cognitive abilities? *Psychological Science*, *26*, 759–774.
90

van Assen, M. A., Stoevenbelt, A. H., & van Aert, R. C. (2022). The end justifies all means: Questionable conversion of different effect sizes to a common effect size measure. *Religion, Brain, & Behaviour*.
69

van den Akker, O., van Assen, M., Bakker, M., Enting, M., de Jonge, M., ... Wicherts, J. (2021). Selective hypothesis reporting - preregistration. *Open Science Framework*. Retrieved from `https://osf.io/z4awv`
75

van Dongen, N., van Doorn, J. B., Gronau, Q. F., van Ravenzwaaij, D., Hoekstra, R., Haucke, M. N., ... Wagenmakers, E.-J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician*, *73*, 328–339.
39, 40, 238, 249

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2020, December). Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's ρ. *Journal of Applied Statistics*, *47*, 2984–3006.
58

van Doorn, J., van den Bergh, D., Dablander, F., van Dongen, N., Derks, K., Evans, N. J., ... Wagenmakers, E.-J. (2021). Strong public claims may not reflect researchers' private convictions. *Significance*. Retrieved from `https://psyarxiv.com/pc4ad`

251

van Elk, M. (2021, January). Assessing the religion-health relationship: Introduction to the meta-analysis by Garssen et al., and two commentaries. *The International Journal for the Psychology of Religion*, *31*, 1–3.

41

van Elk, M., Matzke, D., Gronau, Q., Guang, M., Vandekerckhove, J., & Wagenmakers, E.-J. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology*, *6*, 1365.

41

van Lissa, C. J. (2022). Complementing preregistered confirmatory analyses with rigorous, reproducible exploration using machine learning. *Religion, Brain, & Behaviour*.

67, 69

van Lissa, C. J., Peikert, A., & Brandmaier, A. M. (2021). worcs: Workflow for open reproducible code in science [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=worcs` (R package version 0.1.8)

69

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology–A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2-12. doi: https://doi.org/10.1016/j.jesp.2016.03.004

73, 82, 91, 133, 280

Van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., . . . Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, *5*, 2–19.

82

van den Bergh, D., Bogaerts, S., Spreen, M., Flohr, R., Vandekerckhove, J., Batchelder, W. H., & Wagenmakers, E.-J. (2020). Cultural Consensus Theory for the evaluation of patients' mental health scores in forensic psychiatric hospitals. *Journal of Mathematical Psychology*, *98*, 102383.

286

van der Lans, R., Cremers, J., Klugkist, I., & Zwart, R. (2020). Teachers' interpersonal relationships and instructional expertise: How are they related? *Studies in Educational Evaluation*, *66*, 100902.

139, 148

van der Zee, T., Anaya, J., & Brown, N. J. (2017). Statistical heartburn: An attempt to digest four pizza publications from the Cornell food and brand lab. *BMC nutrition*, *3*, 1–15.

4

van Dongen-Boomsma, M., Vollebregt, M. A., Slaats-Willemse, D., & Buitelaar, J. K. (2013). A randomized placebo-controlled trial of electroencephalographic (eeg) neurofeedback in children with attention-deficit/hyperactivity disorder. *The Journal of clinical psychiatry*, *74*, 821–827.

93

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Bayesian

latent-normal inference for the rank sum test, the signed rank test, and Spearman's ρ. *Manuscript submitted for publication.*
29

van Doorn, J., Matzke, D., & Wagenmakers, E.-J. (2020). An in-class demonstration of Bayesian inference. *Psychology Learning & Teaching*, *19*, 36–45.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.
197, 203

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.
197

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra*, *1*, 1–5.
260

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*, 411–417.
6, 256

Veldkamp, C., Bakker, M., van Assen, M., Crompvoets, E., Ong, H., Soderberg, C., ... Wicherts, J. (2017). Restriction of opportunistic use of researcher degrees of freedom in pre-registrations on the open science framework. In *The human fallibility of scientists: Dealing with error and bias in academic research* (pp. 106–133).
133

Veldkamp, C. L., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS ONE*, *9*, e114876.
138

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, *90*, 614–618.
143

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475.
35, 269

Villani, D., Sorgente, A., Iannello, P., & Antonietti, A. (2019). The role of spirituality and religiosity in subjective well-being of individuals with different religious status. *Frontiers in Psychology*, *10*. doi: 10.3389/fpsyg.2019.01525
38

Vinkers, C. H., Tijdink, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific pubmed abstracts between 1974 and 2014: Retrospective analysis. *BMJ*, *351*, h6467.
251

Vogel, V., Prenoveau, J., Kelchtermans, S., Magyar-Russell, G., McMahon, C.,

Ingendahl, M., & Schaumans, C. B. C. (2022). Different facets, different results: The importance of considering the multidimensionality of constructs. *Religion, Brain, & Behaviour*.
64, 65, 66, 67

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review*, *14*, 779–804.
269

Wagenmakers, E.-J. (2019). *Bayesian Spectacles* [Scientific Blog]. `https://www.bayesianspectacles.org`.
263

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R., ... Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*, 917–928.
90, 102

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*, 830–841.
98

Wagenmakers, E.-J., & Dutilh, G. (2016). Seven selfish reasons for preregistration. *APS Observer*, *29*.
73

Wagenmakers, E.-J., Etz, A., Gronau, Q., & Dablander, F. (2018). *The single most prevalent misinterpretation of Bayes' rule* [Scientific Blog]. `https://www.bayesianspectacles.org/the-single-most-prevalent-misinterpretation-of-bayes-rule/`.
18

Wagenmakers, E.-J., Kucharsky, S., & the JASP Team (Eds.). (2020). *The jasp data library.* Amsterdam: JASP Publishing.
252

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.
274

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... Morey, R. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57.
269

Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., ... Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, *5*, 1473–1480.
64, 83, 203

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, *100*, 426–432.
14

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H., & Kievit,

R. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638.
6, 72, 90, 116, 258

Wainer, H. (1984). How to display data badly. *The American Statistician*, *38*, 137–147.
241

Walker, M. (2017). *Why we sleep: Unlocking the power of sleep and dreams.* London: Penguin, UK: Simon and Schuster.
2, 3

Wan, C., Chiu, C.-y., Tam, K.-P., Lee, S.-l., Lau, I. Y.-m., & Peng, S. (2007). Perceived cultural importance and actual self-importance of values in cultural identification. *Journal of Personality and Social Psychology*, *92*, 337–354. doi: 10.1037/0022-3514.92.2.337
43

Wansink, B. (2016). *The grad student who never said "no"* [Scientific Blog]. `https://web.archive.org/web/20170206132854/http://www.brianwansink.com/phd-advice/the-grad-student-who-never-said-no`.
4

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*–values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.
238, 268

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician*, *73*, 1–19.
238, 253

Webb, M. R., & Lee, M. D. (2004). Modeling individual differences in category learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 26, pp. 1440–1445).
199

Weber, S. R., & Pargament, K. I. (2014, September). The role of religion and spirituality in mental health. *Current Opinion in Psychiatry*, *27*, 358–363.
39

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond bar and line graphs: Time for a new data presentation paradigm. *PLoS Biology*, *13*, e1002128.
240, 245

Wessel, I., Albers, C., Zandstra, A. R. E., & Heininga, V. E. (2020). A multiverse analysis of early attempts to replicate memory suppression with the Think/No-think task. *Memory*, *28*, 870–887.
246

Wetzels, R., Grasman, R., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage–Dickey density ratio. *Computational Statistics & Data Analysis*, *54*, 2094–2102.
147, 148, 206

WHOQOL Group. (1998). Development of the World Health Organization WHOQOL-BREF Quality of Life Assessment. *Psychological Medicine*, *28*, 551–558. doi: 10.1017/s0033291798006667

43

Wicherts, J., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728.
260, 269

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 01–17.
133

Wiebe, D. (2009). Religious biases in funding religious studies research? *Religio: Revue pro Religionistiku*, *17*, 125–140.
41

Wilke, C. O. (2019). *Fundamentals of data visualization: A primer on making informative and compelling figures.* Sebastopol, CA: O'Reilly Media.
242

Wilkinson, L. (1999). *The grammar of graphics.* New York: Springer Science & Business Media.
242

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*.
252

Wilson, B. M., & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science*, *1*, 186–197. doi: 10.1177/2515245918767122
33

Witmer, J. (2017). Bayes and MCMC for undergraduates. *The American Statistician*, *71*, 259–264.

World Bank Group. (2017). *World Bank Group - International Development, Poverty, & Sustainability* [Web Page]. http://www.worldbank.org/.
43

World Values Survey. (2010). *Wave 6 Official Aggregate v. 20150418.* worldvaluessurvey.org. Asep/JDS Madrid.
43

Wulff, D. M. (1998, June). Rethinking the rise and fall of the psychology of religion. *Religion in the Making*, 181–202. doi: 10.1163/9789004379039_013
41

Wyman, A. J., & Vyse, S. (2008). Science versus the stars: A double-blind test of the validity of the neo five-factor inventory and computer-generated astrological natal charts. *The Journal of General Psychology*, *135*, 287–300.
271, 273, 274

Yarkoni, T. (2018). *No, it's not the incentives–it's you.* `https://www.talyarkoni.org/blog/2018/10/02/no-its-not-the-incentives-its-you/`.
251

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in

psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122.
95

Yong, E. (2018). Psychology's replication crisis is running out of excuses [Online News Article]. *The Atlantic*. Retrieved from `https://www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/`
260

Zimmer, Z., Jagger, C., Chiu, C.-T., Ofstedal, M. B., Rojo, F., & Saito, Y. (2016, December). Spirituality, religiosity, aging and health in global perspective: A review. *SSM - Population Health*, *2*, 373–381. doi: 10.1016/j.ssmph.2016.04.009
38

Zou, X., Tam, K.-P., Morris, M. W., Lee, S.-l., Lau, I. Y.-M., & Chiu, C.-y. (2009). Culture as common sense: Perceived consensus versus personal beliefs as mechanisms of cultural influence. *Journal of Personality and Social Psychology*, *97*, 579–597. doi: 10.1037/a0016399
39

English Summary

In this dissertation, entitled “Good Research Practices”, I examined research practices and reform ideas aiming to combat the crisis of confidence (Pashler & Wagenmakers, 2012) in psychology. I did so through theoretical contributions and empirical work, I developed statistical methods and practical guidelines for researchers, and I demonstrated how principles of good research can be conveyed to students. In this dissertation, I divided my efforts thematically into three parts, that is: (1) Revealing Hidden Uncertainty in Data Analysis, (2) Multinomial Order-Restrictions, and (3) Guidelines for Good Research Practices.

Part I: Revealing Hidden Uncertainty in Data Analysis

The first part of the dissertation examined current good research practices in psychological science. This part began with a sobering realization: you do not need to be an expert to be able to predict whether a social science study replicates– Chapter 2 illustrated that laypeople too are able to predict replication success with above-chance performance. We suggested that laypeoples’ predictions may be used to quantify intuitive plausibility of empirical effects and hence contribute to efficiently selecting studies for replication research.

Chapter 3 introduced the Many-Analysts Religion Project. In this project, we recruited 120 analysis teams and had them answer the same two research questions based on the same data. Specifically, the teams investigates (1) whether religious people self-report higher well-being, and (2) whether the relation between religiosity and self-reported well-being depends on perceived cultural norms of religion (i.e., whether it is considered normal and desirable to be religious in a given country). For the first research question, all but 3 teams found evidence that religious people report higher levels of well-being and reported positive effect sizes with credible/confidence intervals excluding zero. For the second research question, this was the case for 65% of the teams.

Chapter 4 contained our reflections and conclusions about the Many-Analysts Religion Project. We addressed the issue of theoretical specificity, highlighted some more in-depth observations, discussed methodological concerns raised by the analysis teams, and reflected on our experience of organizing a many-analysts

project.

Chapter 5 described a survey study which identified the benefits and challenges of preregistration from the researcher's perspective. When preregistering a study, researchers describe their hypotheses, as well as the data collection and data analysis plan, as precisely as possible and register them before data collection begins. This research practice increases the transparency of the research process and forces researchers to adhere to the empirical cycle, that is, to clearly distinguish predicted findings from possible chance findings. The study showed that preregistration has benefits beyond safeguarding the adherence to the empirical cycle, including the improvement of the overall project quality. This survey, however, also illustrated some of the challenges that come with preregistration, such as the increase of the overall project duration and work related stress.

Chapter 6 introduced analysis blinding as an addition or possible alternative to preregistration. In analysis blinding, research teams develop their data analysis plan based on a blinded version of the data, that is, data for which collaborators or independent researchers have removed any information that might jeopardize a fair statistical analysis (e.g., treatment effects or differences between two experimental conditions). The chapter discussed how analysis blinding can be applied in experimental psychology. Specifically, it introduced different methods of analysis blinding, offered recommendations for blinding of popular experimental designs, and introduced the design for an online blinding protocol.

Following this idea, Chapter 7 compared the reported efficiency and convenience of preregistration and analysis blinding in the context of the Many-Analysts Religion Project. The recruited teams answered the same research questions based on the same data either preregistering their analysis or using analysis blinding. The study concluded that analysis blinding does not mean less work but approximately the same amount, but researchers can still benefit from the method since they can plan more appropriate analyses from which they deviate less frequently.

Part II: Multinomial Order-Restrictions

The second part of the dissertation explored how researchers can integrate their theory-based knowledge into statistical models and presents statistical procedures for testing ordinal hypotheses (i.e., hypotheses about increasing or decreasing trends). Our research on this topic has focused primarily on categorical data analysis which is based on a multinomial distribution.

Chapter 8 described a Bayesian technique with which researchers can evaluate ordinal hypotheses concerning the distribution of multinomial proportions. Whenever researchers formulate ordinal hypotheses that entail expectations about increasing or decreasing trends they must rely on methods that are relatively inefficient and computationally expensive. To address this problem, we developed a bridge sampling routine that allows an efficient evaluation of these hypotheses for multinomial variables. An empirical example showed that bridge sampling outperforms current Bayesian methods in terms of accuracy and efficiency.

In order to maximize the accessibility of the proposed bridge sampling routine, we developed the user-friendly software package **multibridge** which was introduced in Chapter 9. The `R` software package implements the bridge sampling routine for multinomial variables and independent binomial variables. The chapter described the core functions in **multibridge** and illustrated its use with two examples one concerning the prevalence of statistical reporting errors across eight different psychology journals, the other using **multibridge** to reveal corrupt statistics on the deficit and debt of the Greek government in the years before the financial crisis.

Chapter 10 applied the evaluation of ordinal hypotheses in the context of multinomial processing tree (MPT) models. In psychology, MPT models are used to test sophisticated theories on memory, judgement and decision making, and reasoning. The chapter highlighted how researchers can refine their Bayesian MPT modeling practices by adequately capturing their theory in the model and testing their ordinal expectations.

Part III: Guidelines for Good Research Practices

The third part of the dissertation provided concrete suggestions on how to facilitate the uptake of good research practice among researchers. This part addressed this challenge on two levels: educating researchers and training students.

Chapter 11 presented the Transparency Checklist which allows researchers in social and behavioural sciences to improve and document the transparency of research reports. The initial set of items in the Transparency Checklist was developed in collaboration with 45 behavioural and social science journal editors-in-chief and associate editors, as well as 18 open-science advocates. The final checklist spans the four study components: preregistration, methods, results and discussion as well as data, code and materials availability. Responses to the checklist items can be submitted along with a manuscript, providing reviewers, editors and, eventually, readers with critical information about the research process allowing them to evaluate the robustness of a finding.

Chapter 12 discussed seven concrete statistical practices which embody the current aspirations in the social and behavioural sciences to increase transparency and reproducibility. These practices are (1) visualizing data; (2) quantifying inferential uncertainty; (3) assessing data preprocessing choices; (4) reporting multiple models; (5) involving multiple analysts; (6) interpreting results modestly; and (7) sharing data and code. We discussed the benefits and limitations of each practice and provided guidelines for its adoption.

The remaining two chapters showed how the concepts of good research practices can be incorporated into the methodological training of students. Chapter 13 described the content of the graduate course "Good Research Practices" which we have designed and taught at the University of Amsterdam. This course gives a general introduction into the crisis of confidence as well as recent methodological reforms proposed in psychological science, such as direct and conceptual replication studies, preregistration, and the public sharing of data and code.

Chapter 14 presented a Bayesian research project that we conducted with undergraduate psychology students. This project aimed to (1) convey the basic mathematical concepts of Bayesian inference; (2) have students experience the entire empirical cycle including collection, analysis, and interpretation of data and

(3) teach both the philosophy behind good research practices and the practical skills needed to apply them.

Nederlandse Samenvatting

In dit proefschrift, getiteld “Good Research Practices”, onderzocht ik onderzoekspraktijken en hervormingsideeën om de vertrouwenscrisis in de psychologie te bestrijden (Pashler & Wagenmakers, 2012). Ik heb dit gedaan door middel van theoretische bijdragen en empirisch werk; ik heb statistische methoden en praktische richtlijnen voor onderzoekers ontwikkeld, en laten zien hoe de principes van goed onderzoek kunnen worden overgebracht op studenten. In dit proefschrift heb ik mijn werk thematisch opgedeeld in drie delen, namelijk: (1) Onthullen van verborgen onzekerheid in data-analyse, (2) Multinomiale volgorde-restricties, en (3) Richtlijnen voor goede onderzoekspraktijken.

Part I: Onthullen van verborgen onzekerheid in data-analyse

Het eerste deel van het proefschrift onderzocht de actuele praktijken van goed onderzoek in de psychologische wetenschap. Dit deel begon met een ontnuchterende constatering: je hoeft geen expert te zijn om te kunnen voorspellen of een sociaal-wetenschappelijke studie repliceert - hoofdstuk 2 illustreerde dat ook leken in staat zijn om replicatiesucces beter dan kansniveau te voorspellen. Wij suggereerden dat de voorspellingen van leken kunnen worden gebruikt om de intuïtieve plausibiliteit van empirische effecten te kwantificeren en zo bij te dragen aan een efficiënte selectie van studies voor replicatieonderzoek.

Hoofdstuk 3 introduceerde het Many-Analysts Religion Project. In dit project wierven we 120 analyseteams om te onderzoeken (1) of religieuze mensen een hoger niveau van welzijn zelfrapporteren, en (2) of de relatie tussen religiositeit en zelfgerapporteerd welzijn afhangt van waargenomen culturele normen van religie (d.w.z. of het in een bepaald land als normaal en wenselijk wordt beschouwd om religieus te zijn). Voor de eerste onderzoeksvraag rapporteerden op 3 na alle teams positieve effectgroottes waarbij nul buiten de geloofwaardheids-/ betrouwbaarheidsintervallen viel. Voor de tweede onderzoeksvraag was dit het geval voor 65% van de teams.

Hoofdstuk 4 bevatte onze overwegingen en conclusies over het Many-Analysts Religion Project. We behandelden de kwestie van theoretische specificiteit, belichtten enkele diepgaandendere observaties, bespraken methodologische problemen, en reflecteerden op onze ervaring met het organiseren van een project met veel analisten.

In hoofdstuk 5 wordt een enquêteonderzoek beschreven waarin de voordelen en knelpunten van preregistratie vanuit het perspectief van de onderzoeker in kaart zijn gebracht. Uit de studie bleek dat preregistratie extra voordelen heeft naast het waarborgen van de naleving van de empirische cyclus, waaronder de verbetering van de algehele projectkwaliteit. Dit onderzoek illustreerde echter ook enkele van de uitdagingen die gepaard gaan met preregistratie, zoals de verlenging van de totale projectduur en werkgerelateerde stress.

In hoofdstuk 6 werd analyseblindering geïntroduceerd als aanvulling op of mogelijk alternatief voor preregistratie. Het hoofdstuk besprak hoe analyseblindering kan worden toegepast in de experimentele psychologie. In het bijzonder werden verschillende methoden van analyseblindering geïntroduceerd, werden aanbevelingen gedaan voor blindering van veelgebruikte experimentele opzetten, en werd het ontwerp voor een online blinderingsprotocol geïntroduceerd.

Naar aanleiding van dit idee werden in hoofdstuk 7 de gerapporteerde efficiëntie en het gemak van preregistratie en analyseblindering vergeleken in het kader van het Many-Analysts Religion Project. De deelnemende teams beantwoordden dezelfde onderzoeksvragen op basis van dezelfde data door hun analyse vooraf te registreren of door gebruik te maken van analyseblindering. In de studie werd geconcludeerd dat analyseblindering niet minder werk betekent, maar ongeveer evenveel, maar dat onderzoekers toch baat kunnen hebben bij de methode omdat zij adequatere analyses kunnen plannen waarvan zij minder vaak hoeven af te wijken.

Part II: Multinomiale Orde-Beperkingen

In het tweede deel van het proefschrift werd besproken hoe op theoretische kennis kan worden gekwantificeerd in statistische modellen en werden statistische technieken geïntroduceerd om ordinale hypothesen te testen in de context van categorische data-analyse.

In hoofdstuk 8 is een Bayesiaanse techniek beschreven waarmee onderzoekers ordinale hypotheses over de verdeling van multinomiale verhoudingen kunnen evalueren. Telkens wanneer onderzoekers ordinale hypothesen formuleren met verwachtingen over stijgende of dalende trends, zijn zij aangewezen op methoden die relatief inefficiënt zijn. Om dit probleem aan te pakken hebben wij een bridge-sampling methode ontwikkeld die een efficiënte evaluatie van deze hypothesen voor multinomiale variabelen mogelijk maakt. Een empirisch voorbeeld toont aan dat bridge-sampling beter presteert dan de huidige Bayesiaanse methoden met betrekking tot nauwkeurigheid en efficiëntie.

Om de toegankelijkheid van de voorgestelde bridge-sampling methode te maximaliseren, ontwikkelden we het gebruiksvriendelijke `R`-pakket **multibridge** dat in hoofdstuk 9 werd geïntroduceerd. Het pakket implementeert de bridge-sampling methode voor multinomiale variabelen en onafhankelijke binomiale variabelen. Het hoofdstuk beschreef de kernfuncties van **multibridge** en illustreerde het gebruik ervan met twee voorbeelden, waarvan er een betrekking had op de prevalentie van statistische rapportagefouten in acht verschillende psychologische tijdschriften.

In hoofdstuk 10 wordt de evaluatie van ordinale hypothesen toegepast in de context van multinomiale procesboom (MPT) modellen. In de psychologie worden MPT-modellen gebruikt om geavanceerde theorieën over geheugen, besluitvorming, en redeneren te testen. Het hoofdstuk belichtte hoe onderzoekers hun Bayesiaanse MPT-modellen kunnen verfijnen door hun theorie adequaat vast te leggen in het model en hun ordinale verwachtingen te testen.

Part III: Richtlijnen Voor Goede Onderzoekspraktijken

Het derde deel van het proefschrift bevatte concrete suggesties over hoe de invoering van goede onderzoekspraktijken onder onderzoekers kan worden vergemakkelijkt. In dit deel werd deze uitdaging op twee niveaus aangepakt: het opleiden van onderzoekers en het trainen van studenten.

Hoofdstuk 11 presenteerde de Transparantiechecklist waarmee onderzoekers in de sociale en gedragswetenschappen de transparantie van onderzoeksartikelen kunnen verbeteren en documenteren. De eerste reeks items van de Transparantiechecklist werd geëvalueerd door 45 hoofdredacteuren en adjunct-redacteuren van tijdschriften in gedrags- en sociale wetenschappen, alsook door 18 voorvechters van Open Wetenschap. De uiteindelijke checklist omvat de vier onderzoeksonderdelen: preregistratie, methoden, resultaten en discussie, alsmede de beschikbaarheid van data, code en materialen. Reacties op de items van de checklist kunnen samen met een manuscript worden ingediend, waardoor reviewers, redacteuren en uiteindelijk ook lezers kritische informatie krijgen over het onderzoeksproces die nodig is om de robuustheid van een bevinding te evalueren.

In hoofdstuk 12 werden zeven concrete statistische praktijken besproken die het huidige streven naar meer transparantie en reproduceerbaarheid in de sociale en gedragswetenschappen belichamen. Deze praktijken zijn (1) het visualiseren van data; (2) het kwantificeren van inferentiële onzekerheid; (3) het beoordelen van data-voorbewerkingskeuzes; (4) het rapporteren van meerdere modellen; (5) het betrekken van meerdere analisten; (6) het bescheiden interpreteren van resultaten; en (7) het delen van data en code. Wij bespraken de voordelen en beperkingen van elke praktijk en gaven richtlijnen voor de toepassing ervan.

De resterende twee hoofdstukken lieten zien hoe de concepten van goede onderzoekspraktijken kunnen worden opgenomen in de methodologische training van studenten. Hoofdstuk 13 beschreef de inhoud van de masteropleiding cursus “Good Research Practices” die wij hebben ontworpen en gegeven aan de Universiteit van Amsterdam. Deze cursus geeft een algemene inleiding in de vertrouwenscrisis en recente methodologische hervormingen die zijn voorgesteld in de psychologische wetenschap, zoals directe en conceptuele replicatiestudies, preregistratie, en het openbaar delen van data, code en analyseplannen.

In hoofdstuk 14 presenteerden we een Bayesiaans onderzoeksproject dat we uitvoerden met studenten psychologie. Dit project had tot doel (1) de wiskundige basisconcepten van Bayesiaanse inferentie over te brengen; (2) studenten de hele empirische cyclus te laten ervaren, inclusief het verzamelen, analyseren en interpreteren van data, en (3) zowel de filosofie achter goede onderzoekspraktijken te onderwijzen als de praktische vaardigheden die nodig zijn om ze toe te passen.

Deutsche Zusammenfassung

In dieser Dissertation mit dem Titel “Gute Forschungspraktiken” habe ich Forschungspraktiken und Reformideen untersucht, die darauf abzielen, die Vertrauenskrise in der Psychologie zu bekämpfen. Im Rahmen meiner Forschung habe ich sowohl theoretische als auch empirische Arbeiten verfasst, statistische Methoden und praktische Leitlinien entwickelt und aufgezeigt, wie die Grundsätze guter Forschung an Studierende vermittelt werden können. Thematisch lässt sich die Dissertation in drei Teile gliedern: (1) Die Aufdeckung verborgener Unsicherheit in der Datenanalyse, (2) Ordinale Beschränkungen bei Multinomialverteilungen und (3) Richtlinien für gute Forschungspraktiken.

Teil I: Aufdeckung verborgener Unsicherheit in der Datenanalyse

Im ersten Teil meiner Dissertation untersuchte ich aktuelle gute Forschungspraktiken in der psychologischen Wissenschaft. Dieser Teil begann mit einer ernüchternden Erkenntnis: Man muss kein Experte sein, um vorhersagen zu können, ob eine sozialwissenschaftliche Studie repliziert werden kann. Kapitel 2 zeigte, dass auch Laien in der Lage sind, den Replikationserfolg sozialwissenschaftlicher Studien überdurchschnittlich gut vorherzusagen. Wir schulgen vor, dass solche Laienvorhersagen verwendet werden können, um die intuitive Plausibilität empirischer Effekte zu erfassen, was wiederum zur effektiven Auswahl von Replikationsstudien beitragen kann.

Kapitel 3 stellte das “Multi-Analysten Projekt zum Thema Religion” vor. Im Rahmen dieses Projekts rekrutierten wir 120 Analyseteams und liessen sie dieselben zwei Forschungsfragen auf der Grundlage derselben Daten beantworten. Konkret beantworteten die Analyseteams die Frage, (1) ob religiöse Menschen nach eigenen Angaben ein höheres Wohlbefinden haben, und (2) ob die Beziehung zwischen Religiosität und subjektiven Wohlbefinden von den wahrgenommenen kulturellen Normen zu Religion abhängt (d.h., ob es in einem bestimmten Land als normal und wünschenswert angesehen, wird religiös zu sein). In Bezug auf die erste Forschungsfrage fanden alle bis auf 3 Teams positive Evidenz dafür, dass religiöse Menschen ein höheres Wohlbefinden haben, und meldeten Effektgrößen mit

Kredibilitäts-/Konfidenzintervallen größer Null. Bei der zweiten Forschungsfrage war dies bei 65% der Teams der Fall.

Kapitel 4 enthielt unsere überlegungen und Schlussfolgerungen über das Multi-Analysten Projekt zum Thema Religion. In diesem Kapitel befassten wir uns mit der Frage der theoretischen Spezifizität, hoben tiefer gehende Erkenntnisse zu beiden Forschungsfragen hervor, setzten uns kritisch mit einigen der methodischen Bedenken auseinander, die von einigen Analyseteams geäußert wurden und berichteten über unsere eigenen Erfahrungen mit der Organisation dieses Projektes.

In Kapitel 5 wurde eine Umfragestudie beschrieben, bei der wir Forscher zu den Vorteilen und Herausforderungen der Präregistrierung befragten. Bei der Präregistrierung werden Hypothesen, sowie der Datenerhebungs- und Datenanalyseplan so genau wie möglich festgehalten und vor Beginn der Datenerhebung registriert. Diese Forschungspraktik erhöht die Transparenz des Forschungsprozesses und zwingt Forscher dazu, den empirischen Zyklus einzuhalten, das heißt, vorhergesagte Befunde klar von möglichen Zufallsbefunden zu unterscheiden. Unsere Studie zeigte, dass die Präregistrierung nicht nur den Vorteil hat, dass Forscher den empirischen Zyklus einhalten, sondern Forscher gaben auch an, dass die Präregistrierung die Qualität ihres Forschungsprojekts insgesamt verbessert. Die Umfrage veranschaulichte jedoch auch einige Nachteile der Präregistrierung, beispielsweise, dass diese Forschungspraktik die Projektdauer verlängert und den arbeitsbedingten Stress erhöht.

In Kapitel 6 wurde die Analyseverblindung als Ergänzung oder mögliche Alternative zur Präregistrierung vorgestellt. Bei der Analyseverblindung entwickeln Forschungsteams ihren Datenanalyseplan auf der Grundlage einer verblindeten Version der Daten, das heißt, Daten, bei denen Mitarbeiter oder unabhängige Forscher alle Informationen entfernt haben, die eine faire Auswertung gefährden (z.B. Behandlungseffekte oder Unterschiede zwischen zwei experimentellen Bedingungen). In diesem Kapitel wurde erörtert, wie die Analyseverblindung in der experimentellen Psychologie angewendet werden kann. Insbesondere wurden verschiedene Methoden der Analyseverblindung vorgestellt, Empfehlungen gegeben, wie gängiger Datenstrukturen in der Psychologie effektiv verblindet werden können, sowie ein Online-Verblindungsprotokoll präsentiert.

Diesem Gedanken folgend wurde in Kapitel 7 die berichtete Effizienz und Zweckmäßigkeit von Präregistrierung und Analyseverblindung im Rahmen des Multi-Analysten Projektes zum Thema Religion verglichen. In diesem Projekt beantwortete die eine Hälfte der Analyseteams die Forschungsfragen, indem sie ihre Analysen präregistrierten, die andere Hälfte indem sie Analyseverblindung anwendeten. Unsere Studie kam zu dem Schluss, dass die Analyseverblindung nicht wie erwartet weniger Arbeitsaufwand bedeutet, sondern ungefähr den gleichen Aufwandsaufwand mit sich bringt wie Präregistrierung. Die Analyseverblindung konnten dennoch von der Analyseverblindung profitieren. Unsere Studie zeigte, dass Forscher die diese Praktik anwendeten, angemessenere Analysen planen konnten, von denen sie später weniger häufig abweichen mussten.

Teil II: Ordinale Beschränkungen bei Multinomialverteilungen

Im zweiten Teil der Dissertation wurde erörtert, wie Forscher ihr theoriebasiertes Wissen in statistische Modellen integrieren können und stellt statistische Verfahren zur Prüfung ordinaler Hypothesen vor (d.h., Hypothesen zu aufsteigenden oder absteigenden Trends). Unsere Forschung zu diesem Thema konzentrierte sich dabei in erster Linie auf die Analyse von kategorischen Daten, die auf einer multinomialen Verteilung beruhen.

In Kapitel 8 beschrieben wir eine Bayes'sche Technik, mit der Forscher ordinale Hypothesen über die Verteilung multinomialer Wahrscheinlichkeiten testen können. Zurzeit müssen Forscher, wann immer sie ordinale Hypothesen formulieren, auf statistische Methoden zurückgreifen, die relativ ineffizient und rechenaufwändig sind. Um dieses Problem zu lösen, haben wir eine Bridge-Sampling-Routine entwickelt, die eine effiziente Auswertung dieser Hypothesen für multinomiale Variablen ermöglicht. Ein empirisches Beispiel hat gezeigt, dass die Bridge-Sampling-Routine verglichen zu den derzeitigen Bayes'schen Methoden sowohl genauere Ergebnisse liefert als auch effizienter ist.

Um unsere Bridge-Sampling-Routine für Forscher in der Psychologie zugänglich zu machen, haben wir das benutzerfreundliche Software-Paket **multibridge** entwickelt, das in Kapitel 9 vorgestellt wurde. Das Software-Paket wurde in der Programmiersprache `R` geschrieben und implementiert die Bridge-Sampling-Routine für multinomialverteilte Variablen und unabhängig-binomialverteilte Variablen. In diesem Kapitel wurden die Kernfunktionen von **multibridge** beschrieben und seine Verwendung anhand von zwei Beispielen veranschaulicht. Im ersten Beispiel verwendeten wir **multibridge**, um die Häufigkeit von Fehlern in der Berichterstattung statistische Ergebnisse (z.B., verursacht durch Tippfehler) in acht verschiedenen psychologischen Fachzeitschriften zu testen. Im zweiten Beispiel verwendeten wir **multibridge**, um korrupte Statistiken zu Defizit und Schuldenstand des griechischen Staates in den Jahren vor der Finanzkrise aufzudecken.

In Kapitel 10 testeten wir ordinale Hypothesen im Rahmen von Multinomialen Prozessbaum Modellen. In der Psychologie werden Multinomialen Prozessbaum Modelle verwendet, um anspruchsvolle Theorien zu Gedächtnisprozessen, Urteils- und Entscheidungsfindungsprozessen und Prozessen rationalen Denkens zu testen. Das Kapitel zeigte auf, wie Forscher ihre Modellierungspraktiken verfeinern können, wenn sie ihr Wissen zu psychologischen Prozessen direkt in ihre statistische Modelle integrieren und ihre Erwartungen als ordinale Hypothesen formulieren und testen.

Teil III: Richtlinien für gute Forschungspraktiken

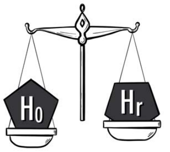
Im dritten Teil der Dissertation machen wir konkrete Vorschläge, wie sich Wissen zu gute Forschungspraktiken leichter verbreiten lassen. Dieser Teil geht die Herausforderung auf zwei Ebenen an: der Schulung von Forschern und der Ausbildung von Studenten.

In Kapitel 11 wurde die Transparenz-Checkliste vorgestellt, mit deren Hilfe Forscher in den Sozial- und Verhaltenswissenschaften die Transparenz ihrer Forschungsberichte verbessern und angemessen dokumentieren können. Die aufgeführten Fragen wurden in Zusammenarbeit mit 45 Chefredakteuren und Mitherausgebern sozial- und verhaltenswissenschaftlicher Fachzeitschriften sowie von 18 Verfechtern der offenen Wissenschaft entwickelt. Sie dient dazu, wichtige Informationen über den Forschungsprozess zu liefern und ist thematisch in vier Teile gegliedert: Präregistrierung, Forschungsmethoden, Forschungsergebnis und Diskussion, sowie Verfügbarkeit von Daten, Code und Materialien. Forscher können den ausgefüllten Fragenkatalog zusammen mit ihrem Manuskript in Fachzeitschriften einreichen und liefern damit den Gutachtern, Redakteuren und den Lesern notwendige Informationen, die eine Bewertung zur Robustheit ihrer Forschungsergebnisse zulässt.

In Kapitel 12 wurden sieben konkrete statistische Verfahren beschrieben, die die derzeitigen Bestrebungen in den Sozial- und Verhaltenswissenschaften zur Erhöhung von Transparenz und Reproduzierbarkeit verkörpern. Diese Praktiken sind (1) Datenvisualisierung; (2) Erwähnung von Messungenauigkeiten in statistischen Ergebnissen; (3) Einschätzung des Einflusses verschiedener Methoden der Datenvorverarbeitung auf das Ergebnis; (4) Einschätzung des Einflusses verschiedener gleichwertiger statistischer Modelle auf das Ergebnis; (5) Einbeziehung mehrerer Analyseteams; (6) die bescheidene Interpretation von Ergebnissen; und (7) das zur Verfügung stellen von Daten und Code. Wir erörterten die Vorteile und Einschränkungen der einzelnen Verfahren und geben Richtlinien für ihre Anwendung.

In den beiden verbleibenden Kapiteln wird aufgezeigt, wie die Konzepte guter Forschungspraktiken in die methodische Ausbildung von Studenten integriert werden können. Kapitel 13 beschrieb den Inhalt des Kurses “Gute Forschungspraktiken”, den wir im Research-Masterstudiengang Psychologie an der Universität Amsterdam entworfen und unterrichtet haben. Dieser Kurs gibt eine allgemeine Einführung in die Vertrauenskrise in der Psychologie sowie in die jüngsten methodologischen Reformen, um die Krise zu bekämpfen, wie beispielsweise direkte und konzeptionelle Replikationsstudien, Präregistrierung und das zur Verfügung stellen von Daten und Code.

In Kapitel 14 wurde ein Bayes’sches Forschungsprojekt vorgestellt, das wir mit Psychologiestudenten im Bachelorstudiengang durchgeführt haben. Dieses Projekt zielte darauf ab, (1) die grundlegenden mathematischen Konzepte der Bayes’schen Inferenz zu vermitteln, (2) Studenten den gesamten empirischen Zyklus erleben zu lassen, einschließlich der Erhebung, Analyse und Interpretation von Daten und (3) Studenten sowohl die Philosophie zu lehren die hinter guten Forschungspraktiken steckt, als auch die praktischen Kompetenzen zu lehren, die für deren Anwendung erforderlich sind.



Contributions

Part I: Revealing Hidden Uncertainty in Data Analysis

Chapter 2:

Conceptualization: S.H., A.S., and E.-J.W.
Data Curation: S.H. and A.S.
Formal Analysis: S.H. and A.S.
Funding Acquisition: A.S. and E.-J.W.
Investigation: S.H., A.S., and E.-J.W.
Methodology: S.H., A.S., and E.-J.W.
Project Administration: S.H., A.S., and E.-J.W.
Supervision: S.H., A.S., and E.-J.W.
Validation: S.H. and A.S.
Visualization: S.H. and A.S.
Writing - Original Draft: S.H., A.S., and E.-J.W.
Writing - Review & Editing: S.H., A.S., and E.-J.W.

Chapter 3:

Conceptualization: S.H., A.S., M.v.E., and E.-J.W.
Data Curation: S.H. and A.S.
Formal Analysis: S.H., A.S., B.A., Y. Aditya, A.J.A., P.J.A., S. Alzahawi, Y. Amir, F.-V.A., O.K.A., Q.D.A., A. Baimel, M.B.-I., M. Balsamo, S.B., F.B., M. Becerra, B.B., J. Beitner, T. Bendixen, J.B.B., M.I.B., J. Billingsley, T. Bortolini, H.B., A. Bret, F.L.B., J. Brown, C.C.B., J. Buczny, J. Bulbulia, S. Caballero, L.C., C.L.C., M.E.G.V.C., S.J.C., S. Claessens, M.C.P., A.B.C., D.L.C., S. Czoschke, C.C., E.D.D.U., Ö.D., A.D.R., H.D., J.D.R., Y.A.d.V., K.K.D., B.J.D., D.J.D., J.K.D., T.D., L.D., M.D., Y.D., T.E., P.A.E., A.E., C.T.E., S. Farahmand, H.F., M.F., A.A.F., K.F., R.F., D.F.-T., Z.F., S. Frick, L.K.F., D.G., E. Gerdin, L.G., O.G., E. Gielens, V.G., H.R.H., I.H., P.H.P.H., C.E.H., B.C.H., L.E.H., M.I., H.I., M.L.I., C.-G.I., O.I., D.I., B.J., K.A.J., J.J., J.A.K., K.K.H., E.K., B.A.K., L.A.K., S. Kelchtermans, J.M.K., R.A.K., B.K., M.L.K., M.K., D.K., J.K., S.

Kritzler, A.-M.K., T.K., T.L.L., R.L., G.A.L.F., A.L., B.J.L., R.F.L., P.L., J.L., P.Ł., A.L.L., E.M., G.M.M.-R., M. Maier, D.R.M., N.M., M. Martinie, I.M., S.E.M., A.L.M., P. McAleer, T.M., M. McCullough, C.M.M., A.A.M., K.K.M., B. Mercier, P. Mitkidis, B. Monin, J.W.M., D.M., J. Morgan, J. Murphy, G.M., C.N., T.N., L.N., N.H., G.N., P.N., A.N., M.B.N., A.O.-C., L.O., Y.G.P., J.O.P., H.I.P., H.P., H.K.P., M. Pinus, S.P., V.P., M. Porubanova, M.J.P., J.M.P., M.A.P., J.P., C.P., B.G.P., J.Q.P., M.L.R., G.R., A. Roberts, L.M.R.L., R.M.R., P.R., N.R., S.-M.K.S., J.Y.S., C. Schaumans, B. Schivinski, M.C.S., S.A.S., M. Schnuerch, M.R.S., V.S., S. Sebben, S.C.S., B. Seryczyńska, U.S., M. Simsek, W.W.A.S., E.R.S., W.J.S., M. Späth, C. Spörlein, W.S., A.H.S., S. Stuber, J.S., C. Suwartono, S. Syropoulos, B. Szaszi, P.S., L.T., R.T.T., B.T., C.M.T., J.T., S.D.T., A.-M.U., R.C.M.V.A., M.A.L.M.v.A., P.V.C., O.R.V.d.A., I.V.d.C., J.V.d.N., N.N.N.v.D., C.J.V.L., V.v.M., D.v.R., C.J.J.v.Z., L.A.V., B. Većkalov, B. Verschuere, M.V., F.V., A.V., V.V., L.V.D.E.V., S. Watanabe, C.J.M.W., K.W., S. Wiechert, Z.W., M.W., C.V.O.W., D.W., X.Y., D.J.Y., O.Y., N.Z., Y.Z., and J.Z.
Funding Acquisition: A.S., M.v.E., E.-J.W., S. Altay, N.L., R.M., and R.M.R.
Investigation: S.H., A.S., M.v.E., E.-J.W., S. Altay, T. Bendixen, R.B., K.H., R.M., L.Q., A. Rabelo, J.E.R., R.M.R., R.W., and D.X.
Methodology: S.H., A.S., M.v.E., and E.-J.W.
Project Administration: S.H., A.S., M.v.E., and E.-J.W.
Supervision: M.v.E. and E.-J.W.
Validation: S.H. and A.S.
Visualization: S.H., A.S., and P.J.A.
Writing - Original Draft: S.H., A.S., M.v.E., and E.-J.W.
Writing - Review & Editing: P.A.E., P.H.P.H., R.M., C.M.M., J. Murphy, T.N., J.E.R., R.M.R., S.C.S., B.T., R.C.M.V.A., M.A.L.M.v.A., and C.J.M.W.

Chapter 4:

Conceptualization: S.H., A.S., M.v.E., and E.-J.W.
Funding Acquisition: A.S. and M.v.E.
Project Administration: S.H. and A.S.
Supervision: M.v.E. and E.-J.W.
Validation: S.H., A.S., M.v.E., and E.-J.W.
Writing - Original Draft: S.H., A.S., M.v.E., and E.-J.W.
Writing - Review & Editing: S.H., A.S., M.v.E., and E.-J.W.

Chapter 5:

Conceptualization: M.K., E.-J.W., and B.A.
Data Curation: A.S., M.K., and B.B.
Formal Analysis: A.S., M.K., and B.B.
Funding Acquisition: A.S. and E.-J.W.
Investigation: A.S., M.K., E.-J.W., and B.A.
Methodology: A.S., M.K., E.-J.W., and B.A.
Project Administration: A.S., E.-J.W., and B.A.

Software: M.K.
Supervision: A.S., E.-J.W., and B.A.
Validation: A.S., M.K., E.-J.W., and B.A.
Visualization: A.S.
Writing - Original Draft: A.S., M.K., E.-J.W., and B.A.
Writing - Review & Editing: A.S., M.K., E.-J.W., and B.A.

Chapter 6:

Conceptualization: G.D., A.S., and E.-J.W.
Funding Acquisition: G.D., A.S., and E.-J.W.
Methodology: G.D., A.S., and E.-J.W.
Project Administration: G.D.
Supervision: E.-J.W.
Validation: G.D., A.S., and E.-J.W.
Visualization: G.D. and A.S.
Writing - Original Draft: G.D., A.S., and E.-J.W.
Writing - Review & Editing: G.D., A.S., and E.-J.W.

Chapter 7:

Conceptualization: A.S., S.H., and E.-J.W.
Data Curation: A.S. and S.H.
Formal Analysis: A.S. and S.H.
Funding Acquisition: A.S. and E.-J.W.
Investigation: A.S., S.H., and E.-J.W.
Methodology: A.S., S.H., and E.-J.W.
Project Administration: A.S., S.H., and E.-J.W.
Supervision: A.S., S.H., and E.-J.W.
Validation: A.S. and S.H.
Visualization: A.S. and S.H.
Writing - Original Draft: A.S., S.H., and E.-J.W.
Writing - Review & Editing: A.S., S.H., and E.-J.W.

Part II: Multinomial Order-Restrictions

Chapter 8:

Conceptualization: A.S., J.H., E.-J.W., and M.M.
Data Curation: A.S.
Formal Analysis: A.S., J.H., and A.L.
Funding Acquisition: A.S., E.-J.W., and M.M.
Methodology: A.S., J.H., A.L., Q.F.G., E.-J.W., and M.M.
Project Administration: A.S.

Supervision: J.H., A.L., E.-J.W., and M.M.
Validation: A.S., J.H., E.-J.W., and M.M.
Visualization: A.S.
Writing - Original Draft: A.S., J.H., A.L., Q.F.G., E.-J.W., and M.M.
Writing - Review & Editing: A.S., J.H., A.L., Q.F.G., E.-J.W., and M.M.

Chapter 9:

Conceptualization: A.S., F.A., M.M., and J.H.
Data Curation: A.S.
Formal Analysis: A.S. and J.H.
Funding Acquisition: A.S., M.M., and E.-J.W.
Investigation: A.S., F.A., and J.H.
Methodology: A.S., M.M., E.-J.W., and J.H.
Project Administration: A.S.
Software: A.S., F.A., and J.H.
Supervision: F.A., M.M., E.-J.W., and J.H.
Validation: A.S., F.A., M.M., E.-J.W., and J.H.
Visualization: A.S., F.A., and M.M.
Writing - Original Draft: A.S., F.A., M.M., E.-J.W., and J.H.
Writing - Review & Editing: A.S., F.A., M.M., E.-J.W., and J.H.

Chapter 10:

Conceptualization: B.K. and J.H.
Data Curation: B.K.
Formal Analysis: A.S. and F.A.
Funding Acquisition: A.S. and J.H.
Investigation: A.S., F.A., and J.H.
Methodology: A.S., F.A., and J.H.
Project Administration: B.K. and J.H.
Supervision: B.K. and J.H.
Validation: A.S., B.K., F.A., and J.H.
Visualization: A.S. and F.A.
Writing - Original Draft: A.S., B.K., F.A., and J.H.

Part III: Guidelines for Good Research Practices

Chapter 11:

Conceptualization: B.A., B.S., A.S., Z.K., and E-J.W.
Formal Analysis: B.A., B.S., A.S., Z.K., and E-J.W.
Project Administration: B.A., B.S., A.S., Z.K., and E-J.W.
Software: Š.K.

Writing - Original Draft Preparation: B.A., B.S., A.S., Z.K., and E-J.W.
Preparation and Conclusion of Checklist Items: D.B., C.D.C., A.F., A.G., M.A.G., J.P.I., E.J., K.J., S.K., S.O.L., D.S.L., C.C.M., M.M., B.R.N., H.P., D.R.S., D.J.S., and J.M.W.
Evaluation of Checklist Items: D.A., N.D.A., J.A., H.A., M.D.B., G.C.B., C.B., A.A.B., W.B., T.W.B., C.C., A.S.C., J.C., C. Clifton, R.M.C., M.C., F.C., N.C., J. Crawford, E.A.C., J. Curtin, R.E., S.F., P.F., M.F., W.F., A.M.F., M.G.G., R.G-S., D.P.G., R.L.G., L.L.H., F.H.G., D.I., J.K., D.L., G.D.L., W.B.M., L.M., B.N., J.P., C.S., and S.V.
Writing - Review & Editing: B.A., B.S., A.S., Z.K., E-J.W., D.B., C.D.C., A.F., A.G., M.A.G., J.P.I., E.J., K.J., S.K., S.O.L., D.S.L., C.C.M., M.M., B.R.N., H.P., D.R.S., D.J.S., J.M.W., D.A., N.D.A., J.A., H.A., M.D.B., G.C.B., C.B., A.A.B., W.B., T.W.B., C.C., A.S.C., J.C., C. Clifton, R.M.C., M.C., F.C., N.C., J. Crawford, E.A.C., J. Curtin, R.E., S.F., P.F., M.F., W.F., A.M.F., M.G.G., R.G-S., D.P.G., R.L.G., L.L.H., F.H.G., D.I., J.K., D.L., G.D.L., W.B.M., L.M., B.N., J.P., C.S., and S.V.

Chapter 12:

Conceptualization: E.-J.W., A. Sarafoglou, and B.A.
Project Administration: B.A.
Writing - Original Draft Preparation: E.-J.W., A. Sarafoglou, C.A., J.A., Š.B., N.v.D., R.H., D.M., D.v.R., A. Sluga, J.T., and B.A.
Writing - Review & Editing: E.-J.W., A. Sarafoglou, S.A., C.A., J.A., Š.B., N.v.D., R.H., D.M., D.v.R., A. Sluga, F.S., J.T., and B.A.

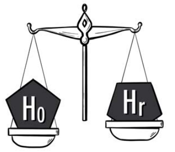
Chapter 13:

Conceptualization: A.S., S.H., D.M., and E.-J.W.
Funding Acquisition: A.S., D.M., and E.-J.W.
Project Administration: A.S., S.H., D.M., and E.-J.W.
Supervision: E.-J.W.
Writing - Original Draft: A.S., S.H., D.M., and E.-J.W.
Writing - Review & Editing: A.S., S.H., D.M., and E.-J.W.

Chapter 14:

Conceptualization: A.S., A.v.d.H., T.D., J.C., E.-J.W., and M.M.
Data Curation: A.S., A.v.d.H., T.D., and J.C.
Formal Analysis: A.v.d.H., T.D., and J.C.
Funding Acquisition: A.S., E.-J.W., and M.M.
Investigation: A.v.d.H., T.D., and J.C.
Methodology: A.S., A.v.d.H., T.D., J.C., E.-J.W., and M.M.
Project Administration: A.S., E.-J.W., and M.M.
Software: A.S., A.v.d.H., T.D., and J.C.
Supervision: A.S., E.-J.W., and M.M.
Validation: A.S., and E.-J.W.

Visualization: A.S., A.v.d.H., T.D., and J.C.
Writing - Original Draft: A.S., A.v.d.H., T.D., J.C., E.-J.W., and M.M.
Writing - Review & Editing: A.S., A.v.d.H., T.D., J.C., E.-J.W., and M.M.



Acknowledgements

I would like to thank my supervisor EJ, and my co-supervisors Julia and Maarten, for their encouragement and patience throughout the past five years. My PhD journey was a wonderful experience and I owe that mostly to you. I am grateful for all that you have taught me, the insightful conversations we had, and the amazing projects we worked on together. I am proud to call myself a member of the Bayesian lab.

I am honored to have my thesis evaluated by excellent experts in methodology, cognitive modeling, and metascience: Beatrice, Jeff, Dora, Denny, and Tom. I would like to thank the members of the Doctorate Committee for agreeing to this task and the time they spend reading and assessing my dissertation. I would also like to thank my paranymphs Selina and Suzanne; on the one hand for your help in successfully completing my dissertation, but also for your friendship. You both helped me to stay mentally stable, cheered me up when I was feeling grumpy or stressed and made me feel at home in the department and in Amsterdam. Thank you Selina for being my friend now for more than 10 years. We have already experienced so much together: from our first vacation together in Italy, to being roommates in the 12-person shared apartment in Tübingen, then as roommates in Amsterdam, to your wedding, the birth of your beautiful daughter, and our PhD graduations. What an adventure! I look forward to experiencing many more with you. Thank you Suzanne for all the projects we started and finished together. The projects with you are by far my favorite: it's so much fun working with you! It's fun to brainstorm and exchange ideas with you, but also to complain about the project and life in general when it's not going well. The numerous times I visited you in Utrecht are among my favorite PhD memories. I hope we can keep up our tradition of publishing one paper a year together (and possibly ending up in newspapers and on the radio) for a while!

Nur mit Hilfe der Unterstützung meiner Familie war es mir möglich meine akademische Laufbahn nachzugehen. Ich möchte mich vor allem bei meinen Eltern und meiner Schwester bedanken. Dafür, dass ihr mich immer in meinen Ambitionen unterstützt habt und mich ermutigt habt, meinen eigenen Weg zu gehen – auch wenn es euch sicherlich nicht leicht fällt, wenn euer Kind/die Schwester so weit weg von zu Hause wohnt. Ich bin immer wieder davon beeindruckt mit

welcher Selbstverständlichkeit ihr mir zur Seite steht, dass ihr zum Beispiel nicht zögert mit vollgepacktem Anhänger über sechs Stunden im Auto zu fahren um mir bei meinen Umzügen zu helfen und meine Wohnung so zu renovieren, dass sie für mich wirklich zu einem Hause wird. Ich habe das grosse Glück Teil einer herzlichen, lustigen und offenen Familie zu sein. Ich geniesse es zu Ostern, im Sommer und an Weihnachten nach Hause zu kommen und Zeit mit der Familie zu verbringen, mit meinen Tanten und Onkeln, Thia Nitsa, Thio Dimitri und Tante Petra, Tante Inge, Tante Doris und Onkel Peter und Onkel Marc und Ceilia sowie mit all meinen Cousins und Cousinen und deren Familien. Ich möchte mich auch bei meiner angeheirateten Familie bedanken, vor allem bei meinem Schwager Georg. Ich kann mir keinen besseren und liebevolleren Ehemann für meine Schwester wünschen. Ich möchte mich auch bei Bruno, Ursel und Johannes dafür bedanken, dass sie mich so herzlich in ihrem Familienkreis aufgenommen haben.

I would like to thank my oldest friends Natascha (I known her since kindergarden), Kirstin (elementary school) and Nadja (high school). I enjoy reminiscing and exchanging news with you each year - and realizing time and again how alike we still are after all these years. I would also like to thank the friends I made during my undergraduate studies in Tübingen, especially Eva and Daniel, Laura and Jörg, Patricia, Dustin and Luisa, Theresa and Harald. Our annual vacation trips and my visits to Tübingen allow me to completely relax, recharge my creative streak, and maintain a strong bond with my homeland and Swabia.

I also want to express my gratitude to my friends and colleagues in Amsterdam. I would like to thank Jason, Juno, Ticho and Barbara for all the dinners, game nights and ping pong tournaments that keep brightening up my weekends. I want to thank Akash and Tim for the countless house parties and barbecue events. To the gang Selina, Don, Frantisek, Angelika, Alex, Charlotte, Quentin, Koen, Waldemar, and Julia for the soccer events and movie nights and to Mauricio for the interesting discussions on politics, academia and weightlifting.

Working on my thesis would be only half as much fun without my office mates. I would like to thank Alexander, Abe, and Joost who shared the office with me at the beginning of my PhD and who left behind a plethora of toys for the coming REC G 0.30 generations. I warmly thank Johnny and Quentin who provided plenty of mental support in the most work intensive and stressful months of my PhD. I am also thankful to Hannes and Ivar –the latest addition to the office– who fill the room with positive energy whenever they arrive.

Hard work cannot be achieved without a good hearty lunch with my colleagues. Thank you Frantisek; your sociability is the reason why we now have such a wonderful lunch group. Thank you Samuel, Ria, Julius, Maarten, Karoline, Adam, Michelle, and Jill. I enjoy our (sometimes controversial and charged) conversations that help me to take my mind off of work for an hour and lift my spirit.

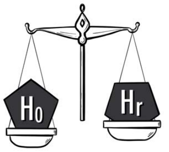
Many thanks goes also to the remaining members of the Bayesian lab, to Frederik, Simon, Alexander, Udo, Fabian, and Lukas, for their expertise, insights and knowledge and for the fun times we had outside of work, for instance, during lab activities, dinners, dancing events, and baby parties. Special thanks goes also to the JASP team Bruno, Joris, Frans, and Tim who never let me down in my countless Mac related problems. I would like to extend my gratitude to all the remaining colleagues in the psychological methods department for their extensive knowledge and guidance over the last years.

I would also like to thank my friends and colleagues from other departments. Thank you Maien for improving my work-life balance and inspiring me to incorporate creative activities into my free time. Thank you Jessica for the nice conversations and shared dinners. Thank you Tamar, Jakub, and Ingmar –my future colleagues– for giving me the opportunity to apply my knowledge to a new research field.

In my PhD journey, I was fortunate enough to meet and work with many great meta-scientists. A special thanks goes to my long-term collaborators Balazs and Marton with whom I have worked on a variety of meta-scientific projects. Your productivity and great visions are a source of motivation for me and I look forward to working with you in the future and to finally visit you in Budapest. I would also like to thank those who have worked hard over the last few years to change the scientific culture in the Netherlands. Engaging the research community in the current developments on Open Science are a crucial step for improving our field and is something very close to my heart. I would like to thank the leading figures of community building, Loek and Anita who have brought open science communities to life, to Anna, Vera, Raul, and Sander who are advancing OSCs on a national level, and Michiel and Anne for building up a national reproducibility network.

Over the course of my PhD I developed a passion for weightlifting, which is fueled primarily by the unique community I have found at my gym DutchStrength. My weightlifting family has taken a special place in my life. I would like to say a big thank you to my coach Randolf. Having you as my coach has helped me so much to develop myself as an athlete. The raw energy you bring to our training sessions have helped to boost my motivation, allowing me to quickly put hard days at work behind me. Ingeborg, thank you for always being such a caring friend. I enjoyed spending hours and hours with you in cafes and restaurants talking about our life plans, training, career, and dating. I am looking forward to moving to the country side with you someday, spending my days with dog training, hiking, and training. Thank you, Martin for being so expressive and infectiously happy at all times. When you are around it feels like a great weight falls off my shoulders and I perceive my own life in a much more positive way. Of course I am also thanking the rest of the Weights and Plates gang Anastasia, Luuk, Elia, Senne, Aleks, Justin, Sonia, Fred, Sven, Nhi, David, Dian, and Safdar for all the fun trainings, testosterone loaded max-out sessions, tasty dinners (and cookies!), and the countless discussions about who has recently gotten bigger quads. Thank you, Parizad; you are my role model when it comes to being a strong, independent, and successful woman, and I hope that one day I can be more like you. Thank you Armin for your warm personality and relaxed attitude towards life. Thank you Ian, Irene and Mario; you were the first people I met in DutchStrength. You welcomed me into the community with open arms and are the main reasons why I decided to stay in DutchStrength. I also want to thank our head coach Tom. Our weightlifting community could not thrive without you.

Last but not least, I would like to thank my neighbors for letting me “borrow” their cat Otto for all these years.



Publications

1. **Sarafoglou, A.**, Hoogeveen, S., & Wagenmakers, E. J. (2022). Comparing analysis blinding with preregistration in the Many-Analysts Religion Project. Manuscript submitted for publication. `https://psyarxiv.com/6dn8f/`

2. **Sarafoglou, A.**, Kovacs, M., Bakos, B. E., Wagenmakers, E. J., & Aczel, B. (2021). A survey on how preregistration affects the research workflow: Better science but more work. Manuscript submitted for publication. `https://psyarxiv.com/6k5gr`

3. **Sarafoglou, A.**, Aust, F., Wagenmakers, E. J., & Haaf, J. M. (2021). multibridge: An R Package To Evaluate Informed Hypotheses in Binomial and Multinomial Models. Manuscript submitted for publication. `https://psyarxiv.com/qk4cy`

4. Schnuerch, M., Haaf, J. M., **Sarafoglou, A.**, & Rouder, J. (2021). Meaningful comparisons with ordinal-scale items. Manuscript submitted for publication. `https://psyarxiv.com/nkydg`

5. The MARP Team. (in press). A Many-Analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behaviour.* `https://psyarxiv.com/pbfye/`

6. Hoogeveen, E. J., **Sarafoglou, A.**, van Elk, M., & Wagenmakers, E. J. (in press). Many-Analysts Religion Project: Reflection and conclusion. *Religion, Brain & Behaviour.* `https://psyarxiv.com/wm3zd`

7. **Sarafoglou, A.**, van der Heijden, A., Draws, T., Cornelisse, J., Wagenmakers, E. J., & Marsman, M. (2022). Combine statistical thinking with open scientific practice: A protocol of a Bayesian research project. *Psychology Learning & Teaching.* Advance online publication. `https://doi.org/10.1177/14757257221077307`.

8. **Sarafoglou, A.**, Haaf, J. M., Ly, A., Gronau, Q. F., Wagenmakers, E. J., & Marsman, M. (2021). Evaluating multinomial order restrictions with bridge sampling. *Psychological Methods.* Advance online publication. `https://doi.org/10.1037/met0000411`

9. Wagenmakers, E. J., **Sarafoglou, A.**, Aarts, S., Albers, C., Algermissen, J., Bahník, Š, van Dongen, N., Hoekstra, R., Moreau, D., van Ravenzwaaij, D., Sluga, A., Stanke, F., Tendeiro, J., & Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, *5*, 1473–1480.

10. van Doorn, J., van den Bergh, D., Boehm, U., Dablander, F., Derks, K., Draws, T., Evans, N. J., Gronau, Q. F., Hinne, M., Kucharsky, S., Ly, A., Marsman, M., Matzke, D., Komarlu Narendra Gupta, A. R., **Sarafoglou, A.**, Stefan, A., Voelkel, J. G., & Wagenmakers, E.–J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*, 813–826.

11. Aczel, B., Szaszi, B., Nilsonne, G., van den Akker, O. R., Albers, C. J., van Assen, M. A., Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., Bringmann, L. F., Busch, N. A., Caruyer, E., Cataldo, A. M., Cowan, N., Delios, A., van Dongen, N. N., Donkin, C., van Doorn, J. B., Dreber, A., Dutilh, G., Egan, G. F., Gernsbacher, M. A., Hoekstra, R., Hoffmann, S., Holzmeister, F., Johannesson, M., Jonas, K., Kindel, A., Kirchler, M., Kunkels, Y. K., Lindsay, D. S., Mangin, J.–F., Matzke, D., Munafò, M. R., Newell, B. R., Nosek, B. A., Poldrack, R., van Ravenzwaaij, D., Rieskamp, J., Salganik, M., **Sarafoglou, A.**, Schonberg, T., Schweinsberg, M., Shanks, D., Silberzahn, R., Simons, D. J., Spellman, B., Starns, J. J., St-Jean, S., Uhlmann, E. L., Wicherts, J. M., Wagenmakers, E.–J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *ELife*, *10*, e72185.

12. Armeni, K., Brinkman, L., Carlsson, R., Eerland, A., Fijten, R., Fondberg, R., Heininga, V. E., Heunis, S., Koh, W. Q., Masselink, M., Moran, N., Baoill, A. O., **Sarafoglou, A.**, Schettino, A., Schwamm, H., Sjoerds, Z., Teperek, M., van den Akker, O. R., van't Veer, A., & Zurita-Milla, R., Zurita-Milla, R. (2021). Towards wide-scale adoption of open science practices: The role of open science communities. *Science and Public Policy*, *48*, 605–611.

13. Dutilh, G., **Sarafoglou, A.**, & Wagenmakers, E. J. (2021). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, *198*, 5745–5772.

14. Ly, A., Stefan, A., van Doorn, J., Dablander, F., van den Bergh, D., **Sarafoglou, A.**, Kucharsky, S., Derks, K., Gronau, Q. F., Raj, A., Boehm, U., van Kesteren, E.–J., Hinne, M., Matzke, D., Marsman, M., & Wagenmakers, E.–J. (2020). The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the p value hypothesis test. *Computational Brain & Behavior*, *3*, 153–161.

15. Aczel, B., Szaszi, B., **Sarafoglou, A.**, Kekecs, Z., Kucharský, Š., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., Gernsbacher, M. A., Ioannidis, J. P., Johnson, E., Jonas, K., Kousta, S., Lilienfeld, S. O., Lindsay, D. S., Morey, C. C., Munafò, M., Newell, B. R., Pashler, H., Shanks, D. R., Simons,

D. J., Wicherts, J. M., Albarracin, D., Anderson, N. D., Antonakis, J., Arkes, H., Back, M. D., Banks, G. C., Beevers, C., Bennett, A. A., Bleidorn, W., Boyer, T. W., Cacciari, C., Carter, A. S., Cesario, J., Clifton, C., Conroy, R. M., Cortese, M., Cosci, F., Cowan, N., Crawford, J., Crone, E. A., Curtin, J., Engle, R., Farrell, S., Fearon, P., Fichman, M., Frankenhuis, W., Freund, A. M., Gaskell, M. G., Giner–Sorolla, R., Green, D. P., Greene, R. L., Harlow, L. L., Hoces de la Guardia, F., Isaacowitz, D., Kolodner, J., Lieberman, D., Logan, G. D., Mendes, W. B., Moersdorf, L., Nyhan, B., Pollack, J., Sullivan, C., Vazire, S., & Wagenmakers, E. J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, *4*, 4–6.

16. Hoogeveen, S., **Sarafoglou, A.**, & Wagenmakers, E. J. (2020). Laypeople can predict which social-science studies will be replicated successfully. *Advances in Methods and Practices in Psychological Science*, *3*, 267–285.

17. van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E.–J., Derks, K., Dablander, F., Gronau, Q. F., Kucharsky, S., Komarlu Narendra Gupta, A. R., **Sarafoglou, A.**, Voelkel, J. G., Stefan, A., Ly, A., Hinne, M., Matzke, D., & Wagenmakers, E.–J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année Psychologique/Topics in Cognitive Psychology*, *120*, 73–96.

18. **Sarafoglou, A.**, Hoogeveen, S., Matzke, D., & Wagenmakers, E. J. (2020). Teaching good research practices: Protocol of a research master course. *Psychology Learning & Teaching*, *19*, 46–59.

19. Dahrendorf, M., Hoffmann, T., Mittenbühler, M., Wiechert, S. M., **Sarafoglou, A.**, Matzke, D., & Wagenmakers, E. J. (2019). "Because it is the Right Thing to Do": Taking Stock of the Peer Reviewers' Openness Initiative. *Journal of European Psychology Students*, *11*, 15–20.

20. Wagenmakers, E. J., Dutilh, G., & **Sarafoglou, A.** (2018). The creativity-verification cycle in psychological science: New methods to combat old idols. *Perspectives on Psychological Science*, *13*, 418–427.

21. Gronau, Q. F., **Sarafoglou, A.**, Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.–J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97.