# TO BELIEVE OR NOT TO BELIEVE

## Open Science and Replication in the Psychology of Religion

Suzanne Hoogeveen

# To Believe or Not to Believe

## Open Science and Replication in the Psychology of Religion

Suzanne Hoogeveen

ISBN 978-94-6421-985-2
This thesis was typeset with LaTeX using the Dissertate class developed by Jordan Suchow and adjusted by Hoàng-Ân Lê.
Printed by Ipskamp Printing B.V.
Cover design by Viktor Beekman

The work described in this thesis is supported by the John Templeton Foundation [60663].
The opinions expressed in this dissertation are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

# To Believe or Not to Believe

## Open Science and Replication in the Psychology of Religion

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.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op vrijdag 10 februari 2023, te 11.00 uur

door

Suzanne Hoogeveen

geboren te Rotterdam

## Promotiecommissie

| | | |
|---|---|---|
| *Promotores:* | prof. dr. E.M. Wagenmakers | Universiteit van Amsterdam |
| | dr. M. van Elk | Universiteit Leiden |
| *Overige leden:* | prof. dr. F. van Harreveld | Universiteit van Amsterdam |
| | prof. dr. H.L.J. van der Maas | Universiteit van Amsterdam |
| | prof. dr. R. McKay | University of London - Royal Holloway |
| | dr. J.M. Haaf | Universiteit van Amsterdam |
| | dr. O. Stavrova | Tilburg University |

Faculteit der Maatschappij- en Gedragswetenschappen

UNIVERSITY OF AMSTERDAM

# Contents

# List of Abbreviations

| | |
|---|---|
| ACC | anterior cingulate cortex |
| BF | Bayes factor |
| CCRRP | Cross-cultural Religious Replication Project |
| CCT | Compensatory Control Theory |
| EEG | electroencephalogram |
| fMRI | functional magnetic resonance imaging |
| MARP | Many-Analysts Religion Project |
| ML4 | Many-Labs 4 |
| TMT | Terror Management Theory |

*If you thought that science was certain – well, that is just an error on your part.*

Richard P. Feynman

*In God we trust. All others must bring data.*
W. Edwards Deming

# 1 General Introduction

Within the scientific context, skepticism is among the highest virtues. Scientists should live by the Royal Society's motto: 'Nullius in verba' - 'Take nobody's word for it'. Rather than faithfully trusting information at face value, the critical scientist should question assumptions and doubt claims unless the empirical evidence convince them otherwise.

But the credibility of psychological science has been shaken over the last decade (Pashler & Wagenmakers, 2012; Simmons et al., 2011). Numerous examples have surfaced showing that in fact we may have put too much faith in the scientific practice. Questionable research practices such as selective reporting of (dependent) variables, publication bias, low statistical power, and post-hoc hypothesizing have threatened the validity of scientific research (G. Francis, 2012; Ioannidis, 2005; John et al., 2012). The sobering realization of the "crisis of confidence" in psychological science seems to be that we cannot even always trust empirical evidence presented in research articles published in respected academic journals, as many findings have turned out not to be replicable.

Against the backdrop of the crisis in psychology, but also inspired by the rapid emergence of reforms for good research practices, my PhD project on assessing the replicability and applying open science in the psychology of religion was started, which resulted in the current dissertation. In this introduction, I will first illustrate the problems with some examples of perhaps amusing yet disturbing cases of flawed articles that somehow slipped through the nets of the scientific peer review system. Then, after dealing with the concerning cases, we can shift focus to the optimism inspiring initiatives and reforms appearing on the horizon.

## 1.1 Science Gone Awry: Examples from the Psychology of Religion

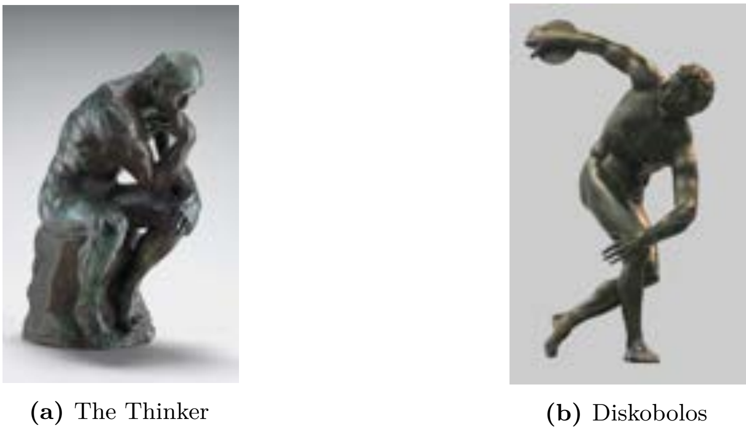
In 2015, psychology professor John Decety and colleagues published a study in *Current Biology* claiming that across six countries, religiosity is negatively related to altruism in children: religious children are less likely to share and more likely to punish than their non-religious counterparts (Decety et al., 2015). The article made headlines around the world and was covered by many newspapers, including the Guardian, the Economist, and the Dutch NRC. The results, however, raised doubts among some scholars of religion. Upon inspection of the data, Shariff, Willard, Muthukrishna, et al. (2016) identified a crucial statistical error: rather than treating the variable country as a categorical predictor with six levels, the authors included it as a continuous covariate of 'country-ness', making Canada twice as much a country as the US. Correctly including country of origin as a categorical variable made the association disappear completely, as the differences in generosity between children were explained by between-country differences rather than religious affiliation.[1] As a result of these contrasting corrected findings, the authors formally retracted the article. Perhaps ironically, the lead author who discovered the mistake in the altruism study was later involved in a retracted paper himself; a societally-sensitive study published in *Psychological Science* claimed that declines in religiosity predict an increase in violent crime, except for countries with a high average level of IQ (Clark et al., 2020). In this case, voiced issues regarding the reliability of the national IQ data and homicide rates, the method of imputing missing values (and perhaps the huge outrage on Twitter), led the authors to retract the article.

While these cases are certainly alarming, they involve clear errors that were discovered and eventually led to retraction of the publications. There is arguably a much larger and more worrisome grey area of studies that are not obviously flawed, yet cannot be replicated. Consider for instance the now infamous priming study on religion and analytical thinking published in *Science* (Gervais & Norenzayan, 2012). This study found that people became less religious after looking at a picture of Rodin's *The Thinker* compared to people looking at the picture of Myron's *Diskobolos* ('discus thrower'), the rationale being that they were primed to think analytically, instead of intuitively, which would suppress religious beliefs (see Figure 1.1 for the pictures). This effect could not be repeated in a large-scale replication project (Camerer et al., 2016). In retrospect, the idea that deep-rooted religious beliefs can be changed by viewing a sculpture of a guy seemingly engaged in deep thoughts for 10 seconds may sound implausible. And indeed, in Chapter 3 of this dissertation we show that laypeople –correctly– considered this study very unlikely to be replicated successfully and even the authors themselves have acknowledged that the study was 'outright silly' (NPR, 2018).[2]

The tendency to strive for significant, positive effects is deep-rooted in psychological research. While the examples above may be somewhat outlandish, there are also more subtle illustrations closer to home. I remember speaking at a conference about a study on the effect of placebo brain stimulation on subjective performance experiences in a cognitive task and neural responses to performance errors (Hoogeveen, Schjoedt, et al., 2018). We found that expected improvement through brain stimulation increased the error-related negativity, a brain marker sensitive to the expectedness of making an error. At the time of the conference, we had just analysed the data of a follow-up study including the same setup but with a larger sample, in which we failed to replicate this neural effect (van Elk et al., 2020). While presenting these new contradictory results, I noticed that the discussion with the audience mainly focused on justifying the first positive result: people offered suggestions for what might have gone wrong in the follow-up study rather than acknowledging that the first study might have

[1] Unfortunately, the original data are not available, so we cannot quantify the evidence for the absence of the negative relationship using a Bayesian reanalysis.

[2] Please note that these examples are mentioned to illustrate the problem, not to make fun of or condemn the associated researchers. I believe the retracted studies involved honest mistakes and the unreplicable work reflects research practices that were simply standard at the time.



**(a)** The Thinker

**(b)** Diskobolos

**Figure 1.1:** Stimuli used to manipulated analytic thinking in the study by Gervais and Norenzayan (2012). Figure (a) is available at https://www.nga.gov/collection/art-object-page.1005.html (public domain) and figure (b) at https://commons.wikimedia.org/w/index.php?curid=547351 (CC BY 2.5).

been a false positive or perhaps even unconsciously exploited the analytic flexibility of EEG research. I understand that the conference attendees did not want to publicly accuse a poor PhD-student of questionable research practices. Yet it felt like I tried to convince them not to put too much faith in the first positive finding, while they tried to convince me that I should disregard the second null-finding.

Add to these anecdotes the proven and alleged fraud cases (e.g., Diederik Stapel, Jens Förster) and disturbingly low replication rates in psychology (Camerer et al., 2018; R. A. Klein, Ratliff, Vianello, Adams, et al., 2014; R. A. Klein et al., 2018; Open Science Collaboration, 2015), and any remaining faith in science and scientists might just crumble away. And –spoiler alert– you will encounter two more failed replication attempts in the upcoming chapters. Surely, these are disheartening facts that do not reflect well on the research culture in psychological science. But luckily, a lot has changed over the last years.

## 1.2 Light at the End of the Tunnel

Once the realization had dawned that the standard research and publication practices were highly vulnerable to biases and misleading results, various calls for enhancing reproducibility and transparency were put forward (Miguel et al., 2014; Munafò et al., 2017; Nosek et al., 2015). Researchers recognized that transparent documentation of the research process, materials, raw data, and analysis scripts is crucial, as it can both protect the researcher from biases and allow the scientific community to validate and build on each other's work. Norms for sharing data and analysis code, for instance, are arguably essential for the self-correcting nature of science; only after thorough inspection and reanalysis of the original data, the mistakes in the articles by Decety et al. (2015) and Clark et al. (2020) were discovered.

To promote transparency and combat the crisis of confidence, various highly effective concrete reforms have been adopted by the research community. *Preregistration*, for instance, has quickly gained popularity and is now widely used (Nosek & Lindsay, 2018). When preregistering a study, the researcher outlines the hypotheses and analysis plan before the data are collected (Nosek et al., 2018; van 't Veer & Giner–Sorolla, 2016). Because the analysis pipeline cannot be tailored to the data, researchers can protect themselves against confirmation bias, hindsight bias, and other questionable research practices that may unwittingly contaminate the results. An extension of preregistration is a *Registered Report*, in which the entire introduction, methods section, and proposed analysis is submitted to a journal (Chambers, 2013). After peer review, the proposed study can get an "in principle acceptance", which gives the green light to collect the data and execute the planned analysis and ensures that the study will be accepted for publication, regardless of the outcomes. A recent empirical investigation of Registered Reports found substantial benefits in methodological rigor, analysis, and overall paper quality compared to traditional articles, while preserving novelty and creativity (Soderberg et al., 2021). In Chapter 4 we describe such a Registered Report study. Moreover, all empirical studies reported in this dissertation were preregistered.

In addition to preregistration, alternative methods have been proposed to inoculate researchers against (unconsciously) biasing their own results. *Analysis blinding*, for instance, involves a temporary distortion of the data when creating an analysis pipeline in order to remove any crucial effects that might bias analytic decisions (Dutilh, Sarafoglou, et al., 2019; MacCoun & Perlmutter, 2015). Blinding data can be achieved, for example, by shuffling the key outcome measure in the real data, hence breaking any potential relation with the independent variable of interest. This method allows the analyst to flexibly develop an analysis pipeline that accounts for (unanticipated) peculiarities in the data without the possibility of being influenced by the hypothesized effect. After the analyst is satisfied with the preprocessing and analysis script, the blind is lifted and the designed pipeline is applied to the real data. In Chapter 11 we empirically compare preregistration and analysis blinding in terms of researchers' experiences (i.e., perceived effort, frustration, and workload) as well as efficiency (i.e., deviations from the planned analysis).

A final crucial reform in psychological science worth highlighting is the trend towards 'team science' (Chartier et al., 2018; Uhlmann et al., 2019). Most prominent are the various collaborative data collection projects such as ManyLabs (Ebersole et al., 2016; R. A. Klein et al., 2019; R. A. Klein, Ratliff, Vianello, Adams, et al., 2014; R. A. Klein et al., 2018), ManyBabies (Frank et al., 2017; The ManyBabies Consortium, 2020), and the Psychological Science Accelerator (S.-C. Chen et al., 2018; Jones et al., 2021; Moshontz et al., 2018). Crowd-sourcing data collection allows researchers to obtain larger samples and hence increase statistical power as well as to reach traditionally less-studied populations (i.e., non-Western participants; Henrich et al., 2010). Chapters 7, 8, and 9 of this dissertation report studies based on data from over ten thousand participants from 24 countries, covering all six populated continents.

In addition to crowd-sourcing data collection, another new initiative is to crowd-source data analysis. In such a 'many-analysts' approach, a given dataset is distributed across multiple analysis teams who are instructed to conduct their own analysis in order to answer a specific research question. As different analysts will naturally choose different analytic strategies, the robustness and variability of the outcomes across a multitude of realistic decisions can be quantified (Silberzahn &

Uhlmann, 2015; Wagenmakers et al., 2022). In Chapters 9 and 10 we describe the results of a many-countries many-analysts project, in which we recruited 120 analysis teams to investigate the much-debated relation between religiosity and well-being.

## 1.3 To Bayes or Not to Bayes?

Actually, to Bayes or not to Bayes is hardly a question. For the analyses for all empirical studies we used the statistical methods of Reverend Thomas Bayes (how else could we, in a project on religion?). Bayesian inference allows us to quantify the extent to which observed data support two competing models (or hypotheses, or parameter values, or accounts of the world). In the context of hypothesis testing and model comparison, we use *Bayes factors* as the main metric of statistical evidence (Berger, 2006; Jeffreys, 1935; Kass & Raftery, 1995). The Bayes factor reflects the change from prior model probabilities to posterior model probabilities and as such quantifies the evidence that the data provide for $\mathcal{M}_1$ versus $\mathcal{M}_2$:

$$\underbrace{\frac{p(\mathcal{M}_1 \mid \text{data})}{p(\mathcal{M}_2 \mid \text{data})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{\text{prior odds}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{M}_1)}{p(\text{data} \mid \mathcal{M}_2)}}_{\text{Bayes factor}} \tag{1.1}$$

For instance, a Bayes factor of four indicates that the data at hand are four times more likely under $\mathcal{M}_1$ than under $\mathcal{M}_2$. If we assume that $\mathcal{M}_1$ and $\mathcal{M}_2$ are equally likely a priori, this also means that $\mathcal{M}_1$ is now four times more plausible than $\mathcal{M}_2$ (i.e., posterior odds).

The motivation for applying a Bayesian rather than frequentist inference is elaborated in Chapter 4. Briefly, Bayes factors allow one to quantify the evidence in favor of or against an effect on a continuous scale and to distinguish between 'absence of evidence' and 'evidence of absence' (Dienes, 2014; Etz & Vandekerckhove, 2018; Wagenmakers, Marsman, et al., 2018). The possibility to obtain evidence in favor of the null-hypothesis is perhaps especially important for replication research (Wagenmakers, Marsman, et al., 2018), where failures to repeat a previous effect are common. In the frequentist framework, we can either reject or fail to reject the null-hypothesis, but we cannot determine the extent to which the data support either hypothesis. When $p > .05$, we conclude that we fail to reject the null-hypothesis assuming no effect of, say, religious affiliation on generosity. However, we cannot quantify how strongly the data favor the null-hypothesis over the alternative hypothesis assuming that religious children are less generous. Alternatively, when $p < .05$, we cannot accept the alternative hypothesis, since predictions by the alternative hypothesis are irrelevant for the statistical test. This is especially pertinent if we have multiple potential hypotheses or models that might account for the observed data. For instance, we might wonder if religious children are (1) equally generous as secular children across all countries, (2) less generous than secular children, to an equal degree across different countries, (3) less generous than secular children, to a varying degree across countries, or (4) less generous than secular children in some countries, and more generous than secular children in other countries. Hypothesis (1) reflects the *null model*: in all countries, the effect is truly zero. Hypothesis (2) reflects the *common effect model* and predicts an overall positive (or negative) effect, without between-country variability in the

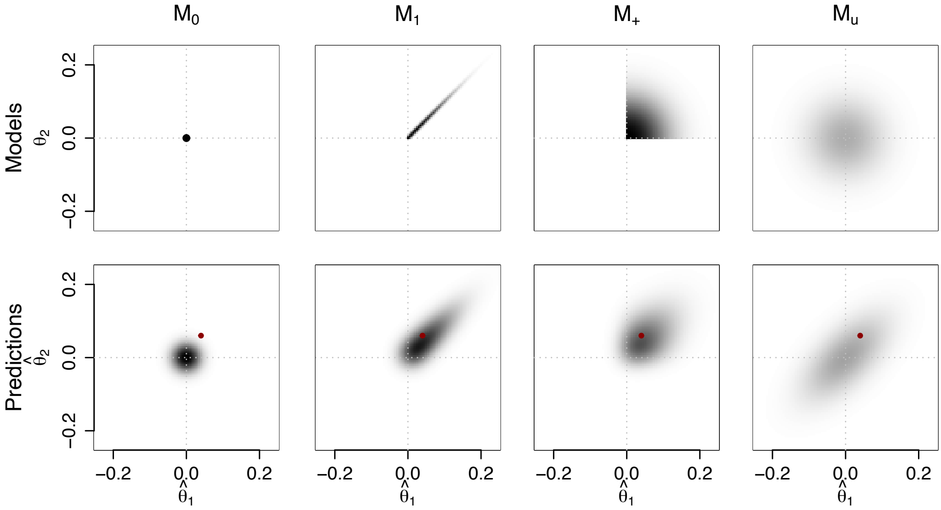


**Figure 1.2:** Illustration of the Bayes factor model comparison approach. The top row shows the model specifications for two example countries, conditional on the prior settings for the average effect of interest, the directional constraint, and the between-country variability. The bottom row shows the corresponding predictions for data, taking into account the sampling noise. The red dots show a hypothetical data point for two countries that is best predicted by the common-effect model (second column). Figure adapted from Haaf and Rouder (2019).

size of the effect. Hypothesis (3) reflects the *positive effects model*, assuming that the effect is truly positive (or negative) in all countries yet varying in size. Finally, hypothesis (4) reflects the *unconstrained model*, allowing the effect to vary in direction and size between countries. Figure 1.2 visualizes the models and predictions for these four different hypotheses. The $\theta_1$ and $\theta_2$ reflect the parameters of interest (e.g., the association between religious affiliation and generosity) for two exemplary countries. In Chapters 7 and 8 we use this Bayes factor model comparison approach (Haaf & Rouder, 2017; Rouder, Haaf, Davis-Stober, et al., 2019) to quantify the evidence for the effects of interest in the cross-cultural data.

## 1.4 In Science We Trust

At the start of my thesis project, I remember actively contemplating whether or not to embark on this PhD project, as I feared the project might be somewhat pessimistic and discouraging. Against the backdrop of the crisis in psychology, "assessing replicability in the psychology of religion" did not necessarily evoke images of sexy ground-breaking findings that would make a nice story at a dinner party. After all, from the start, none of the project members were so naive as to expect perfect replication rates in the psychology of religion, and I was very aware that null findings would constitute a large part of my results. And to be honest, after publishing the two null-results presented in Chapters 4 and 5 I was also pretty excited to have finally found effects that could be supported by the data (Chapters 7, 8, 9). In the end, I think it's only natural to prefer discovery over non-discovery and to wish that one's hypotheses are supported rather than disconfirmed by the data. This intuitive preference for positive results does not necessarily have to be a problem, as long as we put the right mechanisms of checks and balances in place to protect especially ourselves from taking shortcuts and unconsciously exploiting analytic flexibility. As Richard Feynman (1974) said: "The first principle is that you must not fool yourself and you are the easiest person to fool." It seems like the field has finally acknowledged the truth of this statement –although it took some time and a series of shocking events to get there. Luckily, the 'revolution' (Spellman, 2015; Vazire, 2018) or 'renaissance' (L. D. Nelson et al., 2018) in psychological science is happening at full speed and transparency, robustness, and collaborative science are no longer scientific utopia's (Nosek & Bar–Anan, 2012; Nosek et al., 2012; Uhlmann et al., 2019).

So it's probably time to slowly restore our faith in psychological science again. After all, the scientific practice does not function without an appeal on trust and faith; despite the Royal Society's motto, as scientists we must also trust other scientists, as we can hardly reanalyse and replicate the results in every single paper we cite. And although the replication crisis has dealt a blow to the public's faith in psychological science (Anvari & Lakens, 2018; Wingen et al., 2020), trust in scientists is still globally among the highest of all authorities, only rivalled in some countries by the military (Funk, Tyson, et al., 2020).[3]

## 1.5 Chapter Overview

### 1.5.1 Part I: Open science and replication: Why and how?

The first part of this dissertation sketches the context of the replication crisis in the social sciences. In Chapter 2, we introduce the core concepts of open science and offer concrete suggestions to adopt open science practices within the (cognitive) science of religion. The suggestions are illustrated by a 'glimpse behind the scenes' of the cross-cultural religious replication project (CCRRP) that is described in Part III. In Chapter 3, we explore the role of the intuitive plausibility of research outcomes in the context of the replication crisis. By asking laypeople to predict replication outcomes we aimed to address the question: could we have known if we had simply heeded common sense? Research in the social sciences have often put extreme, 'sexy' effects too much into the spotlight. Yet precisely these effects have turned out not to replicate and were in fact also not considered plausible to many scientists and –as shown in Chapter 3– non-scientists alike. We argue that we should not ignore the information we can derive from common sense.

### 1.5.2 Part II: Replicating key effects in the psychology of religion (or not)

In Part II and III, we put our money where our mouth is and describe replication studies targeting some influential effects in the psychology and cognitive science of religion. Chapter 4 reports a direct replication of *compensatory control theory* (CCT), which postulates that religion can serve as an external source of control that can

[3] Let's just ignore that arguably worrisome fact for now.

substitute a perceived lack of personal control (Kay et al., 2008). We found that neither in the Netherlands, nor in the US did an experimental manipulation threatening personal control increase belief in a controlling God. However, while experimental manipulations of control appeared ineffective in shifting belief in God, individual differences in the experience of control may be related to religious beliefs in a way that is consistent with CCT, at least in the US. In Chapter 5 we describe an fMRI study on the relation between religiosity and behavioral and neural conflict processing. This work involves a conceptual replication of the study by Inzlicht et al. (2009). Contrary to the original study, however, we found no evidence that individual differences in religiosity were related to performance on the Stroop task as measured in accuracy and interference effects, nor to neural markers of response conflict (correct responses vs. errors) or informational conflict (congruent vs. incongruent stimuli). In Chapter 6 we report a Bayesian reanalysis of the Many Labs 4 replication study (R. A. Klein et al., 2019) on the mortality salience effect from Terror Management Theory (Greenberg et al., 1995; Greenberg et al., 1994). We conducted a multiverse analysis across theoretically or statistically-motivated data inclusion criteria and prior settings. The results largely converged to the conclusion that the data provide evidence against the mortality salience effect: reminders of one's own death do not seem to strengthen one's cultural identity.

### 1.5.3 Part III: The Cross-Cultural Religious Replication Project

In Part III we describe the outcomes of the CCRRP introduced in Chapter 2. In Part II we conducted a direct and a conceptual replication of two specific influential studies and reanalysed another direct replication. The CCRRP, on the other hand, targeted general effects rather than particular studies.[4] The research reported in this part of the dissertation results from a cross-cultural data collection effort involving 10,195 participants from 24 countries. Chapter 7 describes an experimental study on source credibility effects at play in the context of science and spirituality. We found evidence for what we call the 'Einstein effect': people tend to confer more credibility to incomprehensible claims when attributed to a scientist than when the very same claims are attributed to a spiritual guru. This Einstein effect differed for religious versus non-religious participants: individuals scoring low on religiosity considered the statement from the guru less credible than the statement from the scientist, while this difference was less pronounced for highly religious individuals.

In Chapter 8 we report the results of the second sub-project of the CCRRP. Here we investigate mind-body dualism and the relation with religiosity. Following previous work, we used a vignette describing the passing of the person and subsequently inquired the continuation or cessation of bodily states (e.g., hunger) and mental states (e.g., love). We replicated previous work showing that people tend to reason dualistically as they consider mental states more likely to continue after death than bodily states. While individual religiosity was associated with both overall continuity judgments and mind-body dualism (i.e., the difference between mental and bodily states), a context manipulation emphasising religion did enhance overall continuity but not

[4] Note that the CCRRP included a package of four independent studies. Three of those are reported as separate chapters in this dissertation. The final study has not been written up yet, but is shortly discussed in Appendix A.

mind-body dualism. Contrary to intuitive dualism accounts, however, the pattern of results suggests that cessation rather than continuation is the default response, even for high-level mental processes.

Chapter 9 introduces the many-analysts religion project (MARP), in which we recruited 120 analysis teams to investigate the robustness of the relation between religiosity and well-being in the CCRRP data. Results on the positive association between religiosity and self-reported well-being were remarkably consistent: all but 3 teams reported positive effect sizes with credible/confidence intervals excluding zero. Somewhat more variability was observed for the question 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), though a $^2/_3$ majority of analysis teams again reported positive effect sizes with confidence/credible intervals excluding zero.

In Chapter 10, we reflect on the outcomes of the MARP and put both the answers to the research questions as well as our experiences with a many-analysts approach in a broader perspective. We address the issue of theoretical specificity, highlight some in-depth observations beyond the primary research questions, consider methodological concerns, and discuss our experience of organizing a many-analysts project. Note that the two chapters on the the many-analysts religion project are published as a target article in a special issue of *Religion, Brain & Behavior* that also includes commentary articles by some analysis team members. Therefore, the article in Chapter 10 includes references and responses to these commentaries. It can, however, also be read as a standalone article.

Chapter 11 describes the results of an experimental manipulation applied to the MARP. We assigned all analysis teams participating in the MARP to either a preregistration or an analysis blinding condition. After the teams proposed an analysis based on their assigned preparation method, we compared the teams' experiences and efficiency. We found that subjective experiences and workload are comparable between methods, but that blinding may lead to fewer deviations from the planned analysis.

In Chapter 12, I integrate the findings from all empirical chapters, discuss their relevance for the psychology of religion, and reflect on the state of replicability in this subfield. I return to the replication script introduced in Chapter 2 and highlight some personal experience and insights. Finally, we note that one study from the CCRRP is not yet published. In the interest of transparency, the results of this study are reported in Appendix A.

# Part I

# Open Science and Replication: Why and How?

# 1

# 2 Advancing the Cognitive Science of Religion Through Replication and Open Science

The Cognitive Science of Religion (CSR) is a relatively young but prolific field that has offered compelling insights into religious minds and practices. However, many empirical findings within this field are still preliminary and the reliability of these findings remains to be determined. In this chapter, we first argue that it is crucial to critically evaluate the CSR literature and adopt open science practices and replication research in particular in moving the field forward. Second, we highlight the outcomes of previous replications and make suggestions for future replications in the CSR, with a particular focus on neuroscience, developmental psychology, and qualitative research. Finally, we provide a 'replication script' with advice on how to select, conduct, and organize replication research. Our approach is illustrated with a 'glimpse behind the scenes' of the recently launched Cross-Cultural Religious Replication Project, in the hope of inspiring scholars of religion to embrace open science and replication in their own research.

## 2.1 Introduction

Science is associated with discovery, creativity, and innovation. Thinking outside the box is typically considered a hallmark of the scientific genius. For many researchers pursuing an academic career, it thus seems highly attractive to conduct new, creative studies, to invent new theories, and to postulate novel and crazy hypotheses; this will establish name and fame and may gain publications in high-impact journals. Rigorous verification of existing theories and findings seems far less appealing. Although most scientists would agree that replication is essential for scientific progress, not all of them are eager to commit themselves to this ideal (M. S. Anderson et al., 2007; Baker, 2016). Replication research has the image of being 'boring', 'tedious', non-creative, and disadvantageous for one's career prospects (see for instance Yong, 2012).

This chapter has been adapted from: Hoogeveen, S., & van Elk, M. (2021). Advancing the cognitive science of religion through replication and open science. *Journal for the Cognitive Science of Religion*, *6*(1-2), 158–190. https://doi.org/10.1558/jcsr.39039

Here, we propose a different perspective. First, we believe it is important to be critical of many 'classical' findings in the Cognitive Science of Religion (CSR) and the psychology of religion. Especially non-psychologists in the field of religion may not be sufficiently familiar with the recent crisis and developments in psychology. Accordingly, they may often take published empirical findings at face value. Below we will highlight some examples of high-impact studies in the CSR, most notably from the field of neuroscience and developmental psychology, that still await independent replication.

We offer concrete suggestions of how the CSR should go about replicating these and other studies, including qualitative and field studies as well. Second, we argue that as a field we should acknowledge the merits of open science in general, and replication studies in particular, in moving the CSR forward. Substantiating and validating existing theories and findings is more urgent than developing new 'grand theories' of religion. We should continue to encourage replication attempts in the CSR, and extend replication to different disciplines and methods. Third, drawing on our own experience, we will exemplify that replication studies are in fact highly innovative, require a lot of creative thinking, and foster collaboration between (international) research groups.

## 2.2 Call for Caution

The year 2011 marked a tipping point in psychological science. The publication of a paper demonstrating the possibility of precognition through reverse priming (Bem, 2011), as well as the discovery of severe fraud in the work of social psychologist Diederik Stapel, instigated a process of critical internal scrutiny in psychology.

While the Stapel case had a severe negative impact on the public credibility of psychological science, the Bem study was arguably more worrying as it demonstrated that even with good intentions and adherence to standard practices, psychological research could get derailed. The presented empirical evidence for extrasensory perception was, although statistically significant, in fact not compelling: Bem only provided weak evidence for the extraordinary claim that people can look into the future and did not specify an underlying causal mechanism (Wagenmakers et al., 2011). The study also raised questions about post-hoc hypothesizing and the blurred boundaries between confirmatory research and exploratory research ("fishing"); after seeing the data it may have seemed plausible that precognition only occurred for erotic pictures and extroverted people, but it is hard to believe that this was an a priori hypothesis. Yet the real problem was arguably the lack of transparency in the research process; there was simply no way of knowing whether or not the results were fully anticipated or selectively reported. In retrospect, the study on precognition indeed shook psychological science on its foundations, yet not because the field universally embraced the idea of precognition but because it made people realize that the then-standard research practices lacked rigor and failed to sufficiently inoculate researchers against human biases (e.g., confirmation bias and hindsight bias). Eventually, these issues created a distorted literature with spurious findings.

Attempts to map out the status of the field from a meta-perspective only further lay bare the deep-grained flaws in the methodology and the incentive structure that had become the norm in the field. Voiced issues included publication bias (significant results are more likely to be published than null-results; G. Francis, 2013), underpowered studies (Button et al., 2013; Ioannidis, 2005), and the ubiquity of questionable research practices such as selective reporting of (dependent) variables, of explored analysis paths, and of entire studies, creative inclusion/exclusion criteria, post-hoc hypothesizing (John et al., 2012). Finally, warnings emerged concerning the lack of replication studies to determine the reliability, robustness, and stability of obtained findings (Makel et al., 2012; Schmidt, 2009). Later on, when these replication studies were conducted, they further demonstrated disturbingly low reproducibility rates (Camerer et al., 2018; R. A. Klein, Ratliff, Vianello, Adams, et al., 2014; R. A. Klein et al., 2018; Open Science Collaboration, 2012).

Fortunately, there is light at the end of the tunnel. This rather gloomy picture has inspired a wealth of colorful and creative initiatives to combat the 'crisis of confidence'. Accordingly, some open science advocates have proposed the terms "Credibility Revolution" (Vazire, 2018) or "Revolution 2.0" (Spellman, 2015), to emphasize the constructive reforms rather than focusing selectively on its causes. One popular approach aimed at self-correction in science is *preregistration* (Nosek & Lindsay, 2018; Nosek et al., 2018; van 't Veer & Giner–Sorolla, 2016). This practice entails the detailed delineation of the materials, methods, and analysis plan prior to data collection (see also Kavanagh and Kapitany, 2017). Because the analysis pipeline cannot be tailored to the data, researchers protect themselves against hindsight and confirmation bias and other practices that may unwittingly bias the results (Wagenmakers et al., 2012). Additional open science initiatives focus on open databases and repositories to foster data sharing (e.g., the Open Science Framework; OSF), journal guidelines to promote transparency and reproducibility (PRO initiative; Nosek et al., 2015), novel publishing incentives that focus on quality of design rather than outcomes (e.g., the Registered Reports; Chambers, 2013), badges to award openness and transparency (Kidwell et al., 2016)[1], and large-scale replication projects (Camerer et al., 2018; R. A. Klein, Ratliff, Vianello, Adams, et al., 2014; R. A. Klein et al., 2018; Open Science Collaboration, 2012). These efforts are not restricted to the field of psychology, but also resonate in for instance behavioral economics (Camerer et al., 2016) and empirical philosophy (Cova et al., 2018).

Note that since there are numerous sources introducing general concepts of open science and its application in specific sub-fields, we will here restrict our focus on replication research. We refer the interested reader to Crüwell et al. (2019) for an accessible introduction and annotated reading list on open science practices in general, and to Charles et al. (2019) for an assessment of and recommendations for open science within the psychology of religion in particular.

## 2.3 Replication and the CSR

The CSR has also started to embrace open science initiatives. Various labs now preregister their studies (e.g., Gervais et al., 2017; McPhetres, 2018; C. J. M. White et al., 2018), and data sharing has become far more common (e.g., Maij et al., 2017; Purzycki et al., 2018). At the same time, a recent systematic analysis of the 2017 issues from three psychology of religion journals found that none of the 53 included articles

[1] Though the effectiveness of offering open science badges may not be unequivocal, at least for medical journals (Rowhani-Farid et al., 2020).

were preregistered (Charles et al., 2019). Furthermore, with regard to replication research, most studies conducted so far were aimed at replicating social psychological or priming studies (e.g., Gomes & McCullough, 2015; Sanchez et al., 2017).

Although the debate about the reliability and robustness of some religious priming effects, such as those related to prosociality (Shariff, Willard, Andersen, et al., 2016; van Elk et al., 2016), and cheating behaviour (Lang et al., 2016; Nichols et al., 2020) is not yet settled, it seems that caution is warranted regarding results of religious priming studies. Various replications of priming studies failed to find compelling evidence: subtle reminders of religion do not increase risk taking (based on the divine protection hypothesis; Gervais et al., 2020; Kupor et al., 2015), do not decrease grip endurance (based on the sexual and reproductive religiosity model; Hone and McCullough, 2015; McCullough et al., 2012), religious priming does not increase dictator game allocations (based on the religion and prosociality link; Gomes and McCullough, 2015; Shariff and Norenzayan, 2007), and religious priming does not increase work ethic (based on the implicit puritanism account; Tierney et al., 2021; Uhlmann et al., 2011). In addition, analytical thinking primes do not decrease religiosity (Gervais & Norenzayan, 2012; Sanchez et al., 2017) and threats to personal control do not increase belief in a controlling God (Hoogeveen, Wagenmakers, et al., 2018; Kay et al., 2008). Keep in mind though, that these all concern social priming effects, the efficacy of which has been contested in general (Cesario, 2014; Doyen et al., 2012; Gilder & Heerey, 2018; Pashler et al., 2013; Shanks et al., 2013). Thus, while these failed replications may appear disheartening for the scientific study of religion, we believe there are ample non-priming studies that may have more favourable prior odds of replicability.

We argue that the open science perspective and instruments, including replication studies should also be stimulated in other sub-fields, such as developmental research (e.g., intuitive dualism; Bloom, 2005, teleological thinking; Kelemen, 2004), the neuroscience of religion (e.g., neural correlates of religiosity; S. Harris et al., 2009, palliative attributes of religious beliefs; Inzlicht et al., 2011, involvement of social brain areas in prayer; Schjoedt et al., 2009), and qualitative research branches, including cultural anthropology, history, literary studies and religious studies. Indeed, these fields all face additional difficulties and challenges with regard to replication research.

One of the great merits of the CSR is the interdisciplinary nature of the research, which has successfully been advocated by various scholars of religion (Bulbulia, 2013; Bulbulia & Slingerland, 2012; Slingerland & Collard, 2011; Taves, 2010). In order to continue this progressive movement, we should put the open science principles into action across all sub-fields, by using the tools that were introduced to enhance replicability in the life and social sciences. At the same time, we should of course remain sensitive to the peculiarities of all sub-disciplines in the scientific study of religion. In the following sections we will outline the challenges associated with open science practices and replication research in neuroscience, developmental research, and qualitative research. We will present potential solutions and argue for the importance of taking these into account in moving the field of the CSR forward.

First, however, it is important to clarify the definition and scope of the term 'replication'. Often, a distinction is made between 'direct replications', in which the exact same study protocol is repeated in a new sample, and 'conceptual replications', in which a different methodology is applied to test the same research hypothesis (e.g., Nosek & Errington, 2017). These conceptual distinctions place much emphasis on methodological aspects and are arguably less relevant for non-experimental research. An alternative definition that emphasizes theoretical implications was recently proposed by Nosek and Errington (2020, p.2): "Replication is a study for which any outcome would be considered diagnostic evidence about a claim from prior research. [...] To be a replication, 2 things must be true: outcomes consistent with a prior claim would increase confidence in the claim, and outcomes inconsistent with a prior claim would decrease confidence in the claim." Following this definition, replication research is part of the iterative process of theorizing, predicting, testing, and redefining theories in science. Furthermore, in many cases, it may be worthwhile to adopt a 'replication+' approach, in which the original study is repeated as closely as possible and extended in order to answer additional questions, by including new variables, conditions, samples, or studies (Bonett, 2012). This approach may present the best of both worlds by (dis)confirming existing findings and adding new insights or nuances.

## 2.4 Replication Across Methods and Disciplines

### 2.4.1 Neuroscientific Research

Neuroscience in general has been plagued with problems of low statistical power due its reliance on expensive equipment and the time-intensive nature of data collection and analysis (Button et al., 2013; Szucs & Ioannidis, 2017). This extensive process of data collection may limit the enthusiasm for conducting replication studies. A Web of Science search indicated that of all neuroscience articles published since 2010, only 0.08% contained the word 'replicate(d)' or 'replication' in the title, whereas this was 0.21% for all psychology articles. In addition, few neuroscience journals (6%) explicitly state their interest in replication studies, lowering the incentives for conducting these studies even more (A. W. K. Yeung, 2017). For instance, Huber et al. (2019) describe their experience of attempting and eventually failing to publish a large replication study of neuroimaging work on memory retrieval that appeared in Nature Neuroscience (Potter et al., 2018; Wimber et al., 2015). In addition, the relatively few direct replication attempts that have been undertaken show exceptionally low replication rates. For instance, Boekel et al. (2015) attempted to replicate 17 brain-behaviour effects documented in the literature and found convincing support for only one out of 17.

At the same time, it seems the field has become more aware of the importance of replication research, as evidenced for instance by special issues on replication (e.g., Barch and Yarkoni, 2013; or see this call from the Journal of Memory and Language). Efforts to collect and publicly share large neuroimaging data also contribute to increasing transparency in general and extending possibilities for conducting replication studies in particular. Botvinik-Nezer et al. (2019), for instance, collected a large dataset on decision making, initially aimed at quantifying analytical variability in neuroimaging research (Botvinik-Nezer et al., 2020), but the authors explicitly encourage researchers to use the data to assess replicability of specific findings within this paradigm.

Table 2.1: Relevant Sources and Projects Related to Open Science and Replication for the Cognitive Science of Religion.

| | |
|---|---|
| Introduction to Open Science | |
| Crüwell et al. (2019) | Annotated reading list covering open data, preregistration and registered reports, replication research and more. |
| Charles et al. (2019) | Assessment of OS practices in the CSR and recommendations for scholars of religion. |
| Collaborative (Replication) Projects | |
| Frank et al. (2017) | ManyBabies, a global network for replications in developmental research (osf.io/rpw6d). |
| Moshontz et al. (2018) | Psychological Science Accelerator, a global laboratory network for crowdsourcing research, including replications (psysci-acc.org/). |
| Wagge et al. (2019) | Collaborative Research and Education Project (CREP), global network for involving students in replication research (osf.io/wfc6u/). |
| Open Science in Qualitative Research | |
| Tamminen and Poucher (2018) | Practical advice on data sharing and re-use or replication in qualitative research. |
| Bishop (2009) | Discussion and advice on ethical data sharing en re-use in qualitative research. |
| Introduction to Bayesian Inference | |
| Etz et al. (2018) | Introduction of theoretical and practical concepts for researchers interested in Bayesian statistics. |
| Wagenmakers, Love, et al. (2018) | Practical guide on conducting Bayesian analyses for various standard tests in JASP (JASP Team, 2019). |

*Note.* This is a non-exhaustive list of available interesting references and projects, that is mostly intended to inspire.

#### 2.4.1.1 Prospects for the CSR

Neuroscientific studies have contributed substantially to the CSR, by fostering our understanding of religious beliefs and experiences. As two illustrative examples, studies by Schjoedt et al. (2009) and Inzlicht et al. (2009) gave strong impetuses to the field by providing insights on prayer experiences and conflict detection in religious believers, respectively. Schjoedt and colleagues (2009) used fMRI to demonstrate that brain areas involved in everyday social interaction and mentalizing are also activated when highly religious believers pray to God, substantiating articulated accounts of believers' personal relationship with God. Inzlicht and colleagues (2009) showed that religious believers exhibited a reduced neural response to errors on a Stroop task, potentially reflecting the palliative effects of religiosity on distress responses. Although different interpretations of the results have been put forward (Schjoedt & Bulbulia, 2011), these findings added an interesting theoretical layer to the cognitive science of religion.

Both of these studies have in fact been subjected to replication attempts. First, the study by Schjoedt et al. (2009), which used Danish Christians, was successfully replicated in an American Pentecostal sample (Neubauer, 2014). While replicating the finding that personal prayer involved brain regions related to social cognition, the replication also extended the original study, by showing substantial overlap in neural activation between personal prayer and talking to a loved one. This was taken to indicate that communication with God through prayer and interaction with a real person rely on similar neurocognitive processes related to mentalizing. Second, with respect to the Inzlicht et al. (2009) study, we recently failed to conceptually replicate the main results (Hoogeveen, Snoek, et al., 2020; see Chapter 5). Using data from 193 subjects, we found no association between religiosity and conflict sensitivity, neither at a behavioral nor at a neural level, casting doubt on the reliability of these previous findings. Similarly, van Elk and Snoek (2020) found no evidence for the relation between religiosity and grey matter volume in several brain areas that were identified in the literature as being associated with religiosity or mystical experiences. Both these datasets, as well as the overall neuroimaging project data they were part of, are publicly available and we happily invite researchers to explore whether these data may contribute to answering additional questions in the CSR.[2]

Notably, some of the seminal neuroscientific studies on religiosity are illustrative of bygone times: with samples of $n = 20$ (within-subjects; Schjoedt et al., 2009), $n = 28$; $n = 22$ (between-subjects; Inzlicht et al., 2009), $n = 30$ (between-subjects; S. Harris et al., 2009), and $n = 36$ (between-subjects; Schjoedt et al., 2011) these studies are most certainly underpowered, and thus potentially unreliable.

Importantly, while samples of 20-30 participants may suffice in within-subjects designs with many trials per person, they are most likely inadequate for detecting reliable between-subjects effects or individual differences which are typically targeted in

[2] For the overall neuroimaging project data, see Snoek et al. (2020). For the religiosity and conflict sensitivity study: The preprocessing scripts for the fMRI analysis and the exploratory fMRI analyses can be found at https://github.com/lukassnoek/ReligiosityFMRI. Unthresholded brain maps can be found at https://neurovault.org/collections/6139/. For the structural brain differences and religiosity study: All analysis code can be found at https://github.com/lukassnoek/ReligiosityVBM. Unthresholded brain maps from the whole-brain analysis can be found at https://neurovault.org/collections/5380

the cognitive science of religion, e.g., comparing religious individuals to non-religious individuals or atheists. Indeed, researchers often overestimate the power achieved in these between-subject designs, especially for small effects (Bakker et al., 2016). In general, it may be difficult to formulate standard guidelines for adequate sample sizes as power depends on the specific design and expected effect size. Nevertheless, some have recommended a minimum sample size of $N = 100$ for correlational (neuroimaging) research (Dubois & Adolphs, 2016; Schönbrodt & Perugini, 2013).

Therefore, we would recommend conducting high-powered replication attempts of some of the key neuroscientific studies on religiosity, such as those mentioned above, as these findings had a strong impact on theory development in the field. Replicators could for example, repeat the work by S. Harris et al. (2009) investigating neural correlates of religious and non-religious beliefs. An interesting extension would be to apply multi-voxel pattern analysis (MVPA) to provide insight into how religious concepts are distributed and represented among different brain regions, and compare patterns of brain activation coding for religious agents (e.g., God, angel), for imaginary agents (e.g., Santa Claus, Superman) and for real people (e.g., Napoleon, Bill Gates; cf. Leshinskaya et al., 2017). This would allow one to determine to what extent there is an overlap between the neural representation of real and supernatural agents. In addition, it would be important to establish to what extent these findings generalize across different cultural settings and different religions (e.g., do Muslims show the same pattern of social cognition-related brain activity during prayer as orthodox Christians?).

### 2.4.2 Developmental Research

Similar to neuroscience, the field of developmental psychology has seen many underpowered studies, mostly due to difficulties in recruiting a sufficient number of children or babies (Frank et al., 2017; Schott et al., 2018). This is further complicated by the fact that in most developmental studies there is a high dropout rate, due to fuzziness or distraction of the child. In addition, researchers typically rely on indirect measures of cognition, such as heart rate, EEG responses or gaze pattern (Cristia et al., 2016). Although useful and providing interesting insight into the early stages of cognitive development, there is the recurring problem of providing 'rich interpretations' of the data that are not fully warranted in light of the empirical evidence (Haith, 1998).

So how can we remedy these problems? Developmental research always faces a challenge in terms of subject recruitment and testing, but many hands can make light work. Large-scale collaborative (replication) efforts can play a pivotal role in advancing this field. A notable initiative is for instance the ManyBabies project, which specifically aims to set up multi-lab replication studies to investigate the developmental trajectory of key phenomena such as theory of mind reasoning and infant-directed speech (Frank et al., 2017; see project details via osf.io/rpw6d). The first project, for instance, successfully replicated infants' preference for infant-directed speech over adult-directed speech, and identified developmental, cultural, and methodological moderators (The ManyBabies Consortium, 2020). While the findings of these projects may be interesting for the cognitive science of religion on their own (e.g., how does theory of mind develop for agents and non-agents?), ManyBabies may also serve as an example on how to set up large scale collaborative projects for other developmental effects. The fact that these many-labs style projects enable cross-cultural data collection may be especially valuable for the cognitive science of religion.

#### 2.4.2.1 Prospects for the CSR

The way children develop an understanding of other minds, including God's mind, and believe in an afterlife provides an important argument for the naturalness of religion theory (J. L. Barrett, 2000; Bloom, 2007). Although popular, this account has been criticized for ignoring the role of cultural learning and religious upbringing (Banerjee & Bloom, 2013; Corriveau et al., 2015; E. M. Evans, 2001; Gervais, Willard, et al., 2011; P. L. Harris & Giménez, 2005). The ongoing debate in this domain (e.g., J. L. Barrett, 2018; Sterelny, 2018), further stresses the need for critical evaluation and replication of cornerstone studies, including developmental studies on supernatural mind representation (J. L. Barrett et al., 2003; J. L. Barrett et al., 2001), teleological thinking (Kelemen, 2004), and mind-body dualism (Bering, 2006). Notably, Makris and Pnevmatikos (2007) conducted a replication of the study by J. L. Barrett et al. (2001), in which the authors challenged the finding that the understanding of God's mind precedes the development of theory of mind reasoning about humans. As the debate on the naturalness of religion critically hinges on this type of developmental research, replication of these seminal studies is crucial. Replication studies comparing children from different ages across secular and religious cultures could for instance shed light on the central question whether religious cognition (including dualistic and teleological reasoning) indeed 'comes natural' to children.

### 2.4.3 Qualitative Research

Field studies and qualitative research have made important contributions to the scientific study of religion, for instance, by providing anthropological records of religious communities (e.g., J. L. Barrett, 1998; Power, 2017; Schjoedt et al., 2013; Whitehouse & Lanman, 2014), historical analyses of the cultural evolution of religions (e.g., Norenzayan et al., 2016; Wright, 2010), and phenomenological accounts of religious experiences and rituals (e.g., Hardy, 1981; Luhrmann, 2012; Taves, 1999).

Replication and open science are contentious topics in qualitative research and the humanities. Next to the question of when and how to enhance reproducibility, scholars have also raised precursory questions of whether replication is even *possible* and *desirable* in the humanities. Below we argue that replication is important for any research project that involves empirical data, and thus also for many - though not all - qualitative studies.

#### 2.4.3.1 Conceptual considerations

Proponents of replication studies in the humanities have pointed out that much research in the humanities is empirical in nature (Peels & Bouter, 2018a). Conceptually, being able to answer a question by using the same or similar methods under the same or similar circumstances, is desirable in any scientific field (Peels & Bouter, 2018b). As such, replication should be a pillar in the humanities as well as in the other sciences. Others appear less enthusiastic. In a recent statement in Nature with the title "Resist calls for replicability in the humanities", de Rijcke and Penders (2018) argue that the quality criteria for research are fundamentally different between the (natural) sciences and the humanities. According to the authors, the study objects of the humanities are embedded in a dynamically evolving culture, and therefore by definition cannot be studied separate from the original context (as a replication study implies).

Indeed, qualitative research typically provides an in-depth analysis of a specific study object, which in turn constrains the scope and generalizability of the study. This does not however, preclude replication. Payne and Williams (2005), for instance, argue that humanities often implicitly or explicitly employ "moderatum generalizations" in their research. These are moderate generalizations in terms of the scope of what is claimed and the strength of belief in the claim. These generalizations do not imply universality but have relevance beyond their immediate object, and are presented as hypotheses rather than facts. It is the task of the researcher to explicitly specify the conditions and to demarcate the line between data and interpretation (Payne & Williams, 2005). As argued by Anczyk et al. (2019), replication can be used to verify the "moderatum generalizations" under the assumed conditions, or to investigate how changes in context affect the conclusions. This kind of conceptual replication is indeed recommended to increase reproducibility in the humanities (Peels & Bouter, 2018b). As new data cannot always be acquired, the potential for direct replication in qualitative research is limited. Conceptual replication and triangulation - approaching a claim using independent lines of evidence and different methods - may better serve this purpose (Munafò & Smith, 2018; Peels & Bouter, 2018b).

Note that especially with respect to qualitative research, the distinction between replication and generalization may become fuzzy. However, following the broad definition as proposed by Nosek and Errington (2020), generalization to different contexts can also be considered replication, as long as the original claim presupposes an effect across different contexts, including the newly targeted one.

As argued by Tamminen and Poucher (2018), while some may consider the rationale for engaging in open science problematic for qualitative research, implementing practices could be relatively straightforward. In other words, *if* qualitative researchers want to commit to adopting open science in their work, they could relatively easily do so. We refer the interested reader to Tamminen and Poucher (2018) and Bishop (2009), who address various ethical concerns and provide practical advice on data sharing and re-use or replication in qualitative research.

In general, standardization of protocols and tools may boost reliability and reproducibility in qualitative as much as in quantitative research. Petitmengin (2006) for instance, developed a highly structured interview method to access one's subjective experiences in great detail. This approach, called micro-phenomenology, presents a rigorous method to become aware of and describe an active experience with great precision. This technique, which has been applied to the experience of meditation (Petitmengin et al., 2017; Petitmengin et al., 2018), may be particularly relevant to study religious experiences, which are vulnerable to memory biases and narrative construction (Schjoedt et al., 2013; van Mulukom, 2017; Xygalatas, Schjoedt, et al., 2013).

#### 2.4.3.2 Illustrations from the CSR

Examples from the CSR can exemplify the conceptual concerns related to replicating studies from anthropology and the humanities. For instance, conducting a direct replication of the findings from a Spanish fire-walking ritual (Konvalinka et al., 2011) in Finland, seems silly and invalid, as Finnish people most likely do not engage in fire-walking rituals. However, the underlying conclusions of synchronized arousal contributing to social cohesion can very well be replicated in a sample of practitioners of the high-arousal Brazilian Jiu Jitsu material arts (it did not exactly replicate; Kavanagh et al., 2018) or perhaps in the Finnish ritual of 'wife-carrying' (Eukonkanto, see https://en.wikipedia.org/wiki/Wife-carrying; we happily invite researchers to subject this practice to a replication attempt).

Another example may be found in Slingerland and Chudek (2011), who argue for the presence of folk dualism beliefs in early Chinese culture on the basis of the analysis of historical texts from ancient China (pre-221 BCE). Unless considerable new bodies of text from more than 2000 years ago are discovered, a direct replication seems fruitless here as well. One could, however, draw from different sorts of evidence to investigate the same question for the same population, i.e., conduct a conceptual replication by using a process of triangulation. Pan (2017), for example, used archaeological records of traditional medical practices to shed light on early Chinese mind-body dualism – and reached a different conclusion than Slingerland and Chudek (2011).

The idea of making "moderatum generalizations" and subsequently testing these is also nicely illustrated by the work of Luhrmann (2005, 2012). On the basis of field work and interviews with members of the evangelical Vineyard Christian church in Chicago and Palo Alto, she describes and interprets the primary data. The overall aim is to learn something more general about the way in which believers come to experience the supernatural as real. One of her conclusions reads for instance as follows: "Perhaps the most novel suggestion here [...] is that there may be a shared psychological mechanism – absorption – in the psychiatric response to trauma and in spiritual experience, that the individual capacity for absorption can be trained, and that cultural interest in that training can rise and fall at different times." (Luhrmann, 2005, p.154). The 'absorption hypothesis' has also been tested in groups of participants from India and West-Africa – even though it would probably not be qualified as a 'direct replication' (Luhrmann et al., 2015). This work demonstrates the value of replication beyond confirming or rejecting the original theory and study. By showing both similarities in underlying process and differences in particular content, this research has made significant theoretical contributions to the CSR literature.

## 2.5 A Glimpse Behind the Scenes & A Replication Script

Translating these ideas into action, we recently set up direct replication and conceptual replication studies on CSR-related topics. In addition to previously mentioned replications of the link between personal control and belief in a controlling God (Hoogeveen, Wagenmakers, et al., 2018; Kay et al., 2008) and cognitive control sensitivity and religiosity (Hoogeveen, Snoek, et al., 2020; Inzlicht et al., 2009), we recently launched the cross-cultural religious replication project (CCRRP), a large-scale collaborative project (see www.relcoglab.com/religious-replication-project). We believe a glimpse behind the scenes of this project may illustrate and crystallize our idea that replication can be fun, challenging, innovative and it can create great opportunities for collaboration. Below, we provide a 'replication script' with recommendations and tips for conducting replication research in the CSR, building on our experiences with the CCRRP (see also Table 2.2 on page 25).

First off, deciding which study to select for replication out of the vast literature can be both exciting and fun. We argue that the a priori likelihood of replicability should be taken into account. Assessment of the added value of replication studies, for example by looking at informational gain, is an important first step. For our project, conversations with the project committee highlighted that we should opt for studies with medium chances of being successfully replicated, as these are most interesting and informative. This notion of information gain can be formalized in a Bayesian framework (Hardwicke, Tessler, et al., 2018), but it basically follows a simple intuition; especially for original studies with highly surprising effects (i.e., low prior odds) or small sample sizes (i.e., little evidence; little posterior updating) replications can bring about considerable informational gain. The value of intuitions about a priori chances of replication success is corroborated by recent prediction market studies (e.g., Camerer et al., 2018; Dreber et al., 2015; Forsell et al., 2018; Tierney et al., 2021). In these studies, researchers in psychology were asked to estimate and bet money on the outcomes of large-scale replication projects such as Many Labs 2 (R. A. Klein et al., 2018) and the Science and Nature Replications (Camerer et al., 2018). Interestingly, prediction market bets and survey beliefs about the likelihood of a study being replicated, were highly correlated with actual replication effect sizes (see also https://replicats.research.unimelb.edu.au/ for more information on prediction markets studies). It thus seems important to discuss the selection of to-be-replicated studies with experts in the field beforehand, as their intuitions may be informative (although it seems that even laypeople can to some extent predict replication outcomes; Hoogeveen, Sarafoglou, and Wagenmakers, 2020).

For the CCRRP, we particularly aimed to conduct conceptual replications of influential theories and effects in the psychology of religion that are relevant to be studied cross-culturally. In general, we could recommend the 'replication+' approach, as it not only reinforces previous findings, but can also offer interesting new data and perspectives. For instance, multi-lab projects can assess the cross-cultural universality or variability plus boundary conditions of some phenomenon.

A relevant illustration of such a replication+ approach can be found in the project by Tierney et al. (2021), in which a 'creative destruction approach to replication' is adopted. Here, the original hypothesis is not only compared to the null-hypothesis, but also to various alternative theoretically relevant accounts. This work tested the implicit puritanism theory which holds that Americans are unique in their implicit moralization of work and the link between work and sex ethics, as a heritage of Puritan-Protestant settlers (Uhlmann et al., 2009; Uhlmann et al., 2011). The replication found evidence for some core effects, yet these effects emerged across all included cultures instead of exclusively in the US. For instance, targets who continued working after winning the lottery were evaluated more positively than targets who retired (i.e., a moral praise of needless work effect), and lazy targets are more often misremembered as promiscuous than hard-working targets (and vice versa; a tacit sex and work link effect). In other words, while previous effects were not replicated

**Table 2.2:** Replication Script

| Stage | Recommendation |
|---|---|
| Selecting | Opt for studies with medium chances of replication success.<br>Consult experts in the field for their suggestions and intuitions.<br>Investigate possibilities for replication+ projects that replicate and extend previous work in interesting ways (e.g., boundary conditions or cross-cultural universality). |
| Planning | Possibly: seek collaboration with colleagues in the field, for instance with authors of the original study.<br>In cross-cultural projects: ask for feedback on cultural appropriateness of experimental materials.<br>Preregister the research questions, hypotheses, methods, and analysis plan.<br>Consider a Registered Report format. |
| Executing | Collect data.<br>Possibly: use analysis blinding to retain flexibility yet avoid biases.<br>Conduct analyses according to preregistered plan, and explore data for interesting patterns. |
| Reporting | Write up results and invite discussions from scholars in the field.<br>Share annotated data and code. |

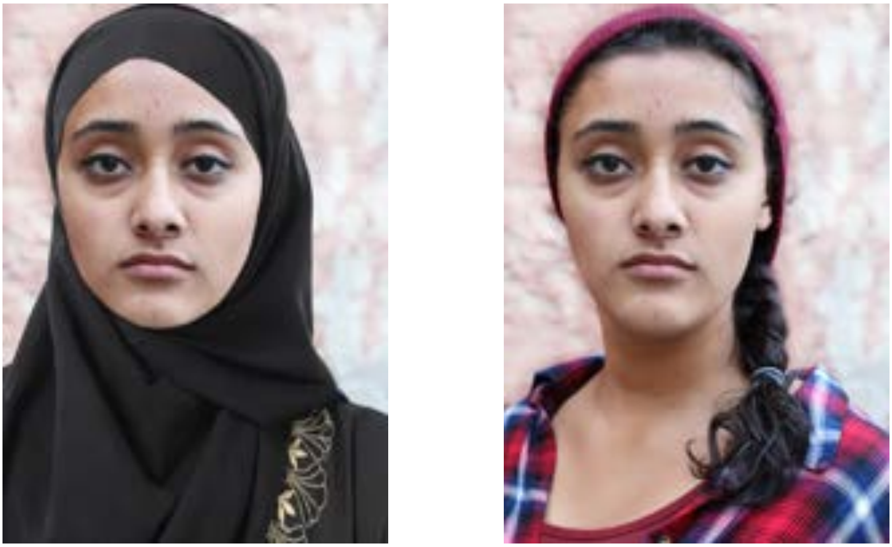
*Note.* This script was inspired by the summary of guidelines reported in van Doorn, van den Bergh, et al. (2020).

successfully according to the most stringent criteria for replication because no cultural differences were found, the study did boost confidence for the -potentially universal-presence of some of the general effects and provided new theoretical insights.

Once settled on a target study and potential extension, we would advise to reach out for collaborations. In our experience, many researchers in the CSR / psychology of religion that we approached were enthusiastic and eager to collaborate on projects like this.

In general, we believe the ideal of collaborative science is increasingly embraced by the scientific community (Chartier et al., 2018), as evidenced by the proliferation of large collaborative projects such as ManyLabs (Ebersole et al., 2016; R. A. Klein, Ratliff, Vianello, Adams, et al., 2014; R. A. Klein et al., 2018), ManyBabies (Frank et al., 2017; The ManyBabies Consortium, 2020), and the Psychological Science Accelerator (Moshontz et al., 2018).

For direct replication attempts it may also be relevant to invite authors who were involved in the original study, in order to optimize the study design and to avoid getting into an argument after the study has been conducted. For what it is worth, we had a very good personal experience of conducting a direct replication in collaboration with the original author for Hoogeveen, Wagenmakers, et al. (2018), although post-hoc disputes have also occurred, see for instance the academic preprint interaction about the ManyLabs 4 replication of the mortality salience effect, where the replicators reported replication failure, proponents of the theory concluded that the effect was



(a) Religious stimulus

(b) Control stimulus

**Figure 2.1:** Example stimuli of a religious and non-religious looking female for Arabic countries as used in the Cross-Cultural Religious Replication Project.

meaningfully present in one specific preregistered subgroup, and opponents argued that there was strong overall evidence against the effect and no evidence for the specific subgroup (Chatard et al., 2020; Haaf et al., 2020; R. A. Klein et al., 2019).

Besides sharing costs and effort, another advantage of multi-lab studies is that more diversity in the populations for data collection is achieved. International collaboration will likely also benefit the quality of the study design and will optimize choices for the phrasing of different questions, statements and stimuli. For instance, we were confronted with the cultural-specificity vs. universality dilemma in the creation of the stimuli for our project. Measures of religious beliefs and behaviours require considerable cultural fine-tuning. Our advice would be to use cross-culturally validated questionnaires and stimulus material where possible. Alternatively, it would be wise to consult scholars in the respective countries and field sites. As an example, we had discussions with an anthropologist whether it would make sense to use the same photographic stimulus of an “ethnically ambiguous-looking person” across all countries – it would not (see Figure 2.1 for an example of the culture-specific stimuli that were subsequently created for the CCRRP). Familiarity with the target population and the possibility to provide feedback with this knowledge in mind can probably improve the study validity substantially. In our case, it most certainly did.

As mentioned before, for any replication study the research questions, hypotheses, methods, and analysis plans should be preregistered. For a detailed argument and guidelines on preregistration, see Kavanagh and Kapitany (2017). The format of a Registered Report may be especially suitable for replication projects (i.e., a ‘triple R’). Here, the preregistration is integrated with the peer-review process; the introduction, methods, and proposed analysis plan are reviewed prior to data collection (Chambers, 2013). Upon approval by the reviewers and editor, the study proposal can receive *In-Principle Acceptance*, allowing it to be published regardless of the eventual study outcomes. Note that the International Journal for the Psychology of Religion now also offers the Registered Report format (van Elk, 2019).

While a preregistration forces one to specify design and analysis choices beforehand, in practice, often one still has to deviate from the plan when conducting the actual data analysis. A recent investigation found that in fact 27 out of 27 preregistered studies published in Psychological Science deviated from their corresponding plan (Claesen et al., 2021). Deviations are not by definition problematic, as there may be valid reasons to change plans, as long as they are transparently reported (DeHaven, 2017). Indeed, we already anticipate deviations from our preregistration for the CCRRP.

In the current project, we also incorporated an additional protective layer against any unconscious data-driven bias affecting the results, namely the notion of *blinded analyses*. We believe this can be of interest in many (large-scale) replication projects, or any study for that matter. Analysis blinding involves the temporary distortion of certain aspects of the data, for example by masking condition labels, adding noise, or shuffling key variables in the data (Dutilh, Sarafoglou, et al., 2019; MacCoun & Perlmutter, 2015). It has been argued that the combination of preregistration and analysis blinding may present an optimal balance between transparency, rigor, and flexibility (MacCoun & Perlmutter, 2015). The crucial idea is that with blinded data, analyses can be conducted flexibly without the risk of contamination by (unintentional) confirmation biases and significance seeking, because the actual outcomes are hidden from the analyst. In the first phase, analyses can be adjusted to unexpected peculiarities in the blinded data, thereby retaining desirable flexibility that may be lost with strictly following the preregistered analysis plans. Only after the data analyst is satisfied with preprocessing and model specification (e.g., outlier removal, choosing the appropriate statistical model given its assumptions), the blind is lifted and the real data are analyzed in the final model. In addition to running prespecified confirmatory analyses, rich data sets may also be used for exploratory analyses - as long as the distinction between confirmatory and exploratory analyses is not blurred.

With respect to data analysis, we would additionally recommend the use of Bayesian statistics in the scientific study of religion (van Elk & Wagenmakers, 2017). We believe the benefits of using a Bayesian approach are huge: Bayesian statistics are intuitive and can easily be implemented with freely available and user-friendly tools such as JASP (JASP Team, 2019). General advantages include the fact that instead of dichotomizing results as ‘significant’ vs. ‘non-significant’, Bayes factors can quantify evidence on a continuous scale and distinguish between “absence of evidence” and “evidence of absence” (Dienes, 2014; Wagenmakers, Marsman, et al., 2018). Furthermore, Bayesian statistics provide an effective method to optimize data efficiency and analysis quality. This is especially advantageous for data collection in hard-to-reach populations such as young children and small-scale societies or data that relies on expensive materials and testing settings such as neuroscience studies (Nakagawa & Hauber, 2011).

In contrast to frequentist statistics, in Bayesian inference, online monitoring of the evidence does not inflate Type I error rates (Rouder, 2014; Wagenmakers, Marsman, et al., 2018). That is, if there is no effect, the *p*-value will randomly fluctuate between

1 and 0. So the more often one inspects the data, the higher the chance that at some point the $p$-value will be lower than .05. In the Bayesian framework, however, if there is no effect, the posterior model odds and Bayes factor will drift towards more and more evidence for the null-model, relative to the alternative model (see the preprint by Wagenmakers et al., 2019 for an accessible explanation of the stopping rule principle in Bayesian inference). We refer the interested reader to Wagenmakers, Love, et al. (2018) for a demonstration of Bayesian analyses for various standard statistical tests such as the $t$-test, ANOVA, contingency tables, and regression.

Finally, a cornerstone of reproducible science is that it is indeed comprehensibly reproducible. Issues with messy, illegible analysis scripts or nonfunctional data formats (e.g., a .pdf file) are prevalent, even with openly shared data (Hardwicke, Mathur, et al., 2018). Careful annotation of data files (i.e., meaningful variable names or an accompanying codebook), analysis scripts (preferably in open-source programs) and workflow is essential for reproducible science (e.g., Gilmore et al., 2018; Wilkinson et al., 2016). The initiative GO FAIR (www.go-fair.org/) picks up on this observation and aims to assist researchers in implementing the FAIR data principles of making data Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016).

## 2.6 Conclusion

We realize that the presented recommendations and suggestions may be somewhat overwhelming for researchers uninitiated in the open science movement. We are not suggesting that scholars in the CSR should completely revolutionize their workflow and adopt all outlined open science practices overnight. Rather, we would invite researchers to try it out. For instance, start by sharing the data of a just-finished project (see Gewin, 2016; O. Klein et al., 2018 for practical advice), or consider submitting unpublished null-findings as a file-drawer report to this journal (JCSR; see this call). When hesitant to immediately lead a large-scale replication project, consider first joining an existing project (e.g., via the Psychological Science Accelerator; https://psysciacc.org/). Another possibility is to integrate replication research with the student curriculum. For instance, the Collaborative Replications and Education Project (CREP; http://osf.io/wfc6u) is an international framework that allows students to work together on conducting direct replications of recent impactful studies (Wagge et al., 2019). This way, the initiative both serves an educational purpose and allows supervisory researchers to get a taste of the replication process and contribute to establishing the reliability of the literature.

Much attention has been paid to large-scale multi-lab replication projects that included various classical effects from psychological studies (e.g., Camerer et al., 2018; R. A. Klein, Ratliff, Vianello, Adams, et al., 2014; Open Science Collaboration, 2015). These projects provided the necessary impetus to fan the flames of the credibility revolution. At the same time, we would like to emphasize the importance of expert scholars actually conducting replication studies in their domain of expertise. Lack of theoretical and methodological knowledge should not be exploited as a – valid or invalid – excuse for replication failure. Moreover, active proponents of replication research are sometimes reproached with taking a skeptical or even cynical stand a priori, i.e., being motivated to show replication *failure*, rather than success. Therefore, replications should also be conducted by ‘expert insiders’.

We would thus like to encourage the community of CSR scholars to take action and invite them to join the bandwagon of open science and replication research. Doing replication studies is innovative, challenging, exciting and it provides a valuable learning experience for all involved.

# 3
# Laypeople Can Predict Which Social-Science Studies Will Be Replicated Successfully

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., 59%). 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.

## 3.1 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. A. Klein et al., 2018), 62% (Camerer et al., 2018) to 85% (R. A. 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 (G. 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

This chapter has been adapted from: 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. https://doi.org/10.1177/2515245920919667.
The first two authors contributed equally.

statistical analysis (i.e., significance seeking; John et al., 2012; Simmons et al., 2011; Wagenmakers et al., 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 predictions 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. A. 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, et al., 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. A. 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 et al., 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.

## 3.2 Disclosures

### 3.2.1 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.

### 3.2.2 Supplemental Material

In the online Appendix (https://osf.io/7cgfw/) we provide additional details on the methods and additional exploratory analyses. Specifically, the online supplement outlines details on the Bayesian reanalyses of the original studies, details on the sampling plan, the statistical models and prior specifications, a table with all study descriptions in English and Dutch as presented to the participants, and two additional exploratory analyses. The first of these analyses concerns the accuracy of predictions derived from the Bayes factors alone, without human evaluation. The second analysis presents a Bayesian logistic regression model that includes random effects for both participants and studies.

### 3.2.3 Reporting

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

### 3.2.4 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.

## 3.3 Methods

### 3.3.1 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 web-shop. 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.

### 3.3.2 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.

### 3.3.3 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. A. Klein et al., 2018).

#### 3.3.3.1 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, 2019), 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 online Appendix (https://osf.io/7cgfw/). 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%).

#### 3.3.3.2 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 et al., 2019). To prevent participants from reading up on 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).

### 3.3.4 Procedure

The survey was generated using the online survey software Qualtrics (Qualtrics, 2019). Participants were randomly assigned to the Description Only or the Description Plus

[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.

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 and 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]

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.

### 3.3.5 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. A. 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

[2] Unfortunately, this explanation fell prey to a prevalent misinterpretation of Bayes' rule (e.g., Wagenmakers, Etz, et al., 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.

[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

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.

### 3.3.6 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 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 online Appendix (https://osf.io/7cgfw/).

## 3.4 Results

### 3.4.1 Descriptive Pattern

Figure 3.1 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 et al., 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,

level).

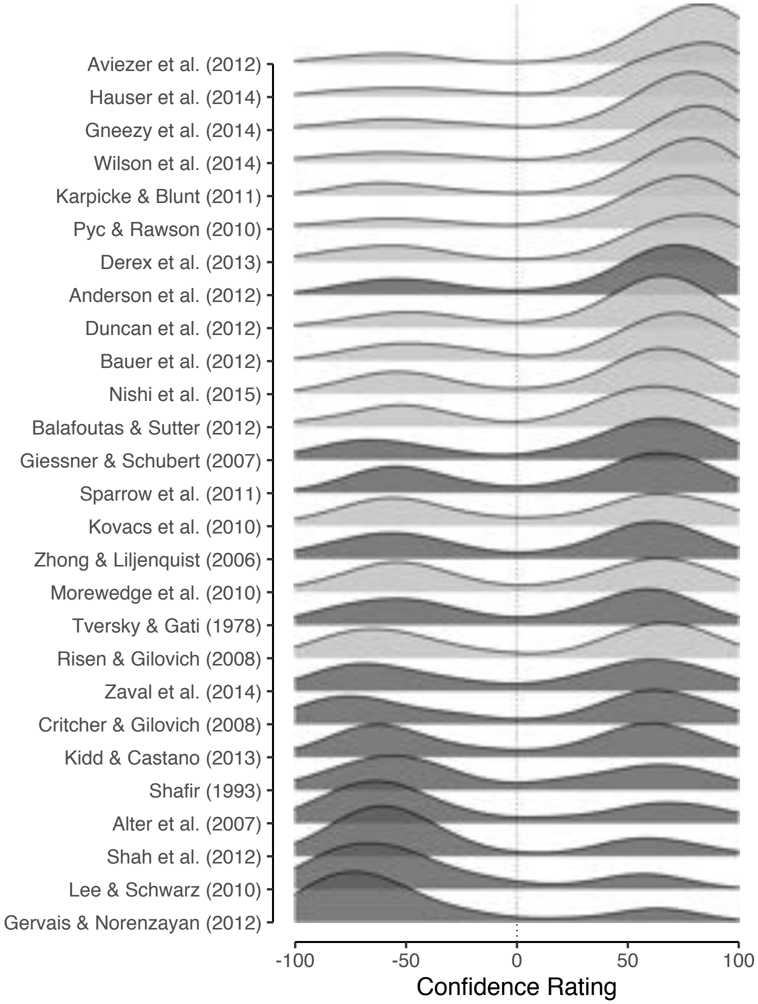
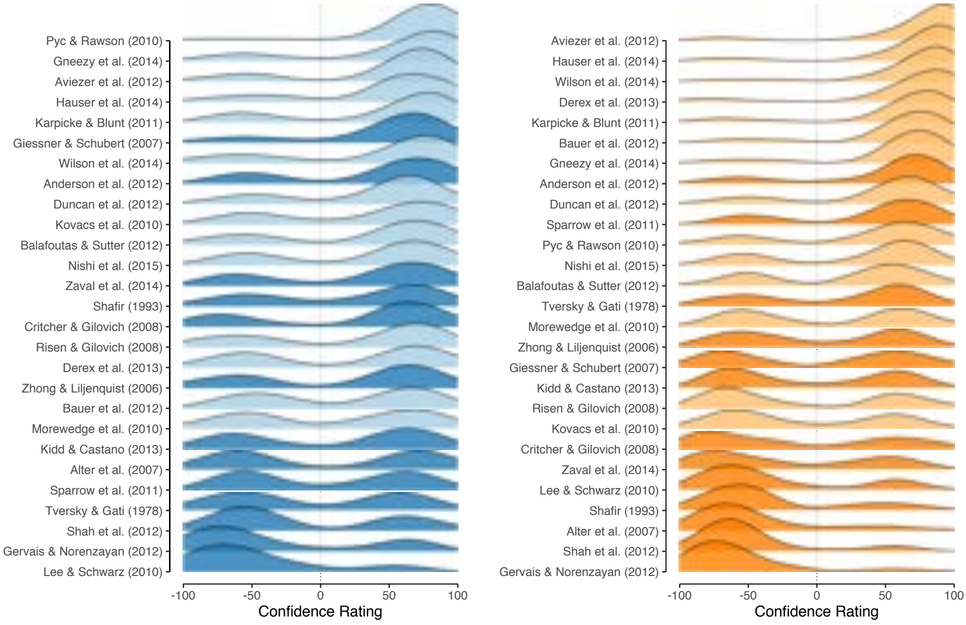


**Figure 3.1:** Laypeople's near unanimous judgments are highly predictive of replication outcomes. Light density distributions reflect studies that successfully replicated, dark grey distributions reflect studies that did not replicate. Confidence ratings are aggregated over both experimental conditions. Negative values reflect the 'does not replicate' prediction, and positive values the 'replicates' prediction.

five replicated successfully and eight failed to replicate successfully. Overall, Figure 3.1 provides a compelling demonstration that laypeople are able to predict whether or not high-profile social science findings will replicate successfully. In Figure 3.2 laypeople's predictions are separately displayed for the Description Only and the Description Plus Evidence condition.

Figure 3.3 provides a more detailed account of the data for three selected studies. For the study in the top panel (i.e., Gneezy et al., 2014), most laypeople correctly predicted that the effect would successfully replicate; for the study in the middle



**(a)** Description Only condition

**(b)** Description Plus Evidence condition

**Figure 3.2:** Laypeople's predictions about replication outcomes, separated per experimental condition. The left panel (in blue) displays predictions of people in the Description Only condition, the right panel (in orange) displays predictions of people in the Description Plus Evidence condition. Light density distributions reflect studies that successfully replicated, dark density distributions reflect studies that did not replicate. Studies are ordered according to the average confidence rating for each study.

panel (i.e., Tversky and 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 et al., 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 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 $\rho$ 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 ($\text{BF}_{+0} = 162$).[5] The median and 95% credible interval for the correlation coefficient $\rho$ were $0.62\,[0.33, 0.78]$, indicating that the Bayes factors of the original studies indeed

[5] The subscripts on the Bayes factor 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.

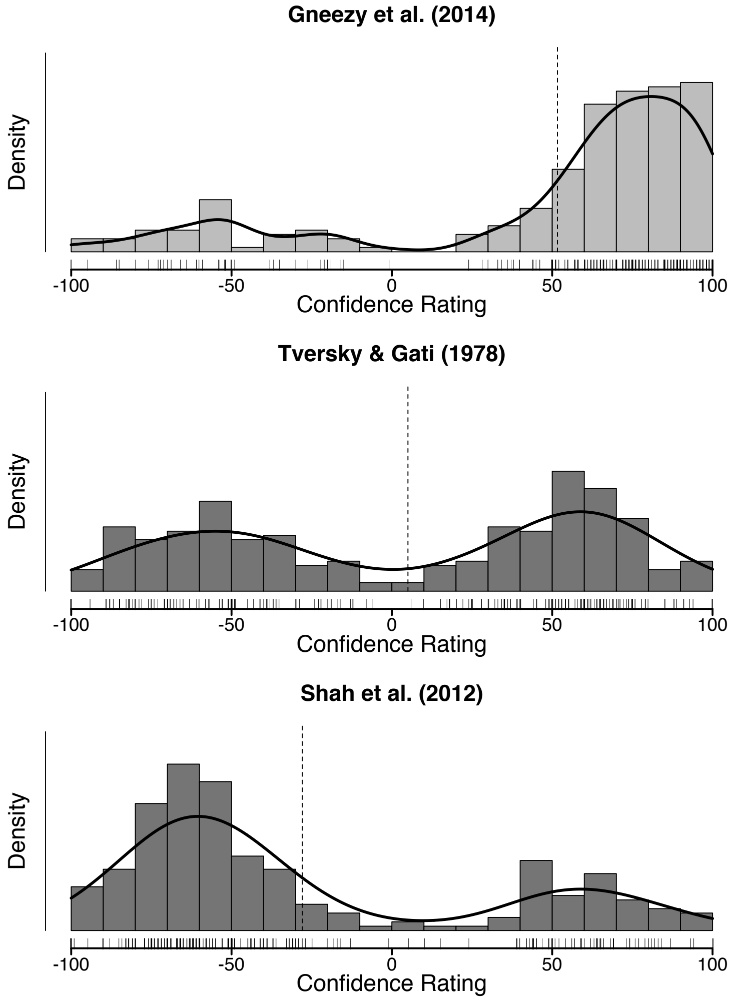


**Figure 3.3:** Histograms of confidence ratings for three studies for which laypeople were nearly unanimous in their belief that the study will either replicate (Gneezy et al., 2014, top panel) or will not replicate (Shah et al., 2012, bottom panel) or for which they are ambiguous (Tversky and Gati, 1978, middle panel). The vertical dotted line shows the average confidence rating for the respective study (i.e., group prediction).

conveyed useful information (see Figure 3.4).

### 3.4.2 Preregistered Analyses

#### 3.4.2.1 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

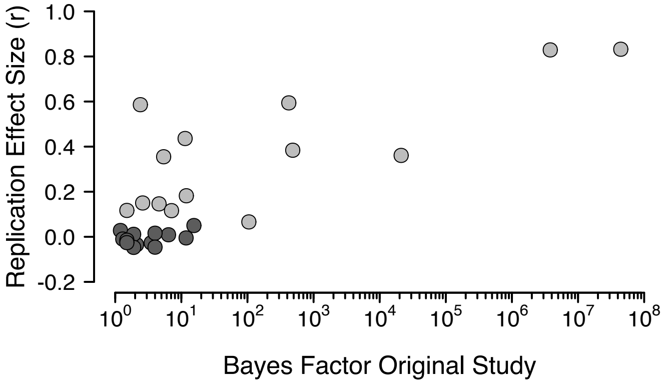


**Figure 3.4:** The evidence of the original studies (quantified by Bayes factors) is positively associated with replication effect sizes. The dark grey dots indicate the studies that did not replicate, the light grey dots indicate the studies that did replicate.

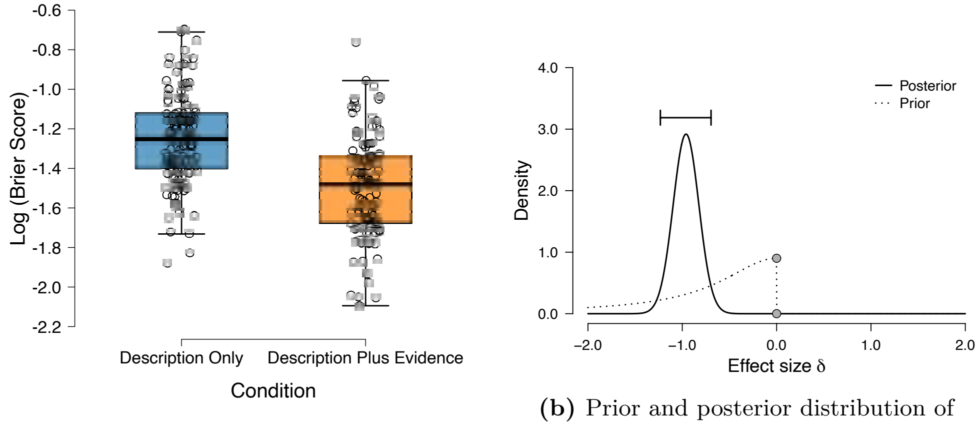
subsequent analyses in case the data provided strong evidence for the hypothesis that overall laypeople performed *worse* than chance level when predicting the replicability of empirical studies. An accuracy rate that is worse than chance level (i.e., less than 50%) indicates that participants either 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 $\omega$ is the mode of the Beta distribution for the group-level accuracy rate. This hypothesis was tested against the encompassing hypothesis $\mathcal{H}_e$ which lets $\omega$ free to vary, that is $\mathcal{H}_e : \omega \sim \text{Beta}(1, 1)$. The Bayes factor in favor for the encompassing hypothesis, $\text{BF}_{er1}$, was computed using the encompassing prior approach (Klugkist et al., 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]

#### 3.4.2.2 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

[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.



**(a)** Boxplot of log Brier scores per condition.

**(b)** Prior and posterior distribution of population effect size $\delta$.

**Figure 3.5:** The data and distribution of effect size $\delta$ 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 (orange boxplot) performed better than laypeople who received the study descriptions only (blue boxplot). Figure (b) was created in JASP (JASP Team, 2019).

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 $\delta$ 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 $[-1.23, -0.68]$ (see Figure 3.5 for a boxplot of the data as well as the prior and posterior distribution of the effect size $\delta$).

### 3.4.2.3 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 $\omega$ (i.e., the mode of the group-level distribution) was 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, $\mathcal{H}_{r2}, \mathcal{H}_{r3} : \omega > 0.5$.

[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).

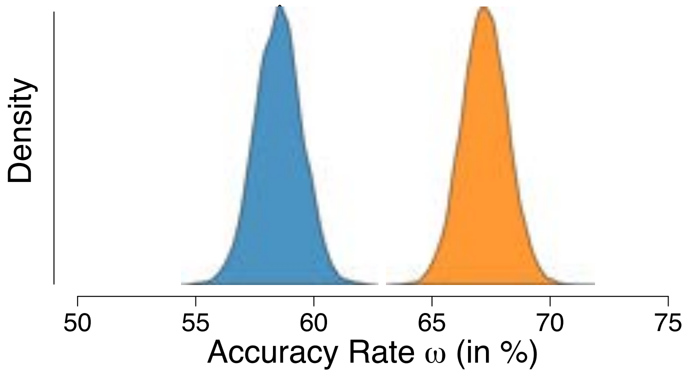


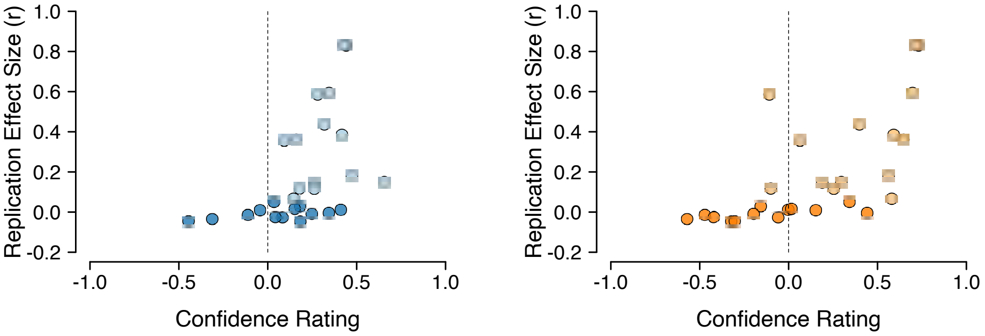
**Figure 3.6:** Accuracy rates of laypeople in both conditions. Posterior distributions of the group-level accuracy rate for laypeople in the Description Only condition are depicted in blue and those of laypeople in the Description Plus Evidence condition are depicted in orange.

The hypotheses $\mathcal{H}_{r2}$ and $\mathcal{H}_{r3}$ were tested against the null hypothesis $\mathcal{H}_0$ stating that $\omega$ 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, $\text{BF}_{r20} = 4.4 \times 10^7$. The median and 95% credible interval for the parameter $\omega$ are $0.59\,[0.57, 0.60]$, which implies a 59% 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, $\text{BF}_{r30} = 5.6 \times 10^{22}$. The median and 95% credible interval for the parameter $\omega$ 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 3.6.

#### 3.4.2.4 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



**(a)** Description Only condition

**(b)** Description Plus Evidence condition

**Figure 3.7:** Relationship between the average confidence rating per study and the replication effect size for the Description Only condition (in blue) and the Description Plus Evidence condition (in orange). 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 refer to studies that did not replicate, and the light dots refer to studies that did replicate.

effect sizes were retrieved from Camerer et al. (2018) and R. A. Klein et al. (2018). The data are plotted in Figure 3.7, displayed per condition.

We used a Bayesian Spearman correlation (van Doorn, Ly, et al., 2020) to test the null hypothesis (i.e., $\mathcal{H}_0 : \rho = 0$) against the one-sided restricted hypothesis that the correlation coefficient $\rho$ 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 ($\text{BF}_{r40} = 523$) and the Description Plus Evidence condition ($\text{BF}_{r50} = 14295$). For the Description Only condition the median and 95% credible interval for the distribution of the Spearman correlation coefficient $\rho$ are $0.61\,[0.34, 0.77]$. For the Description Plus Evidence condition the median and 95% credible interval for the distribution of $\rho$ are $0.77\,[0.57, 0.87]$. 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.

### 3.4.3 Exploratory Analyses

#### 3.4.3.1 Disentangling Discriminability and Response Bias

According to signal detection theory (SDT; Green and Swets, 1966; Tanner Jr and 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

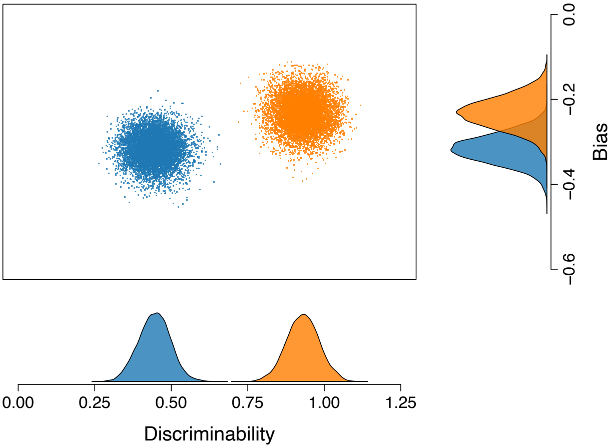


**Figure 3.8:** Laypeople in both conditions are biased towards predicting that a given study will replicate (as indicated by the posterior distributions of the bias parameter in the right panel). In addition, the posterior distributions of the discriminability parameter in the bottom panel show that 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).

*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 & Wagenmakers, 2013, p. 164).

Figure 3.8 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 (orange in the figure) are better at separating replicable studies from unreplicable studies than people in the Description Only condition (blue in the figure). The enhanced discriminability for the Description Plus Evidence condition is also visualized in the top panels of Figure 3.9, 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 3.8, 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 3.9: in both conditions, the adopted criterion is located to the left of the optimal criterion.

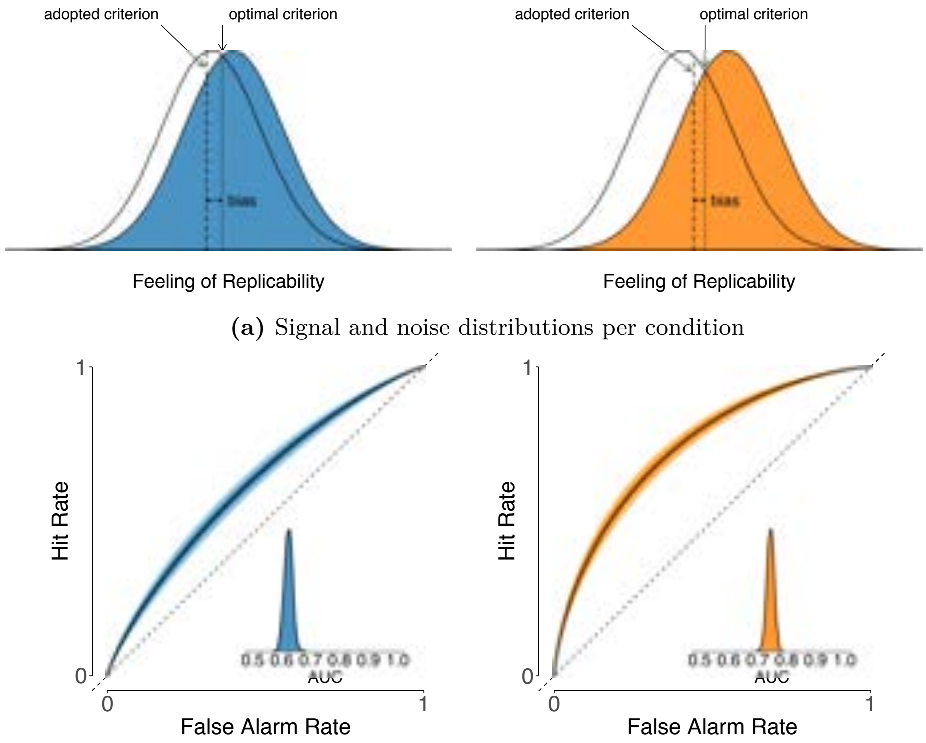
The Receiver Operating Characteristic (ROC) curve is often used to interpret the parameter values of the SDT. This curve reflects the proportion of hits (i.e., replication successes that were deemed replicable) and false alarms (i.e., replication failures that were deemed replicable) as a function of 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 successes 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 3.9 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.

#### 3.4.3.2 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 3.10. 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 successfully. Because of these unequal proportions, that are also not representative for the respective projects,



**(a)** Signal and noise distributions per condition

**(b)** Group-level ROC curves per condition

**Figure 3.9:** 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).

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.

## 3.5 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. Since the predictions were generated by non-experts on the sole basis of simple verbal study descriptions, we took these predictions to reflect intuitions of study plausibility. As such, our results suggest that intuitions

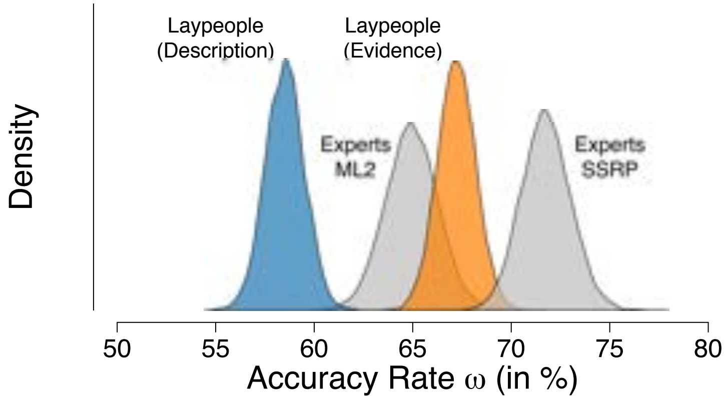


**Figure 3.10:** Accuracy rates of laypeople and experts. Posterior distributions of the group-level accuracy rates for laypeople in the Description Only condition are displayed in blue and for laypeople in the Description Plus Evidence condition in orange. Posterior distributions of the group-level 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's confidence in replicability was associated with replication effect sizes. This may indicate that laypeople were able to derive a sense of the magnitude of the targeted effects from the descriptions. 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. B. M. 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"; B. M. Wilson and Wixted, 2018, p. 191). By asking laypeople about their intuitions regarding the replicability of social science studies, our study sought to shed light on exactly this underlying feature of unreplicable studies derived from the latter source of knowledge, which we called "intuitive plausibility", "surprisingness", or "unexpectedness". Although we did not assess plausibility of the studies directly, we believe laypeople's intuitions regarding the studies' replicability can serve as a close approximation. 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.

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. A. 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, 2015a; 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. A. 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 59% 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 3.1 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 3.1. 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 3.4 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.

### Constraints on Generality

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.

# Part II

# Replicating Key Effects in the Psychology of Religion (Or Not)

# 4

# Compensatory Control and Belief in God: A Registered Replication Report Across Two Countries

Compensatory Control Theory (CCT) suggests that religious belief systems provide an external source of control that can substitute a perceived lack of personal control. In a seminal paper, Kay et al. (2008) experimentally demonstrated that a threat to personal control increases endorsement of the existence of a controlling God. In the current registered report, we conducted a high-powered ($N = 829$) direct replication of this effect, using samples from the Netherlands and the United States. Our results show moderate to strong evidence for the absence of an experimental effect across both countries: belief in a controlling God did not increase after a threat compared to an affirmation of personal control. In a complementary preregistered analysis, an inverse relation between general feelings of personal control and belief in a controlling God was found in the US, but not in the Netherlands. We discuss potential reasons for the replication failure of the experimental effect and cultural mechanisms explaining the cross-country difference in the correlational effect. Together, our findings suggest that experimental manipulations of control may be ineffective in shifting belief in God, but that individual differences in the experience of control may be related to religious beliefs in a way that is consistent with CCT.

## 4.1 Introduction

Why do so many people across the world believe in a supernatural being that can exert a causal influence on human affairs? Why do they engage in time-consuming rituals to ask an invisible entity for a favor or blessing? Take for instance a devout Catholic who prays to God for healing her sick son. Or consider a Hindu who offers valuable goods to his deities in order to obtain their blessing. According to Compensatory

This chapter has been adapted from: Hoogeveen, S., Wagenmakers, E.-J., Kay, A. C., & van Elk, M. (2018). Compensatory control and religious beliefs: A registered replication report across two countries. *Comprehensive Results in Social Psychology*, *3*(3), 240–265. https://doi.org/10.1080/23743603.2019.1684821.

Control Theory (CCT), in all these cases people try to gain a sense of control over their environment through religious beliefs and practices.

The basic rationale of CCT holds that believing in the power of God or other supernatural agents can compensate for the feeling that one lacks personal control over important life outcomes, and hence may partially alleviate the uncomfortable feeling elicited by uncertainty and randomness (Kay, Gaucher, McGregor, et al., 2010; Kay et al., 2008; Landau et al., 2015). Indeed, humans have a deep-rooted desire for personal control; we are reluctant to accept randomness and inclined to believe that we can at least to some extent predict, influence, and control the world around us (Lerner, 1980; S. F. Maier & Seligman, 1976). Yet situational constraints and the complex reality of our environments often substantially reduce the degree to which we can perceive ourselves as being in control. To alleviate this discomfort, individuals can attempt to reaffirm personal control directly through their own actions, e.g., by performing superstitious rituals (Whitson & Galinsky, 2008). Alternatively, when exerting personal control is impossible, individuals might resort to external sources of control, for instance by affiliating with a societal institution, a governmental system, or a religious ideology (Rothbaum et al., 1982).

More specifically, CCT posits that in the face of low or reduced personal control, people will restore their feeling of control by more strongly endorsing belief in the existence and influence of external controlling powers, such as an intervening God. In the classic demonstration of this effect, introduced by Kay et al. (2008) in Study 1, participants were assigned to either a control affirmation or control threat condition in which perceived personal control was strengthened or reduced, respectively, by means of an autobiographic recall task. Subsequently, participants indicated their belief in a controlling God. As predicted, the results of the original study revealed that participants whose perception of personal control was threatened showed a significantly stronger belief in the existence of a controlling God, compared to participants whose personal control was affirmed. Notably, when the controlling nature of God was deemphasized, i.e., God was presented as a creator, the control threat effect was absent. This dissociation underlines the relevance of religious beliefs providing a source of compensatory control, rather than being comforting in general. As noted by the authors, given that profound beliefs such as those associated with religion and supernatural beings are highly stable and difficult to manipulate experimentally (e.g., Yonker et al., 2016), the fact that this simple control manipulation is capable of "shift[ing] these beliefs is rather striking" (Kay et al., 2008, p. 23).

CCT has been supported by many empirical findings and is accepted as an important psychological and motivational account with respect to religious beliefs (Sedikides, 2010). According to Google Scholar, as of September 2018, the paper by Kay et al. (2008) has been cited 602 times. More importantly, the original article inspired a large body of research on compensatory control mechanisms related to a wide variety of structure-restoring tendencies (reviewed in Landau et al., 2015). The breadth of the phenomenon can be illustrated by the variety of research approaches (e.g., temporal fluctuations on a national level, individual differences, experimental manipulations) as well as the range of examined compensation strategies: in correlational designs, lack of personal control has been associated with stronger attraction to astrology (Lillqvist & Lindeman, 1998), stronger endorsement of conspiracy beliefs (Newheiser et al., 2011), higher levels of superstition (Padgett & Jorgenson, 1982), and higher conversion rates to authoritarian relative to non-authoritarian churches (Sales, 1972). In an experimental context, personal control manipulations have been shown to affect illusory pattern perception and conspiracy beliefs (Whitson & Galinsky, 2008), endorsement of horoscope descriptions (C. S. Wang et al., 2012), belief in precognition (Greenaway et al., 2013), support for meritocratic systems (Goode et al., 2014), preference for structured consumption items (Cutright, 2011), belief in the efficacy of rituals (Legare & Souza, 2014) and belief in order-providing theories such as a predictable, non-random version of evolution theory (Rutjens et al., 2010). In a recent meta-analysis by Landau et al. (2015), including 55 studies, it was established that control-threat manipulations exerted a moderate ($r = .24$, $\delta = .494$) though robust effect on different 'epistemic structuring tendencies'.

The primary finding by Kay et al. (2008) with belief in a controlling God as the dependent variable has indeed been replicated (either successfully or unsuccessfully) in seven studies – however always as part of more elaborate designs or additional research questions. Figure 4.1 summarizes the replications of the crucial effect of personal control threat on belief in a controlling God, including a model-averaged Bayesian meta-analysis. The top row of the figure refers to the original study, and the subsequent rows list existing replications. Across all studies, the outcome variable was 'belief in a controlling God', measured by the items specified in the Methods section of this chapter.

Note that the figure displays results of the main experimental effect of the control manipulation on belief in a controlling God, although the listed studies' primary interest in some cases focused on different aspects. The studies investigated for instance the role of specific mediators (defensive reactions towards randomness; Kay et al., 2008, Study 2), moderators (anxiety; Laurin et al., 2008[1]; Kay, Moscovitch, et al., 2010, a personality trait related to independence and desire for autonomy; Alper and Sümer, 2017), included additional conditions (a neutral condition; Verburg et al., 2016[2]) or a different source of control was manipulated (governmental control; Kay, Shepherd, et al., 2010).

We conducted a Bayesian reanalysis and meta-analysis (Gronau, van Erp, et al., 2017; Scheibehenne et al., 2017) of the previous findings to assess the strength of the evidence provided by the replications of the primary effect.[3][4] As can be seen in Figure

[1] The authors only reported statistics for the main effect for anxiety and the anxiety-personal control interaction, but omitted results for the main effect of control on belief in God. Therefore, the result of this replication cannot be quantified. The figure on page 1561, however, suggests that the main effect is not significant.

[2] Importantly, this study was presented on poster that reported only the $F$-values and $p$-values ($F(2, 151) = 30.11, p < .001$), illustrated with a graph of the descriptives. Notably, although belief in God is reportedly measured on a 7-point Likert scale, based on visual inspection of the graph, the mean of the control threat condition appears to be approximately 7.8. We approached the authors to validate these results and request descriptive statistics per group, but we did not receive a reply. Therefore, some caution is warranted in evaluating this finding.

[3] As the paper by Laurin et al. (2008) did not include any statistics on the main effect of the personal control manipulation on belief in God, we were not able to calculate the Bayes factor for this study.

[4] Bayes factors were calculated based on the $F$-value converted to $t$-value and sample size reported in the original studies, using the `meta.ttestBF` function of the package `BayesFactor` (one-sided) in R with default priors (Morey & Rouder, 2015). The exact number of participants per group was not reported in any of the studies, and we therefore assumed that participants were uniformly distributed over conditions.

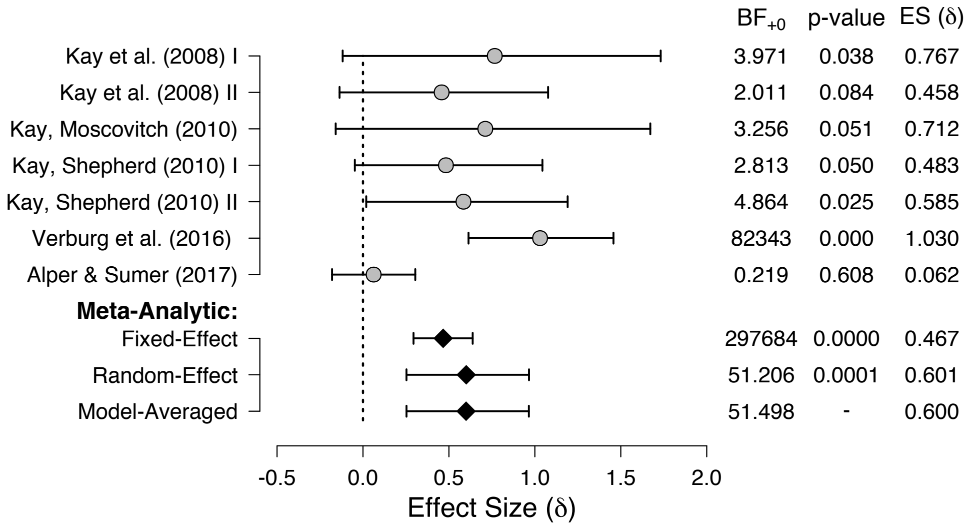


**Figure 4.1:** Summary of previous studies plus meta-analysis on the effects of control threat on belief in a controlling God. The Bayes factor $\mathrm{BF}_{+0}$ quantifies the evidence that the data provide for $\mathcal{H}_+$ (i.e., presence of the compensatory control effect) relative to $\mathcal{H}_0$ (i.e., absence of the compensatory control effect). Created using the `metaBMA` R package (Gronau, van Erp, et al., 2017; Heck & Gronau, 2017). Exact $p$-values were not always given in the articles, but were recalculated based on the reported statistics, converting the one-way ANOVA $F$-values to $t$-values. For Kay, Shepherd, et al. (2010), the misattribution (i.e., no anxiety) condition is excluded.

4.1, the data from most studies provide only weak evidence for the effect of personal control threat on belief in a controlling God. Specifically, based on the commonly used interpretation categories of Bayes factors (e.g., Lee and Wagenmakers, 2013, p. 105; Jeffreys, 1939), the studies by Kay et al. (2008, 2010, 2010) all yielded evidence that is considered anecdotal to moderate. Indeed, only the findings by Verburg et al. (2016) yielded compelling evidence for the control threat effect on religious belief. The study by Alper and Sümer (2017), on the other hand, appears to provide moderate evidence *against* the presence of the effect.

Overall, our Bayesian meta-analysis indicates strong evidence in favor of the presence of a control threat effect on belief in a controlling God. However, our meta-analysis also suggests that there is substantial heterogeneity; the random effects model has far more predictive adequacy than the fixed effects model, hence the averaged model is primarily determined by the random effects model. Furthermore, the credible interval of the average meta-analytic effect size is rather large; CI ranges from 0.250 to 0.962, with a median of $\delta = .600$. This further supports the motivation to conduct the high-powered proposed replication study. In conclusion, in spite of the theoretical and empirical support for the CCT as an overarching framework, the evidence for the primary effect regarding belief in God is not as unequivocal as one might have assumed.

## 4.2 Motivation

Given the impact of the original study, it is quite surprising that, to our knowledge, there have not been any high-powered direct replications of the effect of the personal control manipulation on belief in God. Our primary motivation for the current replication attempt thus naturally arises from the influential status of the study, reflected in the large body of research and theoretical reviews that it inspired. Secondly, all but the last two of the replication experiments used sample sizes smaller than $n = 50$, which translates into a maximum of 25 participants per group. In fact, the original effect was established based on 9 participants per group. We used the meta-analytic effect size of $\delta = .539$ ($r = .24$) reported by Landau et al. (2015), as well as a corrected estimate of $\delta = .379$ ($r = .186$) reported by van Elk and Lodder (2018) to calculate the achieved power of the original study. That is, van Elk and Lodder (2018) report that the standard errors of the studies included in the meta-analysis by Landau et al. (2015) have been overestimated, possibly due to a coding error. As a consequence, the funnel plot asymmetry and hence the amount of missing studies and the extent of publication bias are underestimated. Crucially, whereas the original meta-analysis found no indication for publication bias and reported a final overall effect size of $r = .24$, the reanalysis by van Elk and Lodder yielded an initial effect size of $r = .26$ that should be adjusted to $r = .186, p < .0001$, indeed still reflecting a small to medium but robust effect.[5] The post-hoc power analyses based on the effect size of the meta-analysis, as well as on its corrected version, indicate that the original study was indeed highly underpowered (achieved power = 0.17 or 0.12, respectively).[6]

Moreover, the previous studies all used frequentist significance tests. Although in many cases, both frequentist and Bayesian analyses yield the same conclusions, we have some arguments for why we believe Bayes factors are favorable over $p$-values (see Wagenmakers, Marsman, et al., 2018 for an elaborate argumentation): First, whereas frequentist statistics solely allows one to either reject or fail to reject the null hypothesis, Bayesian analyses can additionally distinguish between 'absence of evidence' and 'evidence of absence' (Dienes, 2014). This seems highly relevant in social psychological research, where effects of interest are generally of a small-to-medium size (Wagenmakers et al., 2016), and perhaps even more so for replication studies (Wagenmakers, Marsman, et al., 2018). Based on the meta-analysis (Landau et al., 2015), we indeed expect a small to medium effect for the current control-threat effect ($\delta = .38$). Second, we believe Bayes factors are intuitive; they arguably do what we desire (and assume) statistical tests to do. That is, they allow us to quantify the evidence that the data provide for $\mathcal{H}_0$ versus $\mathcal{H}_1$. As such, Bayes factors provide a direct comparison between the two hypotheses, conditional on the observed data (e.g., Jeffreys, 1939). Frequentist $p$-values, on the other hand, are calculated conditional on the null hypothesis being true; predictions of the alternative hypothesis are irrelevant and not taken into account in the evaluation. Third, and relatedly, Bayes factors only rely on data that were actually observed, rather than hypothetical data. In contrast, $p$-values are defined as the probability of obtaining the obtained results – *or more*

[5] See van Elk and Lodder (2018, pp. 29-31) for a detailed description on the error in the original meta-analysis and their reanalysis.

[6] Achieved power was calculated with G*Power 3.1, using the sample size ($n = 18$) and $F$-value ($F = 5.12$) of the experiment by Kay et al. (2008) and the converted meta-analytic effect size of $f = 0.247$ (original) and $f = 0.189$ (corrected; Faul et al., 2007).

*extreme results* – given that the null hypothesis is true, thus basically conditioning on the data plus hypothetical data that have not been observed at all (Berger & Wolpert, 1988).

Nevertheless, we are aware of the arguments against the use of Bayesian frameworks, and Bayes factors specifically (e.g., Gelman and Shalizi, 2013). For instance, Bayesian inference does not solve some of the issues associated with null hypothesis significance testing; in large samples, even small and practically meaningless effects will also generate 'strong evidence'. However, meaningfulness can never be resolved by statistical analysis; it is always a context-dependent concept that deserves a scholarly discussions by experts in the field. From the other end of the spectrum, it has been argued that Bayes factors are biased *against* small effects (Simonsohn, 2015b). However, this only applies under the combination of (1) a small sample size; (2) a small true effect size; and (3) a prior distribution that represents the expectation that effect size is large. Indeed, in the present study, we precluded (1) and (3), so we are confident that our analysis is not prejudiced against finding an effect. Therefore, given the listed advantages, we will analyze the data of the present replication study in a Bayesian framework.[7]

Furthermore, it is important to determine the effectiveness and validity of the experimental manipulation (i.e., control threat vs. control affirmation). Particularly, we included manipulation check items (e.g., "To what extent do you feel like you are the one who is in control in your life?") to test whether the control threat manipulation indeed affected feelings of personal control in one's life. In other words, we considered the manipulation effective if the affirmation of control results in higher ratings of feelings of general personal control relative to threats to control, at the group level. Importantly, the inclusion of these manipulation check items additionally allowed us to adopt an individual differences approach in case the experimental manipulation turned out to be ineffective. That is, we hypothesized that a lower feeling of general control in one's life would be related to a stronger belief in a controlling God – irrespective of the experimental manipulation. In this way, any ambiguity regarding the interpretation of an eventual null result, i.e., the inadequacy of the manipulation or the absence of a compensatory control effect, could be eliminated.

Whereas self-reported religiosity was included as a covariate in the original study, we additionally assessed religiosity as a moderator of the experimental effect. That is, while Kay et al. (2008, p. 23) mention that "the manipulation of personal control did not significantly affect this covariate [i.e., religiosity]", we argue that it may nevertheless moderate the effect of control threat on belief in a controlling God. More specifically, belief in a controlling God may be an especially appealing substitute for personal control for those who are (at least somewhat) religious, whereas God's control might not be considered an alternative among atheists, similar to effects of religious priming only affecting religious individuals (Shariff, Willard, Andersen, et al., 2016). However, as we did not find any studies examining the potential moderating role of religiosity on the effect of control threat on belief in God, we left this possibility open and investigated it only exploratorily.[8]

[7] For more details and extended discussion on Bayesian inference we recommend the recent special issue of Psychonomic Bulletin & Review (Etz & Vandekerckhove, 2018).

[8] In an opposite but complementary fashion, Cutright (2011) showed in Study 6 that religiosity moderated the effect of control threat on the tendency to prefer bounded relative to unbounded

Finally, with one exception (i.e., Alper & Sümer, 2017), all previously studied populations are from North America, mostly the United States (US). Besides the moderating effect of religiosity at the individual level, we expected additional cross-country differences between the Netherlands and the US, for two intertwined reasons. That is, the tendency to resort to belief in God as a source of control may vary between countries due to (1) differences in the cultural prevalence of religious beliefs and (2) the availability of alternative secular sources of security and control.

The Netherlands can be defined as a highly secularized country; national statistics indicate that as of 2015, only 12% of the Dutch population regularly attended church, and 32% believed in a personal God or a higher power (Bernts & Berghuijs, 2016; Kregting et al., 2018). Thus, although some people can still be considered religious, the majority of the Dutch population do not endorse traditional religious beliefs in God. This stands in contrast with the US, which can be considered a highly religious country, where the majority of the population endorses traditional religious belief in a powerful, intervening, and controlling God (i.e., as of 2016, 79% of the US citizens indicated to believe in God; Gallup, 2016; Stavrova et al., 2013).

The development of social security has proven to be a relevant factor in explaining secularity over time in Western countries (Kregting et al., 2018; Reitsma et al., 2012). For instance, predictors of existential securities (i.e., political, material, and financial security) and religious socialization and control (i.e., being raised in a religious family or environment) have been shown to partly explain the difference in religious attendance across 60 countries, including the Netherlands and the US (Ruiter & van Tubergen, 2009). Interestingly, it appears that these socioeconomic security factors may also partly explain the "exceptional pattern" of religiosity found in the US. The US have been reported to occupy an outlier position, as a highly modern yet highly religious society, with religion deeply ingrained in culture and social identity (Kelley & de Graaf, 1997; Tiryakian, 1991; Warner, 1993). Taking into account the importance of social security, Ruiter and van Tubergen (2009) argued that the US was no longer exceptional; the persistent strong socio-economic inequalities and strong religious history explain the high prevalence of religiosity in the US.

These country-level differences suggest that religion and belief in God may have an important function for providing a sense of control in people's lives in the US. In the Netherlands however, the social safety net may be so prevalent that it leaves far less room for religious beliefs to compensate for loss of personal control. In addition, based on the differential cultural prevalence of religion in the US and the Netherlands, we expected US participants to resort more easily to belief in a controlling God when lacking control - which is accepted as a socially desirable option in US culture. In contrast, in the Netherlands, strengthening one's belief in a controlling God as a consequence of control threat does not fit with general cultural expectations. Accordingly, we expected the effect of control threat on belief in a controlling God to be stronger in the US than in the Netherlands.

Notably, the original experiment comprised of a 2x2 design, with the emphasized

products. As they interpreted the preference for boundaries as a epistemic structuring tendency, they argued that highly religious individuals do not respond to control threats by compensating though choosing boundaries, as they already have a better alternative for restoring structure, i.e., belief in a controlling or structuring God. Analogously, atheists may use their trust in the government or a societal institution, rather than belief in God to buffer against the feeling of discomfort elicited by the control threat.

aspect of the nature of God as an additional between-subjects factor. That is, half of the participants in the study by Kay et al. (2008) rated their belief in the existence of God as a creator and half rated their belief in the existence of God as a controller. As predicted by CCT, personal control manipulations only affected belief in God when the controlling nature of God is emphasized – only then God serves as a compensation for a lack of personal control. Therefore, in the light of efficiency and relevance, in the present replication we chose to focus solely on the crucial effect with regard to belief in a controlling God.

The decision to omit the control condition with 'God as a creator' as a dependent variable comes at an informational cost. Admittedly, we cannot completely preclude the possibility that any effect of personal control threat causes increased belief in God due to some other characteristics of religious beliefs (e.g., the nature of God as loving, compassionate, all-knowing, or as a designer etc.) rather than belief in a controlling God per se. Interestingly though, later versions of CCT have also included more abstract epistemic structuring tendencies as compensatory strategies (e.g., an ordered, non-random version of evolution theory, stage theories of moral development and Alzheimer's disease, and aesthetically bounded vs. unbounded products; Cutright, 2011; Rutjens et al., 2010; Rutjens et al., 2013). Therefore, even when presented as a creator, belief in God may still serve as a compensatory strategy, by offering an epistemically structured conception of the world. Indeed, as mentioned in Landau et al. (2015), religious beliefs may present an especially well-suited opportunity for restoring feelings of control, exactly because they provide multiple means to this end. "For example, adhering zealously to religious beliefs may bolster external agency (through faith in beneficent intervention), affirm specific epistemic structure (by specifying consequences of moral conduct), and affirm nonspecific epistemic structure (portraying the universe as obeying a few well-observed and immutable laws)" (Landau et al., 2015).

This debate is beyond the scope of the present chapter, however. Instead we currently aimed to focus on investigating the primary effect that has been documented for CCT (i.e., control threat manipulations increase belief in a controlling God), which provides the strongest test of the theory. When evidence for this specific effect has been convincingly reported, this then paves the way for further research on the boundary conditions of the effect or potential extensions.

We thus aimed to conduct a direct replication of the crucial effect of the original Study 1 by Kay et al. (2008), including exactly the same manipulations and measures (excluding the extra control condition). At the same time we extended the original study in five ways, one related to the design, two related to the sampling (power and included population), and two related to the analysis (model and statistical framework). First, we included a measure of generalized feelings of personal control, allowing for a manipulation check and individual differences approach. Second, we increased the sample size ($n$ = 800 in total) to ensure sufficient power for detecting a small to medium effect. Third, we conducted the study in a relatively religious as well as a relatively secular country. Forth, we included religiosity as a potential moderator, rather than solely as a covariate. Fifth, we used a Bayesian hypothesis testing framework, allowing quantification of the evidence for or against the null hypothesis. Importantly, we note that only the first extension changed the experiment itself, yet in no way does it impede the validity of our direct replication attempt, as additional measures were included only *after* the original study had been conducted. Moreover, we believe the outlined extensions of the design allow us to better interpret any obtained results and provide a more sensitive test of the underlying theory.

## 4.3 Hypotheses

The predictions of the current replication attempt were straightforward: we expected that participants would "endorse the existence of [a controlling] God more strongly following the no-control memory task [compared to the control memory task]" (Kay et al., 2008, p. 22). More specifically, the main hypothesis, i.e., the replication hypothesis of primary interest, can be specified as follows:

$\mathcal{H}_{\textbf{exp}}$ : Primary experimental effect: recall of a positively-valenced situation in which one had *no personal control* (e.g., "Describe a pleasant event or situation over which you had absolutely no control") will result in more fervent belief in the existence of a controlling God, compared to recall of a positively-valenced situation in which one did have personal control (e.g., "Describe a pleasant event or situation over which you had total control").

Auxiliary hypotheses that were tested are:

$\mathcal{H}_{\textbf{cov}}$ : Covariate: levels of self-reported religiosity are positively related to belief in the existence of a controlling God.

$\mathcal{H}_{\textbf{man}}$ : Manipulation check: recall of a positively-valenced situation in which one had *no personal control* will result in lower levels of general feelings of personal control in one's life ("To what extent do you feel like you are the one who is in control in your life?"), compared to recall of a positively-valenced situation in which one did have personal control.

$\mathcal{H}_{\textbf{cor}}$ : Correlational effect: levels of general feelings of personal control in one's life are negatively related to belief in the existence of a controlling God.

$\mathcal{H}_{\textbf{cul}}$ : Cross-cultural effect: the primary experimental effect of personal control threat vs. affirmation on belief in a controlling God is moderated by cultural and socioeconomic factors reflected at the country-level; the experimental effect is stronger in the US than in the Netherlands.

The exact sequence of hypothesis testing, as well as the drawn inferences are depicted in Figure 4.2.

These hypotheses, as well as the planned analyses were agreed on by all involved parties and reviewers prior to the start of data collection. All materials, the full preregistered analysis plan, the anonymized raw and processed data, and the analysis scripts to conduct all confirmatory and exploratory analyses (including all figures) are available on the Open Science Framework (OSF; see https://osf.io/49xz3/).

## 4.4 Methods

### 4.4.1 Participants

The study was conducted in collaboration with the independent research agency Kieskompas (Amsterdam, the Netherlands; www.kieskompas.nl). Kieskompas specializes in online tools for assisting a general public in voting choices (e.g., for elections), but also offers panels for scientific research. They are affiliated with the Free University of Amsterdam, and have access to a (largely) representative sample of the 45 countries in which they operate.

Individuals older than 18 were eligible for participation. We specified no a priori exclusion criteria, which is in line with original study and converges with the meta-analysis by Landau et al. (2015) indicating no significant moderating effects of gender, college vs. non-college participants, form of compensation (credits vs. money), or region of data collection (US vs. outside of the US). However, we specified a criterion for the minimum time interval for completing the study. That is, we excluded participants who spent less than a particular number of minutes on the task (also known as "speeders") and whose data are therefore assumed to be invalid. As specified a priori, the criterion was set to 40% of the median of the total duration of the experiment, i.e., participants who spent less than 40% of the median time of the task, were excluded (Greszki et al., 2015). In practice, this resulted in a cutoff of 224 seconds for the Dutch version of the task and 159.4 seconds for the English version of the task. Additionally, as preregistered, we excluded the data from participants who wrote nonsensical stories in the recall task. This led to the exclusion of 34 participants in total; 22 and 9 participants were excluded for speeding in the Netherlands and the US, respectively, and 1 and 2 participants for writing nonsensical stories.[9]

After exclusions, the final samples consisted of 438 (51.6% female) participants in the Netherlands, and 391 (43.0% female) in the US. The average age of the Dutch participants was 58.4 ($SD$ = 15.3; range = $20 - 91$) and 50.2 ($SD$ = 16.1; range = $18 - 89$) for the American participants. We declare that all preregistered methodology was followed exactly unless explicitly stated otherwise.

### 4.4.2 Sampling Plan

Our sampling plan was based on Bayes Factor Design Analysis (BFDA; Schönbrodt and Wagenmakers, 2018; Stefan et al., 2019), a recently-developed method to help balance informativeness and efficiency of planned experiments within a Bayesian framework. We used the `BFDA` R package to compute the required sample size given the corrected effect size of the meta-analysis (i.e., $\delta = .379$; Schönbrodt, 2017). The analysis indicated that we would need 185 observations per group in order to obtain a Bayes factor in favor of $\mathcal{H}_{\text{exp}}$ larger than 10 with a probability of $p = 0.8$.[10] Fol-

[9] Note that we only excluded senseless stories, as preregistered. There were, however, also participants who wrote that they could not recall a situation that fit the particular characteristics that were requested. These participants were retained in the sample for the main analyses, but in the exploratory results section, we additionally report analyses excluding these participants.

[10] We chose the corrected effect size of the meta-analysis, rather than the effect size of the original study ($\delta = .769$) as this provides the most conservative estimate. We realize that BFDA is developed for planning designs in the context of a directed independent-groups $t$-test. Although we will use a one-way ANCOVA instead of a $t$-test, we believe BFDA can still provide a valuable indication of the

lowing this indication, we decided to aim for a final sample of 200 participants per group per country; $n = 400$ per experiment (see online supplementary materials for the distributions of expected Bayes factors generated based on the power analysis; https://osf.io/49xz3/).

### 4.4.3 Materials

Participants received all materials in their respective native language, i.e., English in the US and Dutch in the Netherlands.[11] Dutch materials were translated and back-translated by two different parties.

#### 4.4.3.1 Recall Task

As in the original study, participants were first presented with one of two memory tasks probing them to recall a recent positive event over which they did or did not have control. The Dutch and English items can be accessed on the OSF. The task instruction was taken from Kay et al. (2008) and read as follows: "Please try and think of something positive that happened to you in the past few months that you had [total / absolutely no] control over. Can you remember such a situation or event? Try to briefly describe this [un]controllable event in no more than 100 words. What happened and how did you feel?"

#### 4.4.3.2 Belief in a Controlling God

The dependent variable was equal to the one used in the original study; belief in the existence of God was assessed based on two items:

1. To what extent do you think it is feasible that God, or some type of nonhuman entity, is in control, at least in part, of the events within our universe?
2. To what extent do you think that the events that occur in this world unfold according to God's, or some type of nonhuman entity's, plan?

Following the original study, the items were evaluated on a 7-point Likert scale with descriptive labels at the extremes, hence ranging from *tremendously doubtful* to *very likely*. Ratings for the two items were averaged to reflect the level of belief in a controlling God.

#### 4.4.3.3 Manipulation Check Items

In order to mask the dependent variable and reduce the chances of participants readily discovering the purpose of the study, the items on belief in God were immediately followed by six general questions and four questions on the situation described by the participants in the recall task. The general questions included a manipulation check

desired sample size. That is, since our analysis will also focus on a directed hypothesis comparing two independent groups, we consider BFDA more suitable for the current study than a traditional power analysis.

[11] Although this inevitably creates a language confound – as in any cross-national study – we believe the use of the different, i.e., native languages has higher ecological validity. Moreover, conducting the study in participants' second language induces a probably even larger confound (i.e., it will be more difficult to describe a situation in their second rather than their first language).

on general feelings of control in one's life; the items on the recalled situation served as a check on instruction compliance (i.e., the situation in the control affirmation condition did indeed involve high levels of personal control; the situation in the control threat condition involved low levels of control) and as a reinforcement of the idea that the study supposedly investigated memory. Similar to the items on belief in God, all questions were evaluated on a 7-point Likert scale with descriptive labels at the extremes. The crucial manipulation check items assessing general feelings of control in one's life were:

1. To what extent do you feel like you are the one who is in control of your life?
2. To what extent do you consider yourself the actor in, or the director of, your life?

The ratings on these two items were averaged to reflect general feelings of personal control. The additional personality questions assessed self-esteem (1 item) and mood (2 items; following Kay et al., 2008), and extroversion (1 item). The four items on the recalled situation assessed perceived control, affect, vividness, and significance.

Although inclusion of these additional questions and manipulation check items deviates from the original study, we believe that it did not meaningfully change the crucial effect of the experimental control manipulation on belief in a controlling God. Specifically, because all added questions were presented *after* measurement of the dependent variable, the main study remained a direct replication of the experiment by Kay et al. (2008). Moreover, we believe this deviation was justifiable as it reduces the probability of participants correctly identifying the tested hypothesis, a risk we considered fairly high.

#### 4.4.3.4 Religiosity and Demographics

Finally, participants' age, gender, and level of religiosity were assessed at the end of the experiment. Level of self-reported religiosity was expected to be highly correlated with the dependent variable and was included in the analysis. Again, a 7-point Likert scale was used to measure religiosity ("How religious do you consider yourself?"), ranging from *not at all religious* to *extremely religious*.

### 4.4.4 Procedure

Although the original study administered materials on paper, the current replication used a computerized version presented using the survey software Qualtrics. We believe this adjustment of the original experiment was reasonable in light of the advantages of an online experiment in terms of efficiency and potential for recruiting a large and more representative sample, as well as the fact that we saw no reason to assume that the application of an online version might change the experiment in any meaningful way. Importantly, in the meta-analysis by Landau et al. (2015), 36% of the studies were conducted online. The authors found no effect of method of presentation (called 'region' in the article), corroborating research demonstrating cross-method consistency between lab and online studies in various social-psychological domains (e.g., Buchanan & Smith, 1999; Gosling et al., 2004; Robins et al., 2002; Srivastava et al., 2003).

The experiment was conducted in the order as presented under *Materials*. That is, after a short introduction, participants were presented with the recall task for which they were randomly assigned to either the control affirmation condition or the control threat condition. Subsequently, participants rated their belief in the existence of a controlling God (2 items) and filled out the additional questions on general personality and on the recalled situation (10 items), including the two manipulation check items. Finally, participants provided demographics, including the religiosity item, and completed an awareness check.[12]

### 4.4.5 Data Analysis

All analyses were conducted in a Bayesian framework. The `BayesFactor` R package was used to calculate Bayes factors in order to quantify the evidence for or against the main experimental and the covariate hypothesis (Morey & Rouder, 2015). Specifically, we used the `lmBF` function which allows for the inclusion of categorical (i.e., control condition) and continuous (i.e., religiosity) predictors. Moreover, the statistics software JASP (JASP Team, 2019) was used to calculate the Bayes factors for the manipulation check hypothesis (i.e., a directed independent samples t-test) and the correlational hypothesis (i.e., a Kendall's tau negative correlation test). The full R code as well as the JASP files are published on the OSF (https://osf.io/49xz3/). The online supplement additionally contains the detailed description of all anticipated analysis paths as preregistered, plus the application of these analyses on a simulated data set. As the description included potential outcomes that were not observed and analysis steps that were therefore irrelevant, in the main text we confine ourselves to the relevant analysis paths (see Figure 4.2). We declare that the proposed confirmatory analyses were followed exactly.

#### 4.4.5.1 Prior Specification

A default Jeffreys-Zellner-Siow (JZS) prior for ANOVA / general linear models was used, with an r-scale of fixed effects of 0.5 (for the control condition variable), and r-scale of covariates of .354 (for religiosity; Rouder et al., 2012; Wetzels et al., 2012). For the Kendall's tau correlation, the default uniform prior proposed by (Jeffreys, 1961) was used (van Doorn et al., 2018). That is, a stretched beta prior with width 1.

#### 4.4.5.2 Calculation of Bayes Factor

For all our specified hypotheses, we expected a directed effect, i.e., a one-sided test. Therefore, Bayes factors $\text{BF}_{+0}$ or $\text{BF}_{-0}$ were calculated in order to evaluate the extent to which the data were likely under the alternative hypothesis $\mathcal{H}_+$ or $\mathcal{H}-$ versus the null hypothesis $\mathcal{H}_0$. Note that the subscripts on Bayes factor refer to the hypotheses being compared, with the first and second subscripts referring to the one-sided hypothesis of interest and the null hypothesis, respectively. $\text{BF}_{+0}$ is used in case of a hypothesized positive effect for the reference group or a positive relation between

[12] In our preregistration, we specified that we would investigate whether there was a difference between participants who correctly identified the relation of interest vs. participants who did not. However, analysis of the awareness check (i.e., "What do you think this research was about?") indicated that only 1 person in the Dutch sample and 2 people in the US sample correctly derived that the study investigated whether people tend to more strongly believe in God after recalling a situation in which they did not have control. Therefore, we decided not to run separate analyses.

**Table 4.1:** Descriptive Statistics of Belief in a Controlling God by Country and Control Condition.

| Country | Condition | *n* | Mean | Median | SD |
|---|---|---|---|---|---|
| Netherlands | Control Affirmation | 214 | 2.76 | 2.00 | 1.97 |
| Netherlands | Control Threat | 224 | 2.55 | 1.50 | 1.94 |
| United States | Control Affirmation | 197 | 3.20 | 2.50 | 2.20 |
| United States | Control Threat | 194 | 3.37 | 2.75 | 2.41 |

*Note.* Belief in a controlling God as measured on a 7-point Likert scale and averaged over the two items.

variables; $\text{BF}_{-0}$ is used for a negative effect for the reference group or a negative relation between variables.

The Bayes factor reflects the change from prior model probabilities to posterior model probabilities and as such quantifies the evidence that the data provide for $\mathcal{H}_+$ versus $\mathcal{H}_0$. For the experimental effect, this can be specified as $\mathcal{M}_{exp}$ versus $\mathcal{M}_{cov}$, reflected by:

$$\underbrace{\frac{p(\mathcal{M}_{exp} \mid \text{data})}{p(\mathcal{M}_{cov} \mid \text{data})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{M}_{exp})}{p(\mathcal{M}_{cov})}}_{\text{prior odds}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{M}_{exp})}{p(\text{data} \mid \mathcal{M}_{cov})}}_{\text{Bayes factor}} \tag{4.1}$$

Indeed, the Bayes factor $\text{BF}_{+0}$ then represents the ratio of the marginal likelihoods of the observed data under $\mathcal{M}_{exp}$ and $\mathcal{M}_{cov}$:

$$\text{BF}_{10} = \frac{p(\text{data} \mid \mathcal{M}_{exp})}{p(\text{data} \mid \mathcal{M}_{cov})} \tag{4.2}$$

By default, prior model odds were assumed to be equal for both models. As the evidence is quantified on a continuous scale, we also present the results as such. Nevertheless, we included a verbal summary of the results by means of the interpretation categories for Bayes factors proposed by Lee and Wagenmakers (2013, p.105), based on the original labels specified by Jeffreys (1939).

## 4.5 Results – Preregistered

For the confirmatory analyses, we followed the analysis pipeline as specified in the preregistration. Figure 4.2 represents the pipeline and highlights the route and subsequent conclusions that the results indicated. Below, the results of the individual analysis steps are outlined.

### 4.5.1 Experimental Effect

In the analysis of the original study, a two-way univariate ANOVA was conducted, including the factors *control* (threat vs. affirmation), *nature of God* (controlling vs. creating) and religiosity as a covariate (i.e., an ANCOVA). However, since the replication focused solely on the crucial control threat effect on belief in a controlling

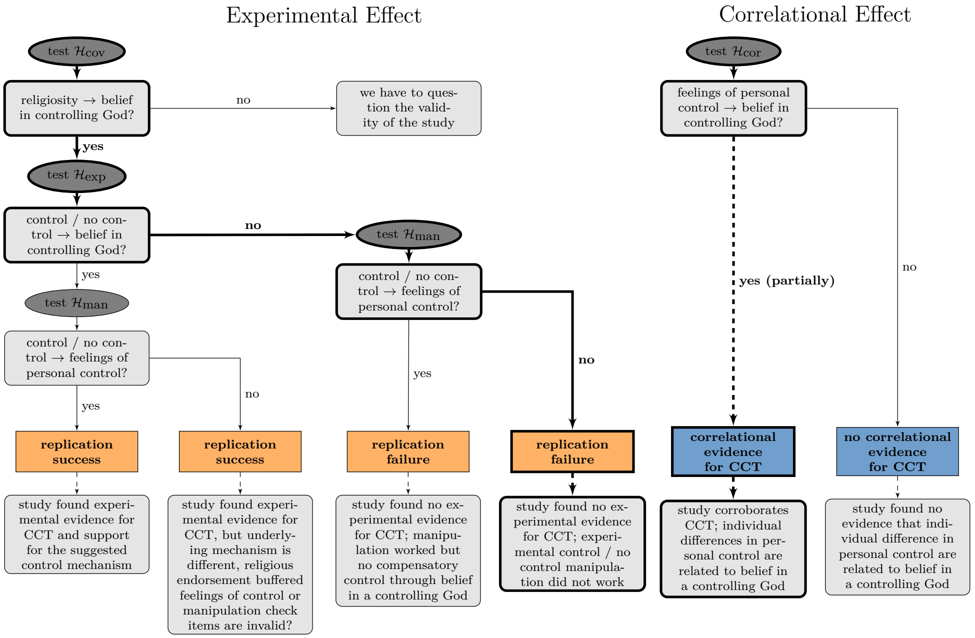

Experimental Effect
Correlational Effect
test $\mathcal{H}_{\text{cov}}$
religiosity → belief in controlling God?
no
we have to question the validity of the study
yes
test $\mathcal{H}_{\text{exp}}$
control / no control → belief in controlling God?
no
test $\mathcal{H}_{\text{man}}$
yes
test $\mathcal{H}_{\text{man}}$
control / no control → feelings of personal control?
control / no control → feelings of personal control?
yes
no
yes
no
replication success
replication success
replication failure
replication failure
study found experimental evidence for CCT and support for the suggested control mechanism
study found experimental evidence for CCT, but underlying mechanism is different, religious endorsement buffered feelings of control or manipulation check items are invalid?
study found no experimental evidence for CCT; manipulation worked but no compensatory control through belief in a controlling God
study found no experimental evidence for CCT; experimental control / no control manipulation did not work
test $\mathcal{H}_{\text{cor}}$
feelings of personal control → belief in controlling God?
yes (partially)
no
correlational evidence for CCT
no correlational evidence for CCT
study corroborates CCT; individual differences in personal control are related to belief in a controlling God
study found no evidence that individual difference in personal control are related to belief in a controlling God

**Figure 4.2:** The preregistered analysis pipeline, displaying tested hypotheses and interpretation of possible results for both the experimental and correlational approach of CCT for religious beliefs. The results of the study suggested to follow the two paths indicated with the thick lines.

God, we preregistered and conducted a one-way ANCOVA instead. Specifically, we calculated the Bayes factor for the hypothesis that the personal control threat induced a higher rating for belief in a controlling God, compared to the personal control affirmation ($\mathrm{BF}_{+0}$), in addition to the effect of religiosity. The descriptive statistics for the experimental hypothesis are given in Table 4.1 and the data are plotted in Figure 4.3.

#### 4.5.1.1 Outcome Neutral Criterion

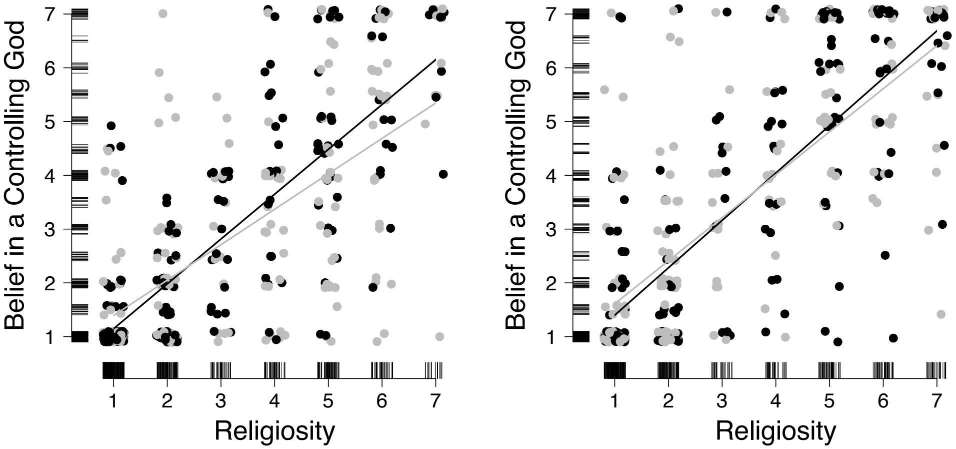
First, we tested the covariate hypothesis to assess whether the outcome neutral criterion was met. That is, we compared the null model ($\mathcal{M}_0$) to the model including religiosity (covariate; $\mathcal{M}_{cov}$) to validate the positive relation between religiosity and belief in a controlling God. Results revealed a Bayes factor of $2.20 \times 10^{71}$ in favor of $\mathcal{M}_{cov}$ relative to $\mathcal{M}_0$; indicating that – given the data – a positive correlation between religiosity and belief in a controlling God is about $2.20 \times 10^{71}$ times more likely than no relation. In the US, a similar relation was observed; here we found a Bayes factor of $8.39 \times 10^{74}$ in favor of $\mathcal{M}_{cov}$ relative to $\mathcal{M}_0$. In order words, for both countries, the data provide overwhelming evidence for the covariate hypothesis.

#### 4.5.1.2 Experimental Effect

In order to quantify the evidence for the control threat effect on belief in a controlling God, we compared the model including only religiosity ($\mathcal{M}_{cov}$) to the model including religiosity and control condition ($\mathcal{M}_{exp}$). In the Netherlands, we found a Bayes factor of 0.18 in favor of $\mathcal{M}_{exp}$ over $\mathcal{M}_{cov}$; $\mathrm{BF}_{+0} = 0.18$ (i.e., the evidence for the null hypothesis was: $\mathrm{BF}_{0+} = 5.41$). This means that the data are about 5.41 times more likely under the null model including only religiosity, compared to the alternative model that also includes the control-threat manipulation. This is constitutes moderate evidence against an effect of control threat on belief in a controlling God. In the US, a similar pattern was observed; $\mathrm{BF}_{+0} = 0.09$ (i.e., $\mathrm{BF}_{0+} = 11.16$) indicates strong evidence for the null hypothesis over the experimental hypothesis that the control threat manipulation resulted in heightened belief in a controlling God. Following the analysis plan, these findings are taken as “replication failure” for the experimental control threat effect on belief in a controlling God. The raw data for both countries are displayed in Figure 4.3 (posterior distributions of the model parameters are plotted in the online supplementary materials).

#### 4.5.1.3 Interaction Effect

Although we did not find a main effect of control condition on belief in a controlling God, there may have been an interaction between religiosity and control condition, e.g., the control-threat effect could be present only for those who are already strongly religious. In order to investigate this possibility, we compared $\mathcal{M}_{exp}$ to the model including religiosity, control condition, and the interaction between religiosity and control condition ($\mathcal{M}_{full}$). This yielded no evidence for an interaction effect: $\mathrm{BF}_{10} = 1.72$ for $\mathcal{M}_{full}$ relative to $\mathcal{M}_{exp}$ in the Netherlands. In the US, we found a $\mathrm{BF}_{10} = 0.10$ for the $\mathcal{M}_{full}$ relative to $\mathcal{M}_{exp}$ (i.e., $\mathrm{BF}_{01} = 10.16$), indicating strong evidence in favor of the no-interaction hypothesis.



**(a)** The Netherlands

**(b)** The United States

**Figure 4.3:** Scatter plots with the relation between religiosity and belief in a controlling God (a) in the Netherlands and (b) in the United States. Light dots represent individuals in the control threat condition and dark dots individuals in the control affirmation condition. Note that the data points are jittered to enhance visibility of overlapping observations.

**Table 4.2:** Posterior Model Probabilities.

| | Netherlands | United States |
|---|---|---|
| Religiosity Only | 0.744 | 0.889 |
| Religiosity + Control | 0.094 | 0.101 |
| Religiosity + Control + Religiosity*Control | 0.162 | 0.010 |

*Note.* All three models were assumed to be equally likely a priori.

#### 4.5.1.4 Posterior Model Probabilities

Assuming equal prior probabilities for all three models and using Bayes' rule (see equation [1]), the posterior model probabilities are 0.744 and 0.889 for $\mathcal{M}_{cov}$, 0.094 and 0.101 for $\mathcal{M}_{exp}$, and 0.162 and 0.010 for $\mathcal{M}_{full}$, for the Netherlands and the US, respectively (see Table 4.2). These results demonstrate again that the religiosity-only model predicted the observed data better than the control threat model and the full model.

### 4.5.2 Manipulation Check

In order to assess the effectiveness of the experimental manipulation, we tested whether the personal control threat condition indeed elicited lower general feelings of personal control, relative to the personal control affirmation. In the Dutch sample, we found no evidence that the control threat manipulation lowers feelings of general control, relative to the affirmation condition, indicating that the manipulation was not successful. The effect size was $\delta = 0.008$, 95% CI [-0.176, 0.193], $\text{BF}_{-0} = 0.11$ (i.e., $\text{BF}_{0-} = 8.79$), which qualifies as moderate evidence for the null hypothesis.[13] Similarly, in the American sample, there was no evidence for the effectiveness of the manipulation: $\delta = 0.081$, 95% CI [-0.116, 0.277], $\text{BF}_{-0} = 0.25$ (i.e., $\text{BF}_{0-} = 4.06$), which qualifies as moderate evidence for the null hypothesis. As specified in the analysis plan, these results indicate that the manipulation was unsuccessful.

### 4.5.3 Correlational Effect

In addition to the experimental hypothesis, we assessed the relationship between feelings of personal control and belief in a controlling God. As we expected a monotonic, but not necessarily linear relation, a one-sided (negative) Bayesian Kendall's tau correlation test was used. In the Netherlands, we found $\tau = -0.010$, 95% CI [-0.074, 0.052], $\text{BF}_{-0} = 0.08$ (i.e., $\text{BF}_{0-} = 12.24$). This qualifies as strong evidence for the null hypothesis. In the US, on the other hand, we found $\tau = -0.144$, 95% CI [-0.210, -0.078], $\text{BF}_{-0} = 1185$. This qualifies as extreme evidence for the presence of an inverse relation between general feelings of personal control and belief in a controlling God.

### 4.5.4 Cross-Cultural Effect

The results of the cross-cultural analysis with combined data from the Dutch and American sample corroborate the findings from the separate analyses; we find $\text{BF}_{10} = 0.08$ (i.e., $\text{BF}_{01} = 12.02$), indicating that the data are 12.02 times more likely under the Religiosity + Country model compared to the Religiosity + Country + Control-threat Condition model. This indicates strong evidence for the null hypothesis that the control threat manipulation did not have an effect on belief in a controlling God. As seen in Table 4.3, adding an interaction between country and condition also did not increase the posterior model probability. The model including only Religiosity and Country outperforms the alternative models. Note that the added predictive

[13] Note that the parameter estimation for the effect size and the confidence interval are based on the unrestricted model, whereas the Bayes factor is derived from the order-restricted model. This applies to all directed tests.

**Table 4.3:** Posterior Model Probabilities For The Cross-Cultural Effect.

| Model | Posterior Probability |
|---|---|
| Religiosity + Control | 0.000 |
| Religiosity + Country | 0.912 |
| Religiosity + Control + Country | 0.077 |
| Religiosity + Control + Country + Control*Country | 0.010 |
| Religiosity + Control + Country + Control*Country*Religiosity | 0.001 |
| Religiosity Only | 0.000 |

*Note.* All six models were assumed to be equally likely a priori.

adequacy of the Country parameter reflects the main effect of country, i.e., belief in a controlling God is higher in the US compared to the Netherlands.

### 4.5.5 Additional Analyses

#### 4.5.5.1 Positive Controls

The relationship between belief in a controlling God and gender was included as a ‘positive control test’ to establish the validity of the dependent variable. The relation between gender and religiosity appears one of the most robust effects with regard to religious beliefs; women consistently report being more religious than men (Bradshaw & Ellison, 2009; Collett & Lizardo, 2009; L. J. Francis, 1997; Miller & Hoffmann, 1995; Roth & Kroll, 2007). Indeed, in both samples we found evidence for the hypothesis that women more strongly believe in a controlling God than men: $\text{BF}_{+0} = 6.07$ (i.e., moderate evidence) in the Netherlands and $\text{BF}_{+0} = 24.42$ (i.e., strong evidence) in the US.

#### 4.5.5.2 Control for Effort

In order to investigate whether there were any differences in the amount of time or number of words participants spent on writing for the experimental manipulation, we conducted two Bayesian default two-sided *t*-tests. The amount of time spent on the memory recall item did not differ between conditions in the Netherlands: $\text{BF}_{10} = 0.19$ (i.e., $\text{BF}_{01} = 5.40$; moderate evidence for the null hypothesis), or in the US: $\text{BF}_{10} = 0.17$ (i.e., $\text{BF}_{01} = 6.02$; moderate evidence for the null hypothesis). Furthermore, the number of words used to describe the memory likewise did not differ between conditions in the Netherlands: $\text{BF}_{10} = 0.12$ (i.e., $\text{BF}_{01} = 8.69$; moderate evidence for the null hypothesis), or in the US: $\text{BF}_{10} = 0.13$ (i.e., $\text{BF}_{01} = 7.53$; moderate evidence for the null hypothesis). See the online supplementary results for descriptives.

## 4.6 Results – Exploratory

### 4.6.1 Instruction Compliance

The data showed that the memory recall manipulation did not substantially affect generalized feelings of personal control. Accordingly, it could be that the manipulation was either insufficiently strong to change feelings of control, or that participants simply did not understand or comply with the instructions to report a personal memory in which they did or did not have control over a situation. To explore this issue, we investigated the item in which participants indicated how much control they had experienced in the described situation. Here, we found extreme evidence for the hypothesis that experienced control was higher in the control affirmation ($M = 5.62$; $M = 5.89$) compared to the control threat condition ($M = 2.85$; $M = 2.32$) in both countries: $\text{BF}_{10} = 8.4 \times 10^{47}$ and $\text{BF}_{10} = 1.1 \times 10^{71}$, in the Netherlands and the US, respectively.

Finally, although in both conditions the valence of the described situation was above the midpoint of the scale (i.e., participants rated the situation as positive), we did observe a difference between conditions; the control affirmation situation was experienced as more pleasant ($M = 6.06$; $M = 6.15$) than the control threat condition ($M = 5.41$; $M = 5.66$). In the Dutch sample, there was extreme evidence for a difference in valence: $\text{BF}_{10} = 568.4$, in the US, the evidence was very strong: $\text{BF}_{10} = 23.58$.

### 4.6.2 Experimental Effect Excluding Unsuccessful Recalls

In the analyses reported above, we followed our preregistration by excluding only those participants who wrote nonsensical stories in the recall task. Whereas nonsensical descriptions were rare ($n = 3$ in total), there were a number of participants who indicated that they could not recall a situation that met the requested characteristics, i.e., being recent and positive and over which they had total control / no control. There were 55 and 15 individuals in the Dutch and in the US sample, respectively, who indicated not being able to access an episode as specified.

We re-ran the models including only the participants who succeeded to recall a specific event, in order to investigate whether the experimental effect would be present in this sub-sample. Again, we collected moderate evidence against the experimental control hypothesis in the Netherlands: $\text{BF}_{+0} = 0.26$, i.e., $\text{BF}_{0+} = 3.89$. Similarly, in the US, the evidence pointed against the experimental effect: $\text{BF}_{+0} = 0.13$, i.e., $\text{BF}_{0+} = 7.97$. In other words, the results as reported for the confirmatory analysis did not change when we additionally excluded participants who attempted but could not describe a situation in line with the experimental control manipulation.

## 4.7 Discussion

In the current replication study, we revisited the initial finding suggesting a causal effect of the loss of experienced personal control in one's life on belief in a controlling God (Kay et al., 2008). Our results indicate moderate to strong evidence for the absence of this effect: belief in (a controlling) God is not modulated by a threat compared to an affirmation of personal control. Using large samples ($N \approx 400$) we did not replicate the original experiment by Kay et al. (2008) in the Netherlands, nor in the US. In a complementary analysis, we assessed the correlational relationship between feelings of personal control and belief in a controlling God. In the Dutch sample no relationship was found. In the American sample, people who experienced lower levels of personal control in their lives, reported a stronger endorsement of belief in a controlling God - although the effect size of this relationship was small.

The data also showed no effect of the personal control manipulation on feelings of personal control in one's life. This manipulation failure is remarkable for several reasons: (1) affecting feelings of personal control is the very purpose of the experimental manipulation (Kay et al., 2008); (2) an effect of the manipulation on generalized personal control has been validated in a separate pilot study reported by Kay et al. (2008); (3) feelings of personal control are the core construct of CCT (Kay et al., 2009; Landau et al., 2015), and (4) these manipulation check items have been successfully used in previous studies (e.g., Cutright, 2011; Goode et al., 2014; Rutjens et al., 2010; Rutjens et al., 2013). It should be noted that Kay et al. (2008) verified the effect of the control threat manipulation on general feelings of personal control in an independent study, rather than adding the manipulation check to the main study. The reason for separating the two effects was that the intervening opportunity to affirm control – via endorsing the existence of a controlling God – should eliminate any effect of the control manipulation on the manipulation check; people who have already restored their sense of structure or order will not report a residual lack of control. Nevertheless, in our study, the control threat manipulation did not influence belief in a controlling God. Therefore, if participants' feelings of control were threatened, the lack of control was not yet buffered and should have been reflected in the manipulation check items.

The lack of an experimental effect may be related to the framing of the autobiographical recall task and/or to the potential inefficacy of experimental control threat manipulations. First, it could be that the specific instruction to recall a recent *positive* memory might be related to the absence of an effect. A positive situation is typically not experienced as *threatening*; prototypical examples of positive situations in which one lacks personal control are the experience of "luck", "happy coincidence", or "fate". Quite a few participants in our study had a hard time recalling a recent positive situation in which they had or lacked control; 70 individuals (i.e., 8.4% of the total sample) reported not being able to recall such a situation. Many of the situations that participants reported would not qualify as 'threatening'. Some of the situations that our manipulation elicited are exemplified by a collection of responses. These were randomly drawn from both the control affirmation and the control threat condition and are displayed in Table 4.4.

Our manipulation was similar to the original study by Kay et al. (2008). The rationale for asking participants to recall a positive situation was to control for the possible confound that any effect might be simply related to the valence of the memory. Research on divine responsibility additionally alludes to the notion that positive episodes may be associated with God. Early studies already suggested that people often tend to make supernatural attributions, also in the case of positive experiences (Gorsuch & Smith, 1983; Ritzema & Young, 1983). For example, Gorsuch and Smith (1983) found that positive outcomes of good fortune were frequently regarded as acts from God's hand. Similarly, Norenzayan and Lee (2010) found that scenarios about winning a lottery or meeting the love of one's life were often attributed to fate, and mostly so for religious individuals, suggesting these individuals inferred divine responsibility to be at play. Following this line of argumentation, it could well be that uncontrollable positive situations as induced in the present autobiographical recall task foster belief in a controlling God as a compensatory source of control and as a

**Table 4.4:** Examples of Recalled Situations in the Control Affirmation and Control Threat Condition as Reported by Participants.

| Condition | Description |
|---|---|
| Control Affirmation | "I completed my first 5K run. While I did not place first, I was nonetheless pleased with my performance. The run itself was exciting and I felt a sense of satisfaction when I was done." |
| Control Affirmation | "I lost 15 lbs by cutting sugar from my diet and controlling my eating for 30 days." |
| Control Affirmation | "I entered several pieces of art into a juried show; they were accepted. And while I didn't have total control over their acceptance, I did over the production of the art pieces. Which is enough." |
| Control Threat | "Insurance Visa card payed up for an item not received or ordered." |
| Control Threat | "Potential for promotion/title change with change in admin and bureaucracy. Happy, to an extent, but not totally consistent with my future goals." |
| Control Threat | "My wife and I went out for dinner with a neighbor. The neighbor paid the bill without us noticing. It was a very thoughtful gesture and I felt appreciated." |

*Note.* These examples are randomly drawn from all responses in the autobiographical recall task (excluding unsuccessful recalls) in the US sample, since these were written in English and did not require translation.

causal agent (indirectly) explaining the occurrence of these uncontrollable events.

At the same time, however, the literature indicates that divine attributions tend to occur more frequently for extraordinary and improbable events that lack alternative explanations (Gorsuch & Smith, 1983; Ritzema & Young, 1983), whereas participants in the current study mostly reported mundane events. According to CCT, people have a fundamental drive to obviate the experience of randomness in the world (Kay et al., 2009). Compensatory strategies such as endorsing belief in an intervening God are triggered when personal control is low, in order to satisfy the basic need to maintain a sense of non-randomness. This assumes that the lack of personal control is experienced as an aversive state. However, the uncontrollable yet positive and mundane situations described by participants in our study likely did not sufficiently activate the need to restore a sense of control through compensatory efforts.

A second reason for our replication failure could be related to the possibility that an experimental recall manipulation may be ineffective in instilling a sufficiently powerful sense of (un)controllability. Autobiographical recall tasks have been used extensively in research on mood induction (e.g., Strack et al., 1985). Although many studies provide supportive evidence for the efficacy of autobiographical recall in inducing basic emotions and mood (e.g., Jallais & Gilet, 2010; Siedlecka & Denson, 2019), other studies have failed to find these effects (Göritz & Moser, 2006). Recalling a particular episode in a lab or behind a computer is probably too subtle to produce an experience that is comparable to that in the original situation and hence may fail to exert causal impact on any outcome of interest (see for instance Schjoedt, 2009 for a similar argument in the context of religious and mystical experiences). This may be a particular concern for manipulations aiming to induce a relatively complex cognitive state (e.g., experience of power or control), rather than an arguably stronger emotional state.

Our findings cohere with those of van Elk and Lodder (2018); across seven experiments they found no support for the effectiveness of various personal control manipulations, including the autobiographical recall task used in the present study. Our suggestion that autobiographical recall manipulations may be ineffective echoes recent discussions in the priming literature, where the effectiveness of behavioral priming generally (Cesario, 2014; Doyen et al., 2012; Pashler et al., 2013; Shanks et al., 2013; Stroebe & Strack, 2014), and religious priming specifically (Gomes & McCullough, 2015; van Elk et al., 2015; van Elk et al., 2016) was called into question. Some contested effects also included autobiographical recall manipulations, for instance with respect to experimental effects of feelings of power (e.g., Galinsky et al., 2008) and morality (Fayard et al., 2009; Zhong & Liljenquist, 2006).

In response to these replication failures, Lammers et al. (2017) argued that ease of retrieval can moderate the effectiveness of recall manipulations of cognitive constructs such as power and control. The authors showed that recall manipulations are ineffective or even counter-effective when the instructed situation or experience is highly inaccessible. Although we did not directly address this possibility, we consider this explanation of our null results unsatisfactory for two interrelated reasons: First, the exploratory analysis which only included participants who managed to successfully recall a situation likewise provided evidence in favor of the null hypothesis. Second, when we added the interaction between time spent on the recall task (as a proxy for ease of retrieval) and control condition to the model, this resulted in very strong evidence against the moderation model: $\mathrm{BF}_{10} = 0.008$ and $\mathrm{BF}_{10} = 0.030$ for the Netherlands and the US, respectively.

We found supportive evidence for a correlational effect consistent with predictions derived from CCT, namely: in the US sample overall feelings of control were related to belief in a controlling God. This finding is in line with previous observations by van Elk and Lodder (2018), who exploratorily found that general subjective feelings of control were associated with different dependent variables related to epistemic structuring tendencies across four of the seven experiments. This again suggests that it is difficult to manipulate feelings of control experimentally, but that relatively stable individual differences in the experience of control are associated with compensatory strategies in a way that is compatible with CCT.

In our study, the correlation between feeling of control and belief in a controlling God was only found in the US and not in the Netherlands. This cross-national difference may be related to country-level differences in the cultural prevalence of religiosity and existential security (Barber, 2011). Religion is deeply rooted in US cultural identity; Christianity presently continues to shape American lives and guide politics (Wald & Calhoun-Brown, 2014). As such, the notion of a controlling God may –unconsciously– be seen as an especially appealing and comforting belief – especially for individuals who experience little personal control. At the same time, strong faith in a controlling God logically implies reduced personal control - as exemplified for instance in the Protestant notion of ‘Predestination’ (M. Weber, 1930). These mechanisms may be mutually reinforcing, together contributing to the negative relation between perceived personal control and belief in divine control as found for the US sample.

In the Netherlands, in contrast, the role of religion in socialization and education has rapidly declined over the last 50 years, curtailing religion's pervasiveness in society (Kregting et al., 2018). Combined with the relatively strong welfare system in the Netherlands, the marginal role of religion makes God a far less likely source for offering a sense of order and control in the world than in the US (Norenzayan & Gervais, 2013). It may well be that in the Netherlands, faith in the government or science constitutes a stronger source for offering compensatory control.

In conclusion, one important general lesson from this work is that caution is warranted in generalizing the effectiveness of experimental manipulations of control across samples and contexts (e.g., Cesario, 2014). Psychological researchers should be sensitive to and explicit about contextual boundaries of the phenomena of interest. In the current study, we anticipated that the cultural religious context would be a boundary condition for the compensatory control effect with respect to religious beliefs. Indeed, we showed that cultural setting affected the relation between feelings of control and belief in God – but only when using an individual differences approach. For the experimental effect, the cultural background appeared to be irrelevant as the manipulation was ineffective across the board; we did not find the experimental effect in a secular country (i.e., the Netherlands), nor in a highly religious country (i.e., the US).

It seems plausible that in periods and places characterized by little personal control some people are drawn to religion to reduce uncertainty and unpredictability in their lives; churches and temples may thrive during times of war or natural disaster, but it remains difficult to investigate this theory by means of experimental and autobiographical priming manipulations.

# 5

# Religious Belief and Cognitive Conflict Sensitivity: A Preregistered fMRI Study

In the current preregistered fMRI study, we investigated the relationship between religiosity and behavioral and neural mechanisms of conflict processing, as a conceptual replication of the study by Inzlicht et al. (2009). Participants ($N = 193$) performed a gender-Stroop task and afterwards completed standardized measures to assess their religiosity. As expected, the task induced cognitive conflict at the behavioral level, and at a neural level this was reflected in increased activity in the anterior cingulate cortex (ACC). However, individual differences in religiosity were not related to performance on the Stroop task as measured in accuracy and interference effects, nor to neural markers of response conflict (correct responses vs. errors) or informational conflict (congruent vs. incongruent stimuli). Overall, we obtained moderate to strong evidence in favor of the null hypotheses that religiosity is unrelated to cognitive conflict sensitivity. We discuss the implications for the neuroscience of religion and emphasize the importance of designing studies that more directly implicate religious concepts and behaviors in an ecologically valid manner.

## 5.1 Introduction

Everywhere across the world, in all times and cultures we find people who believe in supernatural beings. Religious beliefs seem highly successful in offering explanations for various phenomena, ranging from how the world originated, to why one had to switch jobs and what happens after one dies. Yet these beliefs are difficult - if not impossible - to support with empirical evidence. In fact, believers are often confronted with widely supported contradicting evidence, for instance evolutionary explanations of the origins of life or reductionistic explanations of their religious experiences. And

This chapter has been adapted from: Hoogeveen, S., Snoek, L., & van Elk, M. (2020). Religious belief and cognitive conflict sensitivity: A preregistered fMRI study. *Cortex*, *129*, 247–265. https://doi.org/10.1016/j.cortex.2020.04.011.

yet, despite these challenges, most religious believers keep up their faith (Pew Research Center, 2012).

Various scholars have suggested that a mechanism of reduced conflict sensitivity, i.e., detecting the incongruency between two potentially conflicting sources of information, may foster the acceptance and maintenance of religious beliefs. For example, dual-process accounts of religion (Risen, 2016), the predictive processing model (van Elk & Aleman, 2017), and the cognitive resource depletion model (Schjoedt et al., 2013) all assume that religiosity is associated with a reduced tendency for analytical thinking and error monitoring.

Where the dual-process model by Risen (2016) assumes a conflict between intuitive and analytical thinking that is resolved by acquiescing to the intuition, the predictive processing model by van Elk and Aleman (2017) assumes a conflict between prior beliefs and sensory input that is resolved by assigning more weight to priors and suppressing the influence of error signals (and hence mitigating the update of prior beliefs). The cognitive resource depletion model applies the notion of reduced error monitoring specifically to collective religious rituals. According to the model, the combination of a charismatic authority, a high arousal context, and a sequence of causally opaque ritualized behaviors creates optimal circumstances to facilitate a preordained (religious) interpretation of events and reduces the likelihood for idiosyncratic (potentially non-religious) interpretations. These subtle differences seem to predominantly reflect 'a tale of different literatures', possibly due to the fact that the frameworks originate from different disciplines; dual-process models were developed in social psychology, predictive processing in (cognitive) neuroscience, and the cognitive resource depletion model stems from anthropological research. Nevertheless, all three accounts converge on the key idea that a process of reduced conflict detection (or correction) makes individuals less prone to note information that seemingly contradicts their religious worldviews and to update their beliefs in the light of new information. This mechanism could potentially underlie the relative immunity of religious beliefs to criticism based on empirical observations (cf. what Van Leeuwen, 2014 calls 'evidential invulnerability').

Notably, the implicit assumption of most theoretical frameworks appears to be that a mechanism of reduced conflict sensitivity makes people more receptive to being religious. However, it could also be that being religious affects people's sensitivity to conflicting information; religious 'training' inoculates believers against contradictions and violations of their worldview. This notion parallels findings from mindfulness meditation research reporting evidence that meditation training increases cognitive control as it teaches practitioners to suppress irrelevant information (Moore & Malinowski, 2009; Teper & Inzlicht, 2012), with meditation experts showing less activation in brain areas implicated in attention and cognitive control (e.g., the anterior cingulate cortex; Brefczynski-Lewis et al., 2007). As such, mindfulness meditation may train practitioners to flexibly suppress irrelevant information – resulting in increased cognitive control. A similar process may be at play in religious training, in which people also engage in mental practices to maintain attention (e.g., meditative prayer) and to inhibit irrelevant (e.g., sinful) thoughts. On the other hand, naturalness of religion accounts posit that religious concepts (e.g., mind-body dualism, supernatural agents) are highly intuitive and that it is in fact non-religiosity that requires cognitive effort to suppress or reject these intuitions (J. L. Barrett, 2000; Bloom, 2007; Boyer, 2008; Norenzayan & Gervais, 2013). This implies that 'secular training' (e.g., analytic thinking and scientific reasoning), rather than religious training, involves suppressing intuitive information and enhancing the salience of analytic alternatives – resulting in increased cognitive control for non-religious compared to religious individuals.

In line with this latter suggestion, several empirical studies found that increased religiosity is related to a decreased cognitive performance, especially when a logically correct response must override a conflicting intuitive response (e.g., in a base-rate fallacy test; Daws and Hampshire, 2017; Good et al., 2015; Pennycook et al., 2014; Zmigrod et al., 2019). Other behavioral studies correlated individuals' self-reported level of religiosity with their performance on low-level cognitive control tasks such as the Go/No-go task or the Stroop task. These studies present a mixed bag of evidence; some report a positive relationship (Inzlicht et al., 2009), an inconsistent pattern (Inzlicht & Tullett, 2010), or no relationship (Kossowska et al., 2016) between religiosity and cognitive control (in terms of accuracy and reaction times).

In addition to this behavioral research, a few neuroscientific studies have been conducted on the association between religiosity and conflict sensitivity. For instance, an fMRI study investigated brain responses in devoted religious believers who listened to intercessory prayer. When participants believed that the prayer was pronounced by a charismatic religious authority, they showed a reduced activation of their frontal executive network, including the dorsolateral prefrontal cortex (DLPC) and the ACC, which have been associated with conflict detection (Schjoedt et al., 2011). Furthermore, Inzlicht et al. (2009) conducted a series of EEG studies looking at the relation between religiosity and the error-related negativity (ERN; Inzlicht and Tullett, 2010; Inzlicht et al., 2009). Compared to skeptics, religious believers demonstrated a smaller ERN amplitude in response to errors on a color-word Stroop task (Inzlicht et al., 2009). The authors suggest that these findings reflect the palliative effects of religiosity on distress responses: religious believers experience less distress in association with committing an error and this is reflected in a reduced ERN amplitude. There is, however, an open-ended debate on the functional significance of the ERN; while some researchers interpret the ERN primarily as an affective (i.e., distress) signal, others emphasize that it mainly reflects conflict-sensitivity (Botvinick et al., 2001; Bush et al., 2000; Carter et al., 1998; Hajcak et al., 2005; M. E. Maier & Steinhauser, 2016; N. Yeung et al., 2004).

Relatedly, different views have been proposed on how the relation between religiosity and ACC conflict activity should be interpreted; whereas Inzlicht et al. (2011) suggest that ACC activity in this context reflects error distress, Schjoedt and Bulbulia (2011) argue that the interpretation of ACC activity as reflecting purely cognitive conflict sensitivity is more parsimonious. We believe this discussion partly hinges upon the operationalisation of 'conflict'. EEG studies on cognitive conflict have typically studied the ERN as a proxy for ACC activity. The ERN is an *error*-related signal and reflects neural activity associated with incorrect vs. correct responses, i.e., conflict at the level of the behavioral response (hereafter: response conflict). In contrast, fMRI studies on cognitive conflict typically focus on the neural activity associated with incongruent vs. congruent stimulus trials, i.e., conflict at the level of information processing (hereafter: informational conflict). Although there is often a correlation between response conflict[1] and informational conflict, not all incongruent trials re-

[1] Response conflict is here defined as the conflict between the actual and the correct response,

sult in errors, nor do all congruent trials by definition result in correct responses. It is therefore important to dissociate between these two levels of conflict and their associated neural activity (cf. Tang et al., 2006; van Veen & Carter, 2005).

It thus remains unclear to what extent religiosity is related to a reduced sensitivity for response conflict (e.g., responding with 'green' when it should have been 'red') or to a reduced sensitivity for informational conflict (e.g., seeing the word 'green' printed in a red font). An effect for *response conflict* should be reflected in a relationship between religiosity and the strength of the error–correct Stroop contrast in the fMRI data, which would be a direct replication of the study by Inzlicht et al. (2009) and their proposed framework (Inzlicht et al., 2011; Proulx et al., 2012). An effect for *informational conflict* should be reflected in a relationship between religiosity and the strength of the incongruent–congruent Stroop contrast in the fMRI data. Schjoedt and Bulbulia (2011), for instance, indeed seem to interpret Inzlicht et al.'s results as religious believers' inattention to conflict monitoring. In everyday life, both sources of conflict detection could play a role in the maintenance of religious beliefs, e.g., when a believer simply does not detect the incongruency between different sources of information or when he / she fails to suppress an intuitive but objectively incorrect answer.

Taking the distinction between response conflict and informational conflict into account, here we investigated two different hypotheses regarding the relation between religiosity and cognitive conflict sensitivity: (1) there is a negative relationship between religiosity and ACC activity induced by response conflict (i.e., the incorrect–correct response contrast), and (2) there is a negative relationship between religiosity and ACC activity induced by informational conflict (i.e., the incongruent–congruent Stroop contrast). We note that both hypotheses are not mutually exclusive, as religiosity could be related to both mechanisms of conflict detection.[2]

Although earlier studies provide preliminary evidence for the religiosity–conflict sensitivity relation, we believe the present study –including a conceptual replication of the seminal study by Inzlicht et al. (2009)– is important for the following reasons. First, in order to substantiate the notion that religious believers are characterized by a *general* tendency for reduced conflict sensitivity at the neural level, a significant correlation or inter-group difference should be established. So far, only three studies found evidence for an inverse relation between religious beliefs and conflict-induced ACC activity; Inzlicht et al. (2009) showed that religious zeal and belief in God were associated with a reduced ERN response and Kossowska et al. (2016) similarly found that religious fundamentalism was related to a reduced N2 response on the Stroop task, albeit only in the uncertainty condition where participants performed the task under undefined time pressure. Another study failed to find a correlation between neurophysiological measures and religiosity (though the authors did find an experimental effect of priming God's forgiving nature on the ERN; Good et al., 2015). Second, with the exception of Good et al. (2015, $n$ = 108), all experiments

rather than the prepotent and the correct response.

[2] Based on the aforementioned theories addressing believers' failure to notice incompatibility between different sources of contradicting information, we would primarily expect a negative association between religiosity and informational conflict (rather than response conflict). However, from an empirical perspective, our study most closely resembles the design by Inzlicht et al. (2009), who measured and obtained support for a relation between religiosity and neural markers of response conflict.

linking religiosity to ACC activity included small samples and were therefore most likely underpowered (i.e., Inzlicht et al., 2009, $n = 28$ [Study 1], $n = 22$ [Study 2]; Kossowska et al., 2016, $n = 37$) Third, the hypothesized relation between religiosity and cognitive conflict is primarily based on either behavioral or EEG data. EEG studies, however, can offer only indirect evidence for the involvement of specific brain areas (Gazzaniga & Ivry, 2013). The use of fMRI may complement the existing findings, as fMRI allows for a higher spatial specificity, and may thus provide more conclusive evidence regarding the role of the ACC in the acceptance and maintenance of religious beliefs. Finally, the current study design allowed us to dissociate between neural effects related to response conflict (i.e., activity predicted by response accuracy) and to informational conflict (i.e., activity predicted by Stroop congruency). This may help to disentangle the 'conflict sensitivity' accounts of religiosity, and hence affords a more precise theoretical interpretation of the existing data.

## 5.2 Hypotheses

We tested eight hypotheses, four of which were based on our research questions and four that served as 'outcome neutral tests' (Chambers et al., 2014). The four outcome neutral tests were used to validate that our task did indeed induce cognitive conflict (reflected in accuracy and Stroop interference effects), that error commission was reflected in ACC activity, and that informational conflict was reflected in ACC activity. The corresponding outcome neutral hypotheses for the behavioral measures were: ($\mathcal{H}_1$) participants are more accurate on congruent compared to incongruent Stroop trials, and ($\mathcal{H}_2$) participants respond faster on congruent compared to incongruent Stroop trials. Outcome neutral hypotheses for the neural measures were: ($\mathcal{H}_3$) errors on the Stroop task induce more ACC activity compared to correct responses, on average across subjects, and ($\mathcal{H}_4$) incongruent Stroop trials induce more ACC activity compared to congruent trials, on average across subjects.

Conditional on establishing the effects related to hypotheses 1–4, we tested four corresponding hypotheses about the relation between religiosity and conflict sensitivity. For the behavioral measures, we hypothesized that ($\mathcal{H}_5$) Stroop accuracy is negatively related to religiosity, and ($\mathcal{H}_6$) Stroop interference (i.e., the difference in RT for incongruent vs. congruent trials) is positively related to religiosity, indicating decreased cognitive performance. We note that, based on the existing literature one could hypothesize both a positive and a negative relationship between religiosity and conflict detection; on the one hand, religiosity is associated with reduced response conflict and hence smaller interference effects (cf. Inzlicht et al., 2011). On the other hand, religiosity is associated with an increased tendency for intuitive responding, which means that more effort is required to overcome these intuitive response on incongruent Stroop trials, hence larger interference effects should be expected (cf. Pennycook et al., 2014). Despite these divergent theoretical predictions, most studies have not found any association between religiosity and Stroop interference (Inzlicht et al., 2009, Study 1; Inzlicht & Tullett, 2010; Kossowska et al., 2016), except for Study 2 by Inzlicht et al. (2009), in which a positive correlation between religiosity and Stroop interference was reported. Here, in line with the latter finding we hypothesized a positive relationship between religiosity and Stroop interference.

For the neural measures, we hypothesized that ($\mathcal{H}_7$) the size of the error–correct

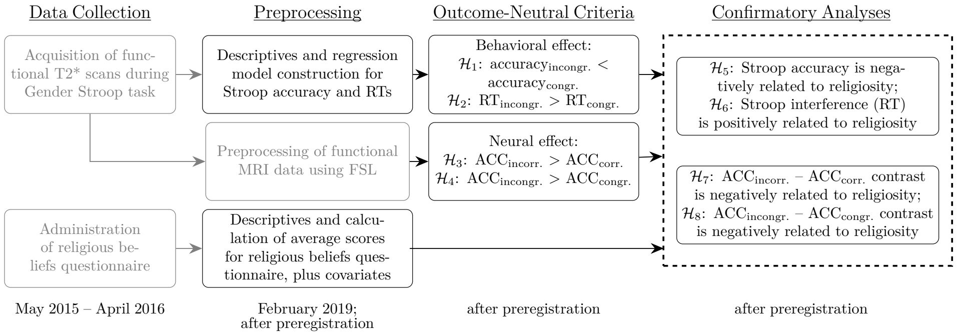

Data Collection
Preprocessing
Outcome-Neutral Criteria
Confirmatory Analyses
Acquisition of functional T2* scans during Gender Stroop task
Descriptives and regression model construction for Stroop accuracy and RTs
Behavioral effect: $\mathcal{H}_1$: accuracy$_{\text{incongr.}}$ < accuracy$_{\text{congr.}}$ $\mathcal{H}_2$: RT$_{\text{incongr.}}$ > RT$_{\text{congr.}}$
$\mathcal{H}_5$: Stroop accuracy is negatively related to religiosity; $\mathcal{H}_6$: Stroop interference (RT) is positively related to religiosity
Preprocessing of functional MRI data using FSL
Neural effect: $\mathcal{H}_3$: ACC$_{\text{incorr.}}$ > ACC$_{\text{corr.}}$ $\mathcal{H}_4$: ACC$_{\text{incongr.}}$ > ACC$_{\text{congr.}}$
$\mathcal{H}_7$: ACC$_{\text{incorr.}}$ − ACC$_{\text{corr.}}$ contrast is negatively related to religiosity; $\mathcal{H}_8$: ACC$_{\text{incongr.}}$ − ACC$_{\text{congr.}}$ contrast is negatively related to religiosity
Administration of religious beliefs questionnaire
Descriptives and calculation of average scores for religious beliefs questionnaire, plus covariates
May 2015 – April 2016
February 2019; after preregistration
after preregistration
after preregistration

**Figure 5.1:** Overview of data acquisition and analysis. Boxes marked in grey had already been completed prior to commencing this project. Boxed marked in black represent the analysis steps for the present study, which were determined in the preregistration.

response BOLD signal contrast (i.e., difference in BOLD signal between errors and correct responses) in the ACC is negatively related to religiosity, on average across subjects (cf. Inzlicht et al., 2009), and ($\mathcal{H}_8$) the size of the incongruent–congruent BOLD signal contrast (i.e., difference in BOLD signal between the incongruent and congruent condition) in the ACC is negatively related to religiosity, on average across subjects. All hypotheses were preregistered on the Open Science Framework (see https://osf.io/nspxb/registrations). Finally, we added exploratory whole-brain analyses to explore whether religiosity is associated with conflict-induced neural activity in any other brain areas besides the ACC.

## 5.3 Methods

### 5.3.1 Overview

The data for this study had already been collected as part of the Population Imaging (PIoP) project (May 2015 - April 2016), conducted at the Spinoza Center for Neuroimaging at the University of Amsterdam (see Appendix 5.A for a description of the project). An overview of the data collection and analysis procedure is presented in Figure 5.1. All hypotheses were formulated independently without any knowledge of the preprocessed data, and the analysis pipeline was developed and preregistered prior to data inspection.[3] The preregistration can be accessed on the OSF (https://osf.io/nspxb/). This folder also contains the anonymized raw and processed data and the R scripts used to preprocess the behavioral data and to conduct the confirmatory analyses (including all figures). The preprocessing scripts for the fMRI analysis and the exploratory fMRI analyses can be found at https://github.com/lukassnoek/ReligiosityFMRI. The (uncorrected) brain maps can be found at https://neurovault.org/collections/6139/.

[3] Specifically, LS was involved in data collection and (pre)processing the MRI data and has no access to the religiosity data. MvE and SH formulated the research questions and hypotheses without any access to the MRI data.

### 5.3.2 Participants

Participants were students who were recruited at the University of Amsterdam and received a financial remuneration. Participants were screened for MRI contraindications before MRI data acquisition. The intended number of participants was 250, but due to technical problems during part of the acquisition process, only 244 participants yielded usable MRI data. Of those 244, data from 20 subjects were excluded due to artifacts in the MRI data due to scanner instabilities or errors during export and/or reconstruction of the data. Additionally, 10 participants were excluded because they did not complete the task of interest (i.e., the gender-Stroop task). These exclusions were known at the time of the preregistration.

We entered the analysis phase with data from $N = 214$ participants. Out of these 214, eight participants were excluded –as preregistered– because they did not complete the religiosity questionnaire or lacked data on the covariates of interest (age, gender, and intelligence). We additionally preregistered to exclude participants whose accuracy was lower than 65%, because this indicates performance at chance level. This means that participants who responded correctly on fewer than 63 out of the 96 trials were excluded. Furthermore, participants who did not respond within the response interval on more than 20% of the Stroop trials were also excluded. As the minimum response interval of 4500ms is assumed to be sufficient for timely responses, missed responses on more than 20% of the trials were taken to indicate that participants did not understand or perform the task adequately. These criteria led to the exclusion of 14 participants, yielding a total sample size of 193. In addition, for the fMRI analyses, there were 21 participants who did not make any mistake during the task, preventing us from calculating the 'incorrect–correct' contrast.[4] As such, the confirmatory ROI and whole-brain analyses of this contrast were based on data from 172 participants. All other analyses were done on a total of $N = 193$ participants with complete data. The final sample consisted of 109 (56.5%) women and 84 (43.5%) men. The average age of the participants was 22.2 years ($SD = 1.9$; range $= 18 - 26$).

The study was approved by the local ethics committee at the Psychology Department of the University of Amsterdam (Project #2015-EXT-4366) and all participants were treated in accordance with the Declaration of Helsinki.

#### 5.3.2.1 Sample Size Justification

The sample size was determined based on the target of the overall project minus exclusions due to artifacts in the data, incomplete data, or preregistered quality criteria. As there were no existing fMRI studies on the relation between religiosity and cognitive conflict processing –only EEG studies– we could not perform a power analysis. However, we note that a sample of $N \approx 200$ is substantially large for an fMRI study (Szucs & Ioannidis, 2017)[5] and exceeds the recommended minimum sample size of $N = 100$ for correlational (neuroimaging) research (Dubois & Adolphs, 2016; Schönbrodt & Perugini, 2013).

[4] Of the 21 excluded participants, 19 made no errors and 2 participants made 1 error, but no reliable signal could be extracted for this error trial.

[5] This meta-analysis reports a median sample size of approximately 22 for fMRI studies.

### 5.3.3 Procedure

The study ran from May 2015 until April 2016. On each testing day, two participants were tested, which took approximately 4 hours and included an extensive behavioral test battery (approximately 2.5 hours) and an MRI session (approximately 1.5 hours). Participants received a financial remuneration of 50 euros. The order of behavioral and MRI sessions were counterbalanced across participants.

### 5.3.4 Study Design

The study involved a mixed design with Stroop congruency as the within-subjects variable and religiosity as the between-subjects continuous individual differences variable. The main part of the study qualified as an observational study; we investigated the correlation between performance on the Stroop task and religiosity, and between BOLD-fMRI activity and religiosity, without manipulating any variables except for trial congruency (congruent vs. incongruent Stroop trials). The fMRI task involved a rapid event-related design; a hypothesized BOLD response was modelled following the presentation of facial stimuli in the congruent or incongruent condition, as well as following correct and incorrect responses.

### 5.3.5 Stroop Task

We used a face-gender variant of the Stroop task (adapted from Egner et al., 2010), often referred to as the 'gender-Stroop' task, in which pictures of faces from either gender are paired with the corresponding (i.e., congruent) or opposite (i.e., incongruent) gender label (see below for details on the task and example pictures of the stimuli). The face-gender variant of the Stroop task (Egner & Hirsch, 2005) has been shown to induce significant behavioral conflict and neural ACC activation (Egner et al., 2008).[6] Each trial consisted of a photographic stimulus depicting either a male or female face, with the gender label 'MAN' or 'WOMAN' superimposed in red, resulting in gender-congruent and gender-incongruent stimuli (see Figure 5.2). The Stroop condition –congruent vs. incongruent– thus formed the within-subjects manipulated variable.

The stimuli set consisted of a total of 12 female and 12 male faces, with the labels 'man', 'sir','woman', and 'lady', both in lower- and uppercase added to the pictures (e.g., 'sir' and 'SIR').[7] All combinations appeared exactly one time, resulting in 96 unique trials (48 congruent and 48 incongruent). Participants were always instructed to respond to the gender of the pictured face, ignoring the distractor word.

The stimuli were presented for 500ms with a variable inter-trial interval ranging between 4000-6000ms, in steps of 500ms. Participants could respond from the beginning of the stimulus presentation until the end of the inter-trial interval (i.e., minimum response interval was 4500, maximum response interval was 6500), using their left

[6] The face Stroop task - instead of the regular word-color variant - was chosen because it offers optimal opportunities for dissociating between perceptual processing of target and distractor dimensions, as processing of the distractor faces can straightforwardly be linked to activation patterns in the fusiform face area (FFA; Egner and Hirsch, 2005). In the current study, however, we were mainly interested in the cognitive conflict aspect rather than perceptual processing, and therefore solely focused on activation in the ACC.

[7] The Dutch labels were 'man', 'heer','vrouw', and 'dame', respectively.

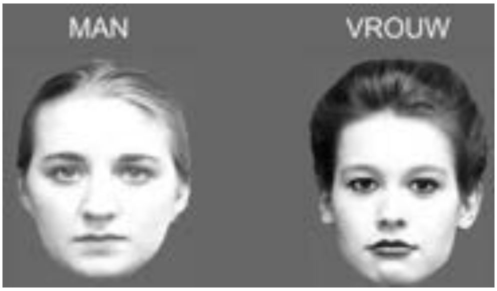


**Figure 5.2:** Stimuli as used in the face-gender Stroop task. Distractor words could be incongruent (left) and congruent (right) with the target face. NB. Translations of the Dutch labels: 'MAN' = 'MAN' and 'VROUW' = 'WOMAN'.

and right index finger. If no button was pressed during this interval, the trial was recorded as a 'miss'. Stimuli were presented using Presentation (Neurobehavioral Systems, www.neurobs.com), and displayed on a back-projection screen that was viewed by the subjects via a mirror attached to the head coil.

### 5.3.6 Religiosity Measures

Our religiosity measure consisted of 7 items that were based on religiosity questions included in the World Values Survey (WVS; World Values Survey, 2010), covering religious identification, beliefs, values, and behaviors (institutionalized such as church attendance and private such as prayer). Besides having high face-validity, these measures have been validated in other studies (Lindeman et al., 2015; Norenzayan et al., 2012; Stavrova, 2015) and the items have been used in previous studies (Maij et al., 2017; van Elk & Snoek, 2020). The items were evaluated on a 5-point Likert scale ranging from *not at all* to *very much*; see Table 5.1 for the exact items. Ratings on the seven religiosity items were tallied to create an average religiosity score per participant ($M = 1.74$, $SD = 0.84$). Cronbach's alpha for the 7-item religiosity scale was .89, indicating good internal consistency. For the analyses, these average scores were standardized. As anticipated in the preregistration, the distribution of the religiosity data was indeed positively-skewed, since our sample consisted of highly secular students. Although non-normality may reduce statistical power (Poldrack et al., 2011), it does not pose a problem for our analysis, since Bayesian linear regression models –like general(ized) linear models in general– do not assume normality of predictors (solely of model residuals).

### 5.3.7 Additional Variables

Gender, age, and intelligence were included as covariates in the analyses of the main hypotheses. Intelligence was indexed by the sum score on the 36 item version (set II) of Raven's Advances Progressive Matrices Test (Raven et al., 1998; Raven, 2000). The rationale for including these measures as covariates in our analysis was to control for the potential confound that any religiosity effect may be driven by other individual differences that are known to be associated with religiosity; females are typically more

**Table 5.1:** Items of the religiosity scale

1. To what extent do you consider yourself to be religious?
2. To what extent do you believe in God or a supernatural being?
3. To what extent do you believe in life after death?
4. My faith is important to me.
5. My faith affects my thinking and practice in daily life.
6. I pray daily.
7. I visit a church or religious meeting on a weekly basis.

*Note.* All items were measured on a 5-point scale ranging from *not at all* to *very much*.

religious than males (Miller & Hoffmann, 1995), older people tend to be more religious than younger people (Argue et al., 1999), and people scoring high on intelligence are on average less religious (Zuckerman et al., 2013). Age and intelligence scores were standardized in the analyses.

Since the proposed study was part of a larger project, a number of extra tasks and questionnaires were administered to the participants (see Appendix 5.A for a description). These measures were not included in the present study.

### 5.3.8 fMRI Data Acquisition

Subjects were tested using a Philips Achieva 3T MRI scanner and a 32-channel SENSE headcoil. A survey scan was made for spatial planning of the subsequent scans. After the survey scan, five functional (T2*-weighted BOLD-fMRI) scans (corresponding to five different tasks, including the gender-Stroop task; see Appendix 5.A for an overview of the other tasks), one structural (T1-weighted) scan, and one diffusion-weighted (DWI) scan were acquired. The DWI scan will not be described further, as it is not relevant to the current study. The Stroop task was done during the second scan of the session (not including the survey scan).

The structural T1-weighted scan was acquired using 3D fast field echo (TR: 82 ms, TE: 38 ms, flip angle: 8°, FOV: 240 × 18 mm, 220 slices acquired using single-shot ascending slice order and a voxel size of 1.0 × 1.0 × 1.0 mm). The functional T2*-weighted gradient echo sequences (single shot, echo planar imaging) were run. The following parameters were used for the MRI sequence during the gender-Stroop task: TR=2000 ms, TE=27.63 ms, flip angle: 76.1°, FOV: 240 × 240 mm, in-plane resolution 64 × 64, 37 slices (with ascending slice acquisition), slice thickness 3 mm, slice gap 0.3 mm, voxel size 3 × 3 × 3 mm), covering the entire brain. During the Stroop task, 245 volumes were acquired.

### 5.3.9 Preprocessing

Preprocessing was performed using `fmriprep` version 1.0.15 (Esteban et al., 2018; Esteban et al., 2019), a Nipype (Gorgolewski et al., 2011; Gorgolewski et al., 2017) based tool. `fmriprep` was run using the package's Docker interface. Each T1w (T1-weighted) volume was corrected for INU (intensity non-uniformity) using `N4BiasFieldCorrection` v2.1.0 (Tustison et al., 2010) and skull-stripped using antsBrainExtraction.sh v2.1.0 (using the OASIS template). Brain surfaces were reconstructed using recon-all from FreeSurfer v6.0.1 (Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (A. Klein et al., 2017). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009) was performed through nonlinear registration with the antsRegistration tool of ANTs v2.1.0 (Avants et al., 2008), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (Zhang et al., 2001; FSL v5.0.9).

Functional data was motion corrected using mcflirt (Jenkinson et al., 2002; FSL v5.0.9). 'Fieldmap-less' distortion correction was performed by co-registering the functional image to the same-subject T1w image with intensity inverted (Huntenburg, 2014; S. Wang et al., 2017) constrained with an average fieldmap template (Treiber et al., 2016), implemented with antsRegistration (ANTs). This was followed by co-registration to the corresponding T1w using boundary-based registration (Greve & Fischl, 2009) with 9 degrees of freedom, using bbregister (FreeSurfer v6.0.1). Motion correcting transformations, field distortion correcting warp, BOLD-to-T1w transformation and T1w-to-template (MNI) warp were concatenated and applied in a single step using antsApplyTransforms (ANTs v2.1.0) using Lanczos interpolation. Functional data was smoothed with a 5 mm FWHM Gaussian kernel. Many internal operations of fmriprep use Nilearn (Abraham et al., 2014), principally within the BOLD-processing workflow. For more details of the pipeline see http://fmriprep.readthedocs.io.

#### 5.3.9.1 Quality Control

After preprocessing, the MRIQC package (Esteban et al., 2017) was used to generate visual reports of the data and results of several intermediate preprocessing steps. These reports were visually checked for image artifacts, such as ghosting, excessive motion, and reconstruction errors. Participants displaying such issues were excluded from further analysis.

#### 5.3.9.2 fMRI First-Level Model

The fMRI timeseries were modelled using a first level (i.e., subject-specific) GLM, using the implementation provided by the nistats Python package (https://nistats.github.io; Abraham et al., 2014; version rel0.0.1b). The GLM included four predictors modelling elements of the task: incongruent trials, congruent trials, correct trials, and incorrect trials. If a participant did not make any mistakes, the 'incorrect trials' predictor was left out. The predictors were convolved with a canonical hemodynamic response function (HRF; Glover, 1999). Onsets for the (in)congruent trial predictors were defined at the onset of the image and had a fixed duration of 0.5 seconds. Onsets for the (in)correct trial predictors were defined at the onset of the response. Additionally, six motion regressors (reflecting the translation and rotation parameters in three dimensions) were included as covariates. GLMs were fit with AR1 autocorrelation correction. After fitting the GLMs, the following contrasts were computed: 'incorrect–correct' and 'incongruent–congruent'. The parameters –beta parameters– and associated variance terms from these contrasts were used in subsequent confirmatory ROI analyses and exploratory whole-brain analyses.[8]

#### 5.3.9.3 fMRI Group-Level Model (Exploratory)

In addition to the confirmatory analyses, we also performed an exploratory whole-brain analysis of the effect of religiosity on fMRI activity associated with response conflict (i.e,. $\mathcal{H}_7$) and informational conflict (i.e., $\mathcal{H}_8$). Similar to the confirmatory analyses, in addition to religiosity, the variables age, gender, and intelligence were added as covariates to the model. In the group-level model and in accordance with the 'summary statistics approach', the first-level 'incorrect–correct' and 'incongruent–congruent' contrast estimates represent the dependent variables, while religiosity, age, gender, and intelligence represent the independent variables. For the participants who did not make any error, we could not compute the 'incorrect-correct' contrast and they were thus excluded from the group-analysis of the 'incorrect-correct' contrast.

We used the FSL tool `randomise` (Winkler et al., 2014) in combination with `threshold-free cluster enhancement` (S. M. Smith & Nichols, 2009) to perform a non-parametric group-analysis of the effect of religiosity. We ran $10,000$ permutations. Specifically, we tested for a non-directional (two-tailed) effect of religiosity variable (controlled for the other covariates). In addition, as 'outcome neutral tests', we computed the average of the first-level contrasts ('intercept-only' model) for both the 'incorrect-correct' and 'incongruent-congruent' first-level contrasts. We corrected for multiple comparisons using the distribution of the 'maximum statistic' under the null-hypothesis (i.e., the default in `randomise`) with a voxel-level $\alpha$ value of 0.025 (i.e., $\alpha = 0.05$ but corrected for two-sided tests; G. Chen et al., 2018). We plotted the significant voxels showing either a negative or positive effect of religiosity on a standard MNI152 brain.

### 5.3.10 ROI Definition

For this study's confirmatory ROI analyses, we used a preregistered ROI based on a conjunction of a functional ROI, derived from fMRI activity preferentially associated with 'error' (for $\mathcal{H}_3$ and $\mathcal{H}_7$) or 'conflict' (for $\mathcal{H}_4$ and $\mathcal{H}_8$) extracted using Neurosynth (Yarkoni et al., 2011), and an anatomical ROI based on the anatomical coordinates of the ACC, taken from the Harvard-Oxford cortical atlas (Craddock et al., 2012). The reasons for using a mask based on both a functional and anatomical ROI are twofold. First, the anatomical ROI of the ACC in the Harvard-Oxford atlas (and many others) consists of several putatively functionally different subregions (Gasquoine, 2013; Holroyd et al., 2004; Vogt, 2005). A functional ROI based on the Neurosynth

[8] We note that the current design was suboptimal in estimating the effect of informational conflict (but not response conflict) in the fMRI data. Due to insufficient 'jittering' of the interstimulus interval, the first-level predictors for congruent and incongruent trails were strongly negatively corrected ($\bar{r} = -0.9$). While this does not bias our results (the generalized least squares estimator we used is still unbiased), it *does* increase the variance of our first-level results, which in turn reduces the power of finding a correlation of religiosity with the first-level effect of informational conflict (operationalized by the 'incongruent-congruent' contrast). This issue only applies to the 'incongruent-congruent' contrast, not the 'incorrect-correct' contrast (as these predictors are much less correlated with each other, $\bar{r} = -0.2$).

database would resolve this issue of functional ambiguity within a single (anatomical) ROI; however, the Neurosynth maps for 'error' and 'conflict' contain more brain areas than just the ACC (such as the bilateral insula). Therefore, by using the conjunction between the functional ROIs based on Neurosynth and the anatomical ROI of the ACC, we restrict our analyses to a single *anatomical* region that is most likely to be *functionally* relevant for the psychological constructs of interest, i.e., response conflict ("error") and informational conflict ("conflict"). We realize that due to the ambiguity of the term 'conflict' (which may refer to informational conflict or response conflict), the Neurosynth map for 'conflict' will likely also be based on studies involving response conflict. Although not ideal, we believe that this method is the most appropriate way to define our ROI.

Specifically, for our functional ROI, we used the Neurosynth Python package to conduct separate meta-analyses of the terms "error*" and "conflict*", with a frequency threshold of 0.001[9]. We used the 'association test map' from the meta-analysis output (FDR-thresholded for multiple comparisons at $p < 0.01$), which reflects voxels which are *preferentially* associated with the term 'error' and 'conflict', rather than other psychological constructs. For our anatomical ROI, we used the 'anterior cingulate cortex' region within the Harvard-Oxford cortical atlas. We will define the ACC within this probabilistic atlas as the set of voxels with a nonzero probability of belonging to the ACC. Our final ROI is based on the logical conjunction of these two ROIs (see Figure 5.3). For the confirmatory ROI analyses, we averaged the GLM parameters ($\hat{\beta}$, 'beta-values') and associated variance parameters (var[$\hat{\beta}$]) separately for the 'incorrect–correct' ($\mathcal{H}_3$ and $\mathcal{H}_7$) and 'incongruent–congruent' (for $\mathcal{H}_4$ and $\mathcal{H}_8$) first-level contrasts for each participant. These ROI-average parameters were subsequently analyzed in a hierarchical Bayesian regression model (see Statistical Models section for details).

### 5.3.11 Statistical Models

We applied hierarchical Bayesian models for all hypotheses to accommodate the hierarchical structure of the behavioral and fMRI data, with trials nested within participants. In the multilevel structure, we allow the overall performance and the effect of condition to vary between participants, by including random intercepts and random slopes, respectively. The random intercepts and slopes are desirable theoretically; we are interested in individual differences, hence we should allow effects to differ between individuals. Statistically, omitting the random slope has been shown to result in overestimation of the cross-level interaction term (i.e., the religiosity × condition effect) and the lower level main effect (i.e., the effect of condition; Heisig and Schaeffer, 2019). Finally, adopting this multilevel structure decreases the influence of trial noise through the process of hierarchical shrinkage (see Discussion; Rouder, Kumar, et al., 2019). We constructed the hierarchical Bayesian models using the R package `brms` (Bürkner, 2017), which relies on the programming language `Stan` (Carpenter et al., 2017). This package incorporates `bridgesampling` (Gronau, Singmann, et al., 2017) for hypothesis testing by means of Bayes factors (BF) and posterior probabilities. The

[9] These maps were generated on February 26th, 2019.

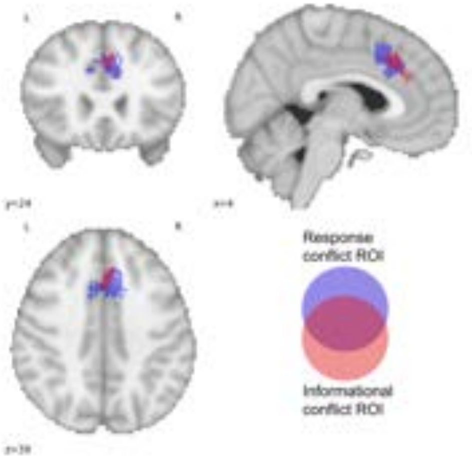


**Figure 5.3:** ROIs used for our confirmatory ROI analyses of the effect of religiosity on response conflict and informational conflict.

general form of our multilevel regression models is:

$$y_{ij} \sim \mathcal{N}(\beta_0 + \beta_{0j} + (\beta_1 + \beta_{1j})x_{ij}, \sigma^2) \tag{5.1}$$

where $y_{ij}$ is the outcome per trial per participant, and $x_{ij}$ the corresponding value of the predictor. The subscript $i$ is for the individual trials ($i = 1...n_{\text{trials}}$) and the subscript $j$ is for the participants ($j = 1...N$).

#### 5.3.11.1 Prior Specification

We note that the most relevant parameter for making inferences in our specified models is the $\beta_1$, i.e., the beta-weight for the (standardized) predictors of interest (e.g., Stroop condition, religiosity). As this parameter is used in the critical tests for our hypotheses, it is important to set appropriate priors particularly for this parameter. We chose $\beta_1 \sim \mathcal{N}(0, 1)$ for the (standardized) predictors. This prior is listed as a recommended 'generic weakly informative prior' in the Stan manual (Betancourt et al., 2015), and has been used in this context before (e.g., Gelman et al., 2015).

On the remaining parameters we used weakly-informative priors, whereby the priors for the regression weights ($\beta$'s) are derived from a normal distribution, and the priors on the scale parameters from a half-Cauchy distribution ($\mathcal{C}^+$; Gelman, 2006): $\beta_0 \sim \mathcal{N}(0, 10)$ for the fixed intercept; $\beta_{0j} \sim \mathcal{N}(0, \tau_0^2)$ for the varying part of the intercept per participant; $\beta_{1j} \sim \mathcal{N}(0, \tau_1^2)$ for the varying part of the predictor effect per participant; $\tau \sim \mathcal{C}^+(0, 2)$ for the participant-level variance. Finally, we used the default LKJ-correlation prior to model the covariance matrices in hierarchical models (Lewandowski et al., 2009). That is, we used $\Omega_\text{k} \sim \text{LKJ}(\zeta)$, with $\Omega_\text{k}$ being the correlation matrix and $\zeta$ set to 1.

#### 5.3.11.2 Interpretation of Evidence

Hypothesis testing was done by means of Bayes factors that evaluate the extent to which the data is likely under the alternative hypothesis (e.g., $\mathcal{H}_1$–$\mathcal{H}_8$) versus the corresponding null hypothesis $\mathcal{H}_0$. The Bayes factor (BF) reflects the change from prior hypothesis or model probabilities to posterior hypothesis or model probabilities and as such quantifies the evidence that the data provide for $\mathcal{H}_1$ versus $\mathcal{H}_0$, reflected by:

$$\underbrace{\frac{p(\mathcal{M}_1 \mid \text{data})}{p(\mathcal{M}_0 \mid \text{data})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}}_{\text{prior odds}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{M}_1)}{p(\text{data} \mid \mathcal{M}_0)}}_{\text{Bayes factor}} \tag{5.2}$$

where $\mathcal{M}_1$–$\mathcal{M}_8$ and $\mathcal{M}_0$ represent the models specified for $\mathcal{H}_1$–$\mathcal{H}_8$ and $\mathcal{H}_0$, respectively. The Bayes factor $\text{BF}_{10}$ then represents the ratio of the marginal likelihoods of the observed data under $\mathcal{M}_1$ and $\mathcal{M}_0$:

$$\text{BF}_{10} = \frac{p(\text{data} \mid \mathcal{M}_1)}{p(\text{data} \mid \mathcal{M}_0)} \tag{5.3}$$

As our hypotheses are directed, we computed order-restricted Bayes factors, i.e., $\text{BF}_{+0}$ in case of an expected positive effect. Note that the subscripts on Bayes factor 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. $\text{BF}_{+0}$ is used in case of a hypothesized positive effect for the reference group or a positive relation between variables; $\text{BF}_{-0}$ is used for a negative effect for the reference group or a negative relation between variables. As Bayes factors are fundamentally ratios that are transitive in nature, we can easily compute an order restricted Bayes factor; by (1) using the BF for the unrestricted model versus the null model, and (2) comparing the unrestricted model to an order restriction, we can then (3) use the resulting BFs to evaluate the order restriction versus the null model (Morey, 2015).

By default, prior model odds were assumed to be equal for both models that are compared against each other. As the evidence is quantified on a continuous scale, we also present the results as such. Nevertheless, we included a verbal summary of the results by means of the interpretation categories for Bayes factors proposed by Lee and Wagenmakers (2013, p. 105), based on the original labels specified by Jeffreys (1939). In addition to Bayes factors, we present the posterior model probabilities that are derived from the generated posterior samples.

For all outcome neutral tests we preregistered that a Bayes factor of at least 10 –the minimum value for strong evidence– was required to meet the criteria.

We declare that all models that are described below were constructed before the data were inspected. Additionally, all analyses were run as preregistered. Any deviations are explicitly mentioned in the chapter.

## 5.4 Results – Outcome Neutral Tests

### 5.4.1 Behavioral Stroop Effect – Accuracy

A hierarchical logistic regression model with varying intercepts for the participants and a varying slope for the effect of Stroop congruency was constructed to model response accuracy. In order to validate the presence of a congruency effect on accuracy, i.e., a Stroop effect, we compared the model for $\mathcal{H}_0$ containing only the varying intercept, to the model for $\mathcal{H}-$ containing the varying intercept and the negative effect of congruency. $\mathcal{H}-$ thus indicates that the incongruent condition *decreases* the probability of responding correctly on the Stroop task, relative to the congruent condition.

Results revealed a Bayes factor of $8.43 \times 10^{11}$ in favor of the alternative model ($\mathcal{M}_-$) relative to the null model ($\mathcal{M}_0$). That is, $\text{BF}_{-0} = 8.43 \times 10^{11}$, indicating that the data are about $10^{11}$ times more likely under the model assuming lower accuracy for incongruent Stroop trials than for congruent Stroop trials. In order words, the data provide strong evidence for the Stroop effect indexed by accuracy ($\mathcal{H}_1$). See Table 5.2 for a summary of the results of all four outcome neutral tests.

### 5.4.2 Behavioral Stroop Effect – Response Times

We used a similar hierarchical regression model with varying intercepts for the participants and a varying slope for the effect of Stroop condition to model reaction times. Note that only correct trials are included in the RT analysis. To account for the typical positive skew in RT data, we modelled reaction times as an ex-Gaussian distribution, i.e., a mixture of a Gaussian and an exponential distribution, which has been shown to fit empirical RT data well (Balota & Spieler, 1999; Balota & Yap, 2011; Whelan, 2008). This distribution is incorporated in the brms package, and thus only needed to be specified. Here we expected RTs to be longer for incongruent vs. congruent trials, hence the Bayes factor $\text{BF}_{+0}$ was calculated for ratio between the marginal likelihoods of the observed data under $\mathcal{H}_+$ versus $\mathcal{H}_0$. Again, we expected a Bayes factor of at least 10.

We obtained a Bayes factor of $3.53 \times 10^{67}$ in favor of $\mathcal{M}_+$, that is $\text{BF}_{+0} = 3.53 \times 10^{67}$. In other words, we collected strong evidence for the Stroop interference effect on reaction times ($\mathcal{H}_2$).

### 5.4.3 Neural Processing – Response Conflict

The hierarchical nature of the fMRI data –being derived from multiple trials– was already taken into account in the calculation of the ‘incorrect–correct’ contrast and the ‘incongruent–congruent’ contrast in FSL; we exported the beta-values for each contrast per participant, as well as the variance for the contrasts, i.e., $\hat{\beta}$ and var$[\hat{\beta}]$. The inclusion of the variance parameter in the Bayesian models is important, because it allows one to retain the uncertainty associated with the activation level contrast, which is typically lost or ignored when extracting fMRI data for ROI-analyses.[10] In order to test $\mathcal{H}_3$ that the average contrast of ACC activation – the average ‘intercept’ or $\hat{\beta}$ – was substantially different from 0, we used the function hypothesis which allows for directed hypothesis test of the specified parameters.[11] $\hat{\beta}$ is calculated as

[10] The possibility to include the variance of the observations in the regression model formula was added for the purpose of meta-analyses (Vuorre, 2016). However, it also serves the current purpose very well.

[11] The term intercept may be somewhat confusing here. Since the outcome variable is the contrast between the incongruent and congruent condition (i.e., the difference), we only include the intercept in this model, and hence look at the effect of the parameter ‘intercept’.

**Table 5.2:** Results Outcome Neutral Tests.

| Hypothesis | Bayes factor | $p(\mathcal{M}_\text{a})$ | Estimated coefficient |
|---|---|---|---|
| $\mathcal{H}_1$: accuracy$_{incongr.}$ < accuracy$_{congr.}$ | $10^{11}$ | 1 | $-0.64\,[-0.85, -0.46]$ |
| $\mathcal{H}_2$: RT$_{incongr.}$ > RT$_{congr.}$ | $10^{67}$ | 1 | $0.03\quad[0.02, 0.03]$ |
| $\mathcal{H}_3$: ACC$_{incorr.}$ > ACC$_{corr.}$ | $\infty$* | 1 | $3.26\quad[2.89, 3.64]$ |
| $\mathcal{H}_4$: ACC$_{incongr.}$ > ACC$_{congr.}$ | 157.7 | 0.99 | $0.15\quad[0.03, 0.26]$ |

*Note.* *Estimated to approach "infinity" as all posterior samples were in accordance with the order-restricted hypothesis. Bayes factors are the order-restricted Bayes factors for the alternative hypothesis of interest; BF$_{-0}$ for $\mathcal{H}_1$ and BF$_{+0}$ for $\mathcal{H}_2$–$\mathcal{H}_4$. $p(\mathcal{M}_\text{a})$ gives the posterior model probabilities of the alternative model versus the null model. Coefficients are the medians of the posterior distributions for the parameter of interest (i.e., Stroop condition or response accuracy) with 95% credible intervals in square brackets.

$(\hat{\beta}_{incorr.} - \hat{\beta}_{corr.})$, therefore the hypothesis states that $\hat{\beta}$ is larger than 0 (i.e., increased ACC activity for errors compared to correct responses). Here we calculated the Bayes factor for $\mathcal{H}_+$ stating that $\hat{\beta} > 0$.

We note that analyses that took the 'incorrect–correct' fMRI contrast as the dependent variable ($\mathcal{H}_3$ and $\mathcal{H}_7$) include data from 172 participants rather than 193, since some participants made no errors on the Stroop task.

The results showed evidence for the alternative hypothesis to approach "infinity", that is BF$_{+0} = \infty$. Note that this Bayes factor was estimated by testing the proportion of posterior samples that satisfy the hypothesis that the intercept > 0. When all posterior samples are in accordance with the hypothesis, a Bayes factor of "infinity" can be obtained. In this case that means that the Bayes factor is at least $60,000$ since the model included $60,000$ samples. In other words, the neural data provide strong evidence that the ACC is sensitive to response accuracy on the Stroop task.

### 5.4.4 Neural Processing – Informational Conflict

A similar procedure was used to test $\mathcal{H}_4$, this time with the ACC activity contrast for Stroop congruency instead of response outcomes. That is, a hierarchical regression model with a varying intercept for the participants was constructed. The Bayes factor was calculated for the hypothesis that $\hat{\beta}$ is larger than 0, since we expected $\hat{\beta}_{incongr.}$ to be larger than $\hat{\beta}_{congr.}$, resulting in a positive contrast. Again, a Bayes factor of at least 10 was required to pass the outcome neutral criterion test.

A Bayes factor of 157.7 in favor of the alternative hypothesis was obtained (i.e., BF$_{+0}$ = 157.7), indicating that the data provide strong evidence that the ACC is sensitive to informational conflict on the Stroop task.

The results of these four analyses indicate that all prespecified outcome neutral criteria were met.

## 5.5 Results – Main Preregistered Analyses

### 5.5.1 Behavioral Stroop Effect and Religiosity – Accuracy

In order to test $\mathcal{H}_5$ whether self-reported religiosity of individuals is related to their performance on a conflict-inducing Stroop task, an extended Bayesian hierarchical logistic regression model was constructed, by adding religiosity as second-level predictor. Specifically, the model for $\mathcal{H}_0$ included varying intercepts and varying slopes for Stroop condition (as before) per participant, plus the participant-level variables gender, age, and intelligence (i.e., the covariates). The model for the alternative hypothesis was identical plus the inclusion of religiosity as an additional participant-level predictor. Notably, an interaction term for religiosity × congruency was also included, as the effect of religiosity might be specific for performance in the conflict condition (i.e., the incongruent Stroop condition). As we expected a negative relation between religiosity and performance on the gender-Stroop task, we restricted the coefficient for religiosity to be negative in calculating the Bayes factor, i.e., we performed a one-sided test.[12] The ratio of marginal likelihoods for the data under $\mathcal{H}-$ versus $\mathcal{H}_0$, i.e., the Bayes factor, was calculated to determine the evidence for the predictive value of religiosity in explaining Stroop performance.

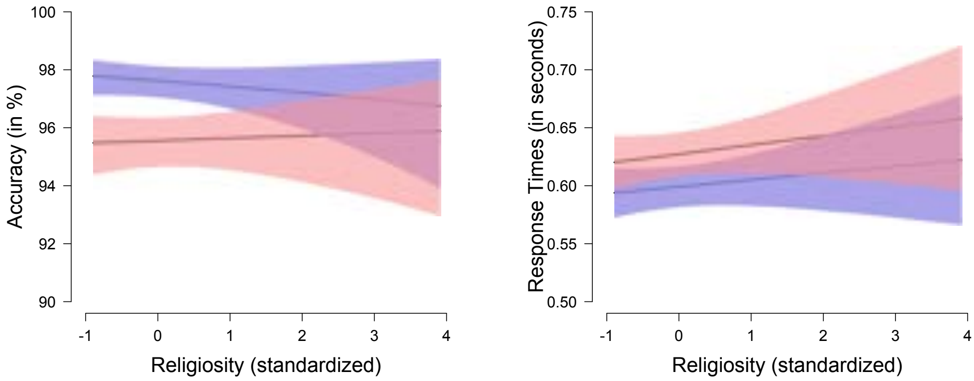
A Bayes factor of 0.022 was obtained (i.e., $\text{BF}_{-0} = 0.022$, $\text{BF}_{0-} = 44.8$), indicating that the data provided more support for the null model than for the religiosity model. This result qualifies as strong evidence that religiosity is not negatively related to accuracy on the Stroop task. The posterior medians and the 95% credible interval for the coefficients of religiosity ($-0.08\,[-0.25, 0.09]$) and of religiosity × Stroop condition ($0.10\,[-0.04, 0.24]$) indicate that neither religiosity, nor the interaction between religiosity and Stroop condition was related to performance on the Stroop task (see also Figure 5.4a). The results of all main hypotheses are also summarized in Table 5.3. The parameters in the regression models for the four main analyses are displayed in Figure 5.7 in Appendix 5.C.

### 5.5.2 Behavioral Stroop Effect and Religiosity – Response Times

We constructed a similar model with RT as the dependent variable; the model for $\mathcal{H}_0$ was a hierarchical ex-Gaussian regression model for RT with varying intercepts and a varying slope for Stroop condition – including participant gender, age, and intelligence as covariates. For $\mathcal{H}_+$, the model was identical with the added religiosity predictor and the religiosity × congruency interaction term. Again, we hypothesized that religiosity would be negatively related to Stroop performance, hence we expected a *positive* effect of religiosity on Stroop response times.

A Bayes factor of $3.93 \times 10^{-5}$ was obtained (i.e., $\text{BF}_{+0} = 3.93 \times 10^{-5}$, $\text{BF}_{0+} = 25461$). Similar to the accuracy analysis, this indicates that the data do not provide support for the hypothesis that religiosity is related to longer response times on the Stroop task. Rather, we obtained strong evidence for the null hypothesis. The posterior medians for the coefficients of religiosity ($0.01\,[-0.01, 0.02]$) and of religiosity × Stroop condition ($0.00\,[-0.00, 0.01]$) corroborate that there was no main effect of religiosity on response times, nor was there an interaction of religiosity × Stroop condition on response times (see also Figure 5.4b).

[12] The coefficient for the interaction term was not order-restricted.



**(a)** Stroop accuracy **(b)** Stroop response times

**Figure 5.4:** The marginal effect of religiosity on Stroop accuracy and response time, displayed per Stroop condition. The line with the blue 95% credible interval band indicates performance on congruent Stroop trials, the line with the red 95% credible interval band indicates performance on incongruent Stroop trials.

### 5.5.3 Neural Processing and Religiosity – Response Conflict

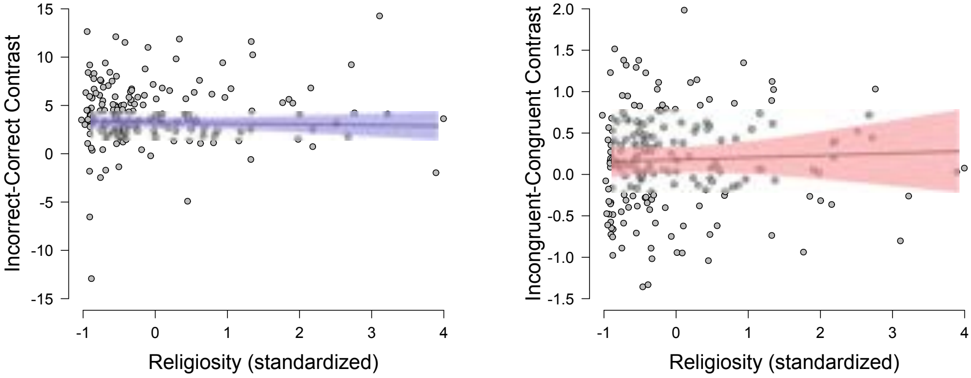
A Bayesian linear regression was performed in order to test $\mathcal{H}_7$ whether self-reported religiosity is related to the ACC sensitivity to incorrect vs. correct responses on the Stroop task. The beta-values for the BOLD contrast in our specified ROI served as the dependent variable, i.e., the extracted $\hat{\beta}$'s. Again, the variance of the individual beta-values was included to take the uncertainty of the contrast estimation into account. Religiosity served as the predictor of interest and gender, age, and intelligence were added as covariates. That is, we compared the model including the contrast-intercept and the covariates ($\mathcal{H}_0$) to the model additionally including the religiosity predictor. Based on the findings by Inzlicht et al. (2009), we expected a *negative* relation between religiosity and ACC activity induced by response conflict.

The results showed more evidence for the null model than for the model including religiosity as a predictor: $\text{BF}_{-0} = 0.286$ (i.e., $\text{BF}_{0-} = 3.49$). This Bayes factor is interpreted as moderate evidence against the hypothesis that religiosity is associated with reduced ACC sensitivity to response conflict in the Stroop task (i.e., the 'incorrect–correct' contrast). The posterior median and credible interval for the religiosity predictor were $-0.09\,[-0.44, 0.26]$. The scatterplot in Figure 5.5a illustrates the (absence of an) association between religiosity and sensitivity of the ACC to response conflict.

### 5.5.4 Neural Processing and Religiosity – Informational Conflict

The same model comparison was performed with regard to the stimulus congruency contrast (i.e., $\mathcal{H}_8$). Here, we used the $\hat{\beta}$'s of the incongruent–congruent BOLD contrast as the dependent variable. Again, we expected ACC activity to be negatively related to religiosity, while taking into account the effects of gender, age, and intelligence.

A Bayes factor of 0.046 ($\text{BF}_{-0} = 0.046$, $\text{BF}_{0-} = 21.9$) was obtained, indicating



**(a)** Religiosity and ACC sensitivity to response conflict

**(b)** Religiosity and ACC sensitivity to informational conflict

**Figure 5.5:** The relation between religiosity on the BOLD signal contrast for incorrect vs. correct responses on the Stroop task (left panel) and on the BOLD signal contrast for incongruent vs. congruent trials in the Stroop task (right panel). The plots display raw individual data points and Bayesian estimated linear effect of religiosity on the conflict-induced BOLD contrasts with 95% credible interval bands.

that the data provide strong evidence against the hypothesis that religiosity is related to reduced ACC sensitivity to informational conflict in the Stroop task (i.e., the ‘incongruent–congruent’ contrast). The posterior median and credible interval for the religiosity predictor were $0.03\,[-0.09, 0.15]$. The scatterplot in Figure 5.5b illustrates the (absence of an) association between religiosity and sensitivity of the ACC to informational conflict.

## 5.6 Results – Exploratory Whole-Brain Analyses

In addition to the confirmatory ROI analyses, we conducted an exploratory (non-parametric) whole-brain analysis of the effect of religiosity on both response conflict and informational conflict. In addition, we ran an ‘intercept-only’ model (estimating the average effect of response and informational conflict) as an outcome neutral test. All whole-brain *t*-value maps and associated ‘1-*p*-value’ maps can be viewed at and downloaded from Neurovault (https://identifiers.org/neurovault.collection:6139).

### 5.6.1 Outcome Neutral Tests

In Figure 5.6, we visualized the whole-brain results (as *t*-values) of the ‘intercept-only’ model for both the response conflict data (i.e., using the ‘incorrect–correct’ contrast; Figure 5.6A) and the informational conflict data (i.e., using the ‘incongruent–congruent’ contrast; Figure 5.6B).

Both whole-brain maps show widespread effects in areas known to be involved in error monitoring and cognitive conflict (such as the ACC and insula). Note that the effects (i.e., *t*-values) are much larger in the response conflict analysis, presumably due

**Table 5.3:** Results Main Analyses.

| Hypothesis | Bayes factor | $p(\mathcal{M}_\text{a})$ | Estimated coefficient |
|---|---|---|---|
| $\mathcal{H}_5$: Religiosity ↑ – Stroop performance (accuracy) ↓ | 0.022 (44.82) | 0.012 | $-0.08\,[-0.25, 0.09]$ |
| $\mathcal{H}_6$: Religiosity ↑ – Stroop response times ↑ | $10^{-5}$ (25461) | 0.000 | $0.01\,[-0.01, 0.02]$ |
| $\mathcal{H}_7$: Religiosity ↑ – ACC activity (response conflict) ↓ | 0.286 (3.49) | 0.172 | $-0.09\,[-0.44, 0.26]$ |
| $\mathcal{H}_8$: Religiosity ↑ – ACC activity (informational conflict) ↓ | 0.046 (21.87) | 0.064 | $0.03\,[-0.09, 0.15]$ |

*Note.* Bayes factors are the order-restricted Bayes factors for the alternative hypothesis of interest; $\text{BF}_{-0}$ for $\mathcal{H}_5$, $\mathcal{H}_7$, and $\mathcal{H}_8$ and $\text{BF}_{+0}$ for $\mathcal{H}_6$. Evidence for the null hypothesis is given between brackets. $p(\mathcal{M}_\text{a})$ gives the posterior model probabilities of the alternative model versus the null model. Coefficients are the medians of the posterior distributions for the parameter of interest (i.e., religiosity) with 95% credible intervals in square brackets.

to the relatively high variance in the first-level analysis stage due to high predictor correlation.

### 5.6.2 Neural Processing and Religiosity – Response Conflict

After multiple comparison correction, no voxels were significantly associated with religiosity in the response conflict analysis.

### 5.6.3 Neural Processing and Religiosity – Informational Conflict

Similar to the response conflict analysis, no voxels were significantly associated with religiosity after multiple comparison correction in the informational conflict analysis.

## 5.7 Discussion

In the current preregistered study we investigated whether religiosity is associated with a reduced sensitivity to cognitive conflict as measured through behavioral performance on the Stroop task and neural activation in the anterior cingulate cortex (ACC). The data from the outcome neutral tests provided strong evidence that the gender-Stroop task induced cognitive conflict at the behavioral level ($\mathcal{H}_1$ and $\mathcal{H}_2$) and that this was reflected in increased ACC activity. The neuroimaging data showed that the ACC was responsive to both response conflict (incorrect vs. correct responses; $\mathcal{H}_3$) and informational conflict (incongruent vs. congruent trials; $\mathcal{H}_4$). However, individual differences in religiosity were not related to performance on the Stroop task as measured in accuracy ($\mathcal{H}_5$) and response times ($\mathcal{H}_6$). We also did not observe the hypothesized relation between religiosity and neural activation related to response conflict ($\mathcal{H}_7$) or informational conflict ($\mathcal{H}_8$). Overall, we obtained moderate to strong

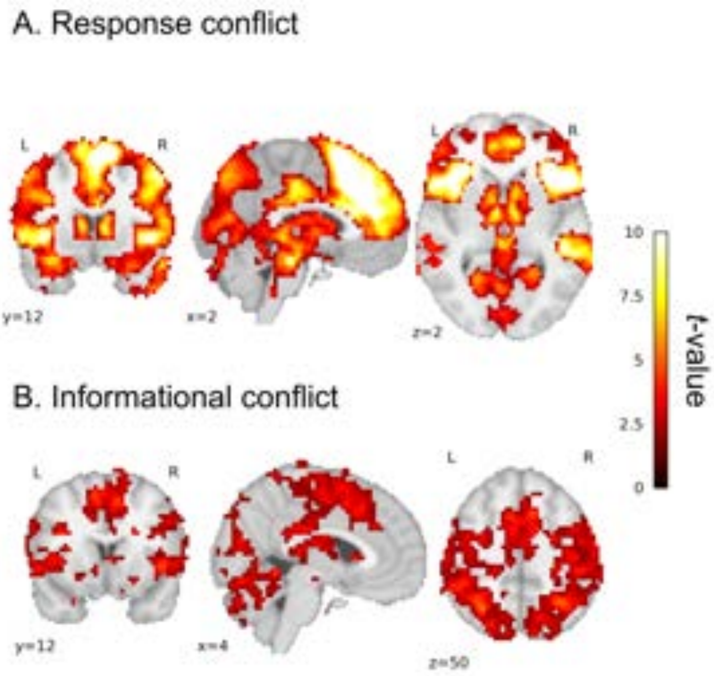


**Figure 5.6:** Brain maps with $t$-values corresponding to the outcome neutral ('intercept-only') test for both the (A) response conflict analysis and (B) informational conflict analysis. The brain maps were masked using $p$-values computed using FSL's *randomize* with threshold-free cluster enhancement, which we thresholded at $p < 0.05$.

evidence in favor of the null hypotheses according to which religiosity is unrelated to sensitivity to cognitive conflict. Exploratory whole-brain analyses similarly showed that conflict-induced neural activity was not associated with religiosity.

These results cast doubt on the theoretical claim that religiosity is related to a reduced process of conflict sensitivity. Although this idea is central to various theories about religious beliefs (e.g., Inzlicht & Tullett, 2010; Schjoedt et al., 2013; van Elk & Aleman, 2017), our study shows that religious believers may not be characterized by a *general tendency of* attenuated conflict sensitivity. An important motivation for conducting the current study was to address and overcome the limitations of previous studies in the field. We did so by increasing statistical power (i.e., we used a large sample) and by minimizing degrees of freedom (i.e., we preregistered all hypotheses, methods, and analyses and a priori specified a region of interest (ROI) for the fMRI analysis). Moreover, we curtailed the possibility of (unconscious) biases, as we separated the preprocessing of the fMRI data from the statistical analysis and only combined the fMRI data with the critical variable of interest (i.e., religiosity) in the final analysis steps.

It is important to note that our sample consisted largely of highly secular students; the average religiosity score was 1.74 on a 5-point scale and only 43% considered themselves at least somewhat religious. It could be that the number of religious believers in the sample was simply insufficient to detect an effect. Although this is a serious limitation that nuances the conclusiveness of the current findings, we still believe our study contributes to the existing literature. The fact that the Bayesian analyses showed evidence of absence rather than absence of evidence for the effect, strengthens our belief that previous claims about the association between religiosity and cognitive conflict sensitivity should be interpreted with caution.

Our null findings are perhaps not surprising in light of the recently voiced concerns about the replicability and reliability of neuroscientific findings, often related to problems of insufficiently powered studies (Button et al., 2013; Cremers et al., 2017; Szucs & Ioannidis, 2017) and general challenges in studying individual differences using neuroimaging (Dubois & Adolphs, 2016). For instance, Boekel et al. (2015) attempted to replicate 17 findings relating behavior to brain structures and found convincing evidence for only one out of the 17 included effects. Similarly, van Elk and Snoek (2020) recently failed to find support for the hypothesized relation between religiosity and grey matter volume in several brain areas that were identified in the literature as being associated with religiosity.

The current study employed the face-gender word variant of the Stroop task rather than the classical color-word Stroop task that has mostly been used in research on religiosity and cognitive conflict sensitivity. Both tasks rely on inhibition of the automatic reading process in order to name the semantic category, with the key distinction that competition takes place either between different features of the same item (i.e., the meaning and the printed color of the word) or between two different items (i.e., the meaning of the word and the 'meaning' of the picture), though also presented within the same visual field. Theoretically, we see no reason to assume that this small difference should be consequential for the religiosity – conflict sensitivity relation; previous claims are based on a general sensitivity for conflicting information, not exclusively for conflicting features within the same item (as in the color-Stroop task) or in superimposed items (as in the gender-Stroop task). Furthermore, based on the close similarities between the neurocognitive effects associated with both tasks, the picture-word and the color-word Stroop task are often assumed to reflect the same underlying process (e.g., MacLeod, 1991; Starreveld and La Heij, 2017; van Maanen et al., 2009, but see Dell'Acqua et al., 2007). Finally, the results of our outcome-neutral tests also provide no indication for substantially different mechanisms at play relative to the classical Stroop task; we find interference effects in the same order of magnitude (i.e., 50.5 ms; Haaf and Rouder, 2019; MacLeod, 1991; Stroop, 1935), and observe the same implicated brain areas (i.e., the ACC, the dorsolateral prefrontal cortex; MacLeod and MacDonald, 2000).

The fact that we did not find behavioral evidence for impaired nor for enhanced Stroop performance among religious believers might indicate that religiosity is unrelated to low-level cognitive control processes. At the same time, the null finding may also reflect the paradox that highly robust experimental effects –such as the Stroop effect– are often difficult to relate to reliable individual differences, irrespective of the specific individual difference construct of interest (Hedge et al., 2018; Rouder, Kumar, et al., 2019). That is, because these effects are very robust and automatic ("everybody Stroops"), the between-subjects variability is by definition relatively small. For correlational designs, this 'problem' of small between-subjects variability is further complicated by the presence of measurement error. Rouder, Kumar, et al. (2019) demonstrated that the ratio of true variability (i.e., true differences between individuals) to trial noise (i.e., measurement error) is 1 : 7. This unfavorable ratio renders the mission to uncover individual differences in cognitive tasks difficult, if not even impossible. Hierarchical models could mitigate these problems, as these models minimize the effect of trial noise by pulling the trial-level estimates toward the individual's mean effect (known as hierarchical shrinkage). In the current study, we did apply hierarchical modeling for the response time models, as well as the neural ACC models (incorporated in the first-level fMRI models in FSL and by adding the variance parameter of the beta's in the statistical models). Nevertheless, as acknowledged by Rouder, Kumar, et al. (2019), characterizing the degree of measurement error does not imply that the real underlying individual differences can be recovered. This casts doubt on the feasibility to detect true individual variation in cognitive control tasks, and hence to uncover associations with other measures. For example, Hedge et al. (2018) reported correlations of Stroop performance with other measures of cognitive control (e.g., Flanker task, Go/No-go task) ranging from −.14 to .14, none of which were significant. If we cannot even establish correlations between two tasks designed to measure exactly the *same* underlying phenomenon (i.e., cognitive control), the quest for reliable correlations between Stroop performance and more distant constructs such as religiosity seems all the more futile.

Although we obtained moderate to strong evidence for all null hypotheses related to religiosity and cognitive conflict, the current study does not imply that we should reject the notion of reduced conflict sensitivity as a defining characteristic of religious beliefs all together. It could well be that the relationship between religiosity and conflict sensitivity is restricted to specific instances or contexts and hinges strongly on the specific measures and operationalizations that are used. For example, in the study by Good et al. (2015) participants read a sermon about different qualities of God and then performed a Go/No-Go task with alcohol-related stimuli for which responses should be inhibited. As all participants refrained form alcohol consumption in their daily lives based on religious grounds, errors on the Go/No-Go task were seen as 'religious' errors, exposing participants' ostensible pro-alcohol tendencies. The results showed that emphasizing the loving and forgiving nature of God reduced the ERN amplitude in response to religious errors, while emphasizing divine punishment did not affect the ERN compared to a control condition. In other words, it could well be that when participants first contemplate on the comforting nature of their religious beliefs, this may reduce conflict-related ACC activity as induced by a task that includes religion-relevant items and responses. Such a task has much higher ecological validity than the Stroop task that we employed in the current study following the work by Inzlicht et al. (2009). Similarly, the observed reduction of activity in religious believers' DLPC and ACC while listening to a charismatic religious authority (Schjoedt et al., 2011), may specifically depend on the religious content of the speech (and may disappear when the same religious authority would talk about public transport or gardening). It is thus important to do justice to the subjective nature of religious practices and experiences, when studying these topics. This resonates with concerns about the lack of ecological validity in many neuroscience studies on religion (e.g., Schjoedt & van Elk, 2019): while studies such as the present one offer high experimental control, the measures do not capture the 'true stuff' that most psychologists and neuroscientists of religion are interested in, namely lived religious beliefs and experiences.

We see two broad future directions for the field. First, the development of new and sophisticated techniques in neuroscience could allow for interesting new hypotheses and measures. For instance, the use of multi-voxel pattern analysis (MVPA) may provide insight into the representational nature of religious concepts endorsed by believers; a question could be whether the neural representations of religious agents such as 'God', 'angels', or 'Satan' are more similar to real people such as 'Napoleon' and 'Donald Trump' or to imaginary agents such as 'Santa Claus' and 'Superman' (cf. Leshinskaya et al., 2017).

Novel methods for assessing brain connectivity also allow for the investigation of new questions (e.g., Huntenburg et al., 2018; Margulies et al., 2016). One could assess for instance the relationship between religiosity and the integration of information from sensory cortical areas and the default mode network (DMN), a network that is implicated in abstract, high-level thinking. A hypothesis could be that religious believers are more likely to show a dissociation between the DMN and primary sensory areas. This could be studied in a correlational resting-state design, or alternatively, one could assess believers' brain connectivity while engaging in contemplation of their (religious) beliefs or actions. For instance, intense personal prayer may be associated with a decoupling of internal self-referential processing in the DMN and perceptual processing in the sensory cortices specifically during the prayer experience, similar to what was found for shamanic trance-experiences (Hove et al., 2015).

Second, and relatedly, we believe there is much promise in future endeavours that focus on the application of paradigms and tasks that have higher ecological validity and more closely implicate religious concepts, as in the examples given above. Such an approach can hopefully do more justice to the multifaceted nature of religious beliefs and practices and can pave the way for a truly better understanding of the mechanisms and processes involved in religiosity.

## Appendix 5.A Population Imaging of Psychology project

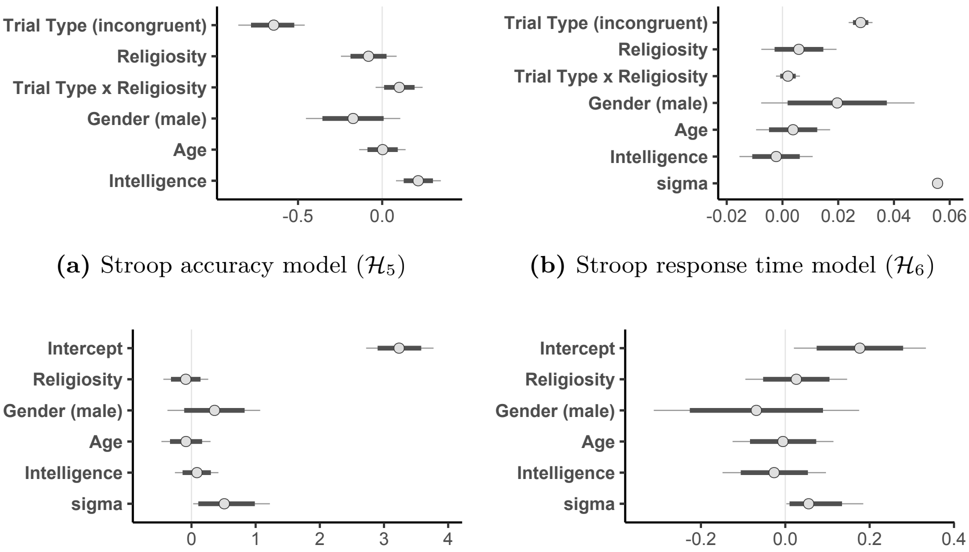
The data for this study were collected as part of the Population Imaging of Psychology project (PIoP), which was conducted at the Spinoza Center for Neuroimaging at the University of Amsterdam. The aim of the PIoP was to offer researchers the opportunity to collect brain-imaging data from a large sample of participants (intended $N = 250$), in association with their individual difference measure of interest. The data were collected between May 2015 and April 2016.

Standard measurements that were collected for each participant included a structural T1 MRI scan, task-free resting state fMRI (6 minutes), a diffusion tensor imaging (DTI) scan, and different functional localizer scans that were collected using EPI sequences, including the Gender Stroop task, an emotional matching task (Hariri et al., 2000), a working memory task (Pessoa et al., 2002), and the anticipation of positively vs. negatively valenced stimuli (Oosterwijk, 2017). In addition, demographic variables were recorded (gender, age, socio-economic status) for each participant, as well as two personality questionnaires, the NEO-FFI (Costa & McCrae, 1992) and the SCID (First et al., 1997), and an intelligence test (Raven's matrices; Raven, 2000). Finally, measures on religiosity and religious experiences were included (see Methods for details on the religiosity scale that was used in the present study).

## Appendix 5.B Additional Religiosity Items

1. To what extent do you consider yourself to be spiritual?
2. To what extent do you believe in paranormal phenomena (e.g., astrology or telepathy)?
3. To what extent are your parents religious?
4. To what extent do your parents frequently visit a church or religious meeting?
5. Do your parents have a religious lifestyle (e.g., don't go shopping on Sunday, pray before dinner)?

## Appendix 5.C Coefficient Plots



**(a)** Stroop accuracy model ($\mathcal{H}_5$)

**(b)** Stroop response time model ($\mathcal{H}_6$)

**(c)** Response conflict ACC model ($\mathcal{H}_7$)

**(d)** Informational conflict ACC model ($\mathcal{H}_8$)

**Figure 5.7:** Coefficients of the fixed effects on Stroop accuracy (top left panel), Stroop response times (top right panel), response conflict ACC activity (bottom left panel), and informational conflict ACC activity (bottom right panel), derived from the Bayesian regression models. For each predictor, points represent the median estimates, thick lines the 80% credible interval and thin lines the 95% credible interval. Note that predictors in the accuracy model are on a linear scale and should be transformed by the inverse logit link to reflect probabilities. In the accuracy and response time models, the intercepts are omitted to enhance visibility of the predictors.

# 6

# A Bayesian Multiverse Analysis of Many Labs 4: Quantifying the Evidence against Mortality Salience

Many Labs projects have become the gold standard for assessing the replicability of key findings in psychological science. The Many Labs 4 project recently failed to replicate the mortality salience effect where being reminded of one's own death strengthens the own cultural identity. Here, we provide a Bayesian reanalysis of Many Labs 4 using meta-analytic and hierarchical modelling approaches and model comparison with Bayes factors. In a multiverse analysis we assess the robustness of the results with varying data inclusion criteria and prior settings. Bayesian model comparison results largely converge to a common conclusion: We find evidence against a mortality salience effect across the majority of our analyses. Even when ignoring the Bayesian model comparison results we estimate overall effect sizes so small (between $d = 0.03$ and $d = 0.18$) that it renders the entire field of mortality salience studies as uninformative.

## 6.1 Introduction

Many Labs is a crowd-sourcing project that collects data from many different sites across the globe to answer questions about replicability and variability of effects, and it has become the gold standard for assessing the robustness of key findings in the psychological literature. Many Labs 4 (R. A. Klein et al., 2019), the most recent implementation of this idea, is a large scale attempt to replicate the Mortality Salience Effect from Terror Management Theory (Greenberg et al., 1994): reminders of one's own death strengthen one's cultural identity. In the classical demonstration of this effect, participants from the United States who were prompted to imagine their own death expressed more pro-American (i.e., in line with their worldview) beliefs than participants who were prompted to imagine watching TV. In addition to the question of replicability, R. A. Klein et al. (2019) wanted to assess the impact of involving

This chapter has been adapted from: Haaf, J. M., Hoogeveen, S., Berkhout, S., Gronau, Q. F., & Wagenmakers, E.-J. (2020). A Bayesian multiverse analysis of Many Labs 4: Quantifying the evidence against mortality salience. https://doi.org/10.31234/osf.io/cb9er.

the original authors in the study design. Therefore, some studies followed a standard protocol that was agreed upon by experts in the field (Author-Advised) while other studies were designed by the labs conducting them (In-House). After data collection from over 2,000 participants in 21 labs with and without involvement of the original authors the project could not replicate the original finding of Study 1 of Greenberg et al. (1994), and reported an overall effect size of $g = 0.03$, 95%CI $= [-0.06, 0.12]$.

Soon after the preprint of the Many Labs 4 project was posted, a critique of the analysis emerged. Chatard et al. (2020) pointed out that R. A. Klein et al. (2019) did not follow their own preregistered analysis. Chatard et al. (2020) argued that the preregistration specified a minimum of 40 participants per experimental cell as the threshold for sufficient power of any individual study, and therefore determined a total of 80 participants as target sample size for each lab. When reanalyzing the data from the Many Labs 4 project only including studies with 40 participants per condition Chatard et al. (2020) found a significant effect in line with the original results.

### 6.1.1 Include or Exclude?

Which of these analyses is the correct one? Based on theoretical arguments and (interpretations of) the preregistered plan, there may be several valid answers to this question, and several levels of exclusion criteria that ought to be considered. Both R. A. Klein et al. (2019) and Chatard et al. (2020) agreed on three *participant-level* exclusion criteria (the last two are suggested by the original authors – Greenberg, Pyszczynski, and Solomon –, who were consulted by the Many Labs 4 team):

1. Exclude participants who did not respond to all prompts of the dependent variable (leaving $N = 2211$).

2. In addition to exclusion criterion 1, participants who did not self-identify as white and/or who reported not to be born in the United States were also excluded (leaving $N = 637$).[1]

3. In addition to exclusion criteria 1 and 2, participants who responded below 7 on the 9-point American Identity item were also excluded (leaving $N = 277$).

In addition to these three participant-level exclusion criteria, power considerations motivated three different study-level exclusion criteria. We refer to these exclusion criteria as *N-based* criteria.

1. Include data from all labs (leaving $K = 21$ studies).

2. Exclude labs with fewer than 60 participants (leaving $K = 17$ studies).

3. Exclude labs with fewer than 40 participants per condition (leaving $K = 13$ studies).

[1] The argument is that the effect may only be present for participants who strongly identify with pro-American worldviews. We included participants who did identify as white in addition to another ethnicity, i.e., who are multiracial. We consider this the most appropriate interpretation of the preregistered ethnicity criterion.

Note that N-based exclusion criterion 2 was preregistered by R. A. Klein et al. (2019): "Samples will be included as long as they collect at least 60 participants by the time data collection ends" (see preregistration document, osf.io/4xx6w). In contrast, Chatard et al. (2020) derive exclusion criterion 3 from the *target* sample size specified in the preregistration document, although it is never mentioned as a criterion for exclusion. We decided to add both exclusion criteria for the sake of comparison.

Lastly, Greenberg et al. (1994) suggested that the effect may only emerge in Author-Advised studies as the mortality salience effect is highly sensitive to nuances in the study implementation. Therefore, the following distinction may constitute an additional set of *study-level* exclusion criteria. We refer to these exclusion criteria as *Protocol* criteria.

1. Include all studies (leaving $K = 21$).
2. Exclude all In-House studies (leaving $K = 9$).

These three levels of exclusion result in $3 \times 3 \times 2 = 18$ constellations of exclusion criteria. Table 6.1 shows all constellations, the resulting number of studies and total number of included participants. In the preprint, R. A. Klein et al. (2019) based their main conclusions on three of these constellations (blue rows): Including all studies, but varying the participant-level exclusion criteria.[2] Similarly, even though Chatard et al. (2020) conducted a variety of analyses in their comment, they based their key conclusions on three different constellations of criteria (pink rows): Excluding studies with fewer than 40 participants per condition, excluding In-House studies, with varying participant-level exclusion criteria.

In the following we will first report a reanalysis for the three exclusion constellations from R. A. Klein et al. (2019), and then for the three exclusion constellations from Chatard et al. (2020). Subsequently, lacking compelling argumentation for or against any of the criteria, we decided to conduct an analysis based on the entire set of 18 constellations as a multiverse analysis (Steegen et al., 2016). Note that some of the participant-level and study-level criteria are completely overlapping (e.g., only Author-Advised labs recorded American Identity, hence all In-House labs are excluded for the third participant-level exclusion set). As a result, there are 11 instead of 18 unique constellations (only rows that are not grey in Table 6.1).

### 6.1.2 A Bayesian Multiverse Reanalysis

We opt for a Bayesian analysis using Bayes factor model comparison (Jeffreys, 1939; Kass & Raftery, 1995). In short, Bayes factors quantify the relative evidence for a model (e.g., the alternative) over another model (e.g., the null). For an introduction to Bayes factor model comparison we refer the reader to Wagenmakers, Marsman, et al. (2018) and Rouder et al. (2018).

[2] We note that the eventual published article of Many Labs 4 may adopt different study-level criteria in order to adhere to the preregistration (i.e., exclude labs where $N < 60$). Furthermore, close examination of the preregistration document also revealed that some In-House labs had already started data collection prior to the registration and were therefore solely to "be included in clearly labelled supplemental and exploratory analyses". However, as these data were not accessed by the lead researchers and concerned In-House studies that were free to design their own protocols, we see no reason to exclude these observations.

**Table 6.1:** Exclusion constellations and resulting sample sizes

| Participant-level | N-based | Protocol | Sample Size | Number of Studies |
|---|---|---|---|---|
| All | All | All | 2,211 | 21 |
| White & US-born | All | All | 637 | 12 |
| US-Identity > 7 | All | All | 277 | 9 |
| All | All | AA | 799 | 9 |
| White & US-born | All | AA | 463 | 9 |
| US-Identity > 7 | All | AA | 277 | 9 |
| All | N > 60 | All | 2,053 | 17 |
| White & US-born | N > 60 | All | 549 | 9 |
| US-Identity > 7 | N > 60 | All | 229 | 7 |
| All | N > 60 | AA | 700 | 7 |
| White & US-born | N > 60 | AA | 386 | 7 |
| US-Identity > 7 | N > 60 | AA | 229 | 7 |
| All | N > 80 | All | 1,852 | 14 |
| White & US-born | N > 80 | All | 549 | 9 |
| US-Identity > 7 | N > 80 | All | 229 | 7 |
| All | N > 80 | AA | 700 | 7 |
| White & US-born | N > 80 | AA | 386 | 7 |
| US-Identity > 7 | N > 80 | AA | 229 | 7 |

*Note.* Blue rows refer to Klein et al.'s key analyses; pink rows refer to Chatard et al.'s key analyses; grey rows are repeated data sets and not included in the multiverse analysis; AA = Author-Advised.

The main advantage of Bayesian statistics in light of the current debate around the Many Labs 4 results is that it allows us to distinguish between evidence for the absence of the mortality salience effect and the absence of evidence for or against the effect. We decided to conduct two alternative analyses: Bayesian model-averaged meta-analysis (Gronau, van Erp, et al., 2017), and Bayesian hierarchical modeling (Haaf & Rouder, 2017; Rouder, Haaf, Davis-Stober, et al., 2019). The key distinction between these two approaches is that they operate on two levels of the data: For meta-analysis, the data from each lab are summarized with an effect size estimate and standard error, and these statistics are then analyzed using a linear model. For hierarchical modeling, the linear model is extended to the participant level, and participants' data are analyzed as nested within labs. Despite these differences, both analyses should provide comparable results. Subsequently, we briefly outline the two modeling approaches.

## 6.2 Methods

### 6.2.1 Bayesian Model-averaged Meta-analysis

Both classical and Bayesian meta-analysis typically consider four different models: (1) fixed-effect null model, (2) fixed-effect alternative model, (3) random-effects null model, and (4) random-effects alternative model. In Bayesian model comparison, we may now compute Bayes factors to compare any two of these models. Bayesian model averaging (e.g., Hinne et al., 2020) allows for broader inference when considering several models simultaneously. Using model averaging one can calculate the evidence for the presence of an effect while taking into account uncertainty with respect to choosing a specific model. For the application here, this logic implies that we can assess evidence for the mortality salience effect without committing to the fixed-effect or random-effects models.

Specifically, the model-averaged Bayes factor in favor of the presence of an effect is obtained by comparing the models that allow for the presence of an effect (i.e., (2) and (4) above) to the models that state the effect is absent (i.e., (1) and (3) above). In a similar fashion one can calculate the model-averaged Bayes factor in favor of the presence of between-study heterogeneity by comparing the models that allow for the presence of between-study heterogeneity (i.e., (3) and (4) above) to the models that state between-study heterogeneity is absent (i.e., (1) and (2) above).

We follow Gronau et al. (2021) for the specification of our Bayesian model-averaged meta-analysis. To conduct such an analysis, one needs to specify priors for the overall effect size across labs and the between-study standard deviation. For the between-study standard deviation we follow Gronau, van Erp, et al. (2017) and use an Inverse-Gamma(1, 0.15) prior. This prior is based on the empirical assessment of effect sizes from meta-analyses reported in *Psychological Bulletin* in the years 1990–2013 (van Erp et al., 2017). Van Erp et al. (2017) gathered all non-zero between-study standard deviation estimates for meta-analyses on standardized mean differences (e.g. Cohen's $d$), and the histogram approximately followed this distribution. For the overall effect size, we considered three different prior settings: (1) a zero-centered Cauchy distribution with scale $1\sqrt{2} \approx 0.707$ (*default* prior, Morey and Rouder, 2021), (2) a $t$-distribution with location 0.35, scale 0.102, and 3 degrees of freedom (*Oosterwijk* prior[3]), and (3) a normal distribution with mean 0.3 and standard deviation 0.15 (*Vohs* prior[4]). In line with the mortality salience hypothesis, all prior distributions on the overall effect size were truncated below at zero to allow only effect sizes in the expected direction. Readers interested in Bayesian model-averaging in meta-analysis may consult Gronau, van Erp, et al. (2017), Scheibehenne et al. (2017), and Landy et al. (2020).

### 6.2.2 Bayesian Hierarchical Modeling

For Bayesian hierarchical modeling we take advantage of the open availability of all collected data from the Many Labs 4 project. The dependent variable is the same across all studies, and participants are nested in studies resulting in a hierarchical data structure. We used the development by Rouder, Haaf, Davis-Stober, et al. (2019) with models similar to the ones used for the embodied cognition reanalysis by Rouder, Haaf, Davis-Stober, et al. (2019). There are four models under consideration: (1) The null model corresponds to the notion that none of the studies show an effect; this model is similar to the fixed-effect null model from the model-averaged meta-analysis. (2) The common-effect model corresponds to the notion that all studies show the same effect in the expected direction; this model is similar to the fixed-effect alternative model from the model-averaged meta-analysis. (3) The positive-effects model corresponds to the notion that all studies show an effect in the expected direction; and (4) the unconstrained model refers to the notion that the overall effect and study effects may vary freely; this model is similar to the random-effects alternative model from the model-averaged meta-analysis. We compute Bayes factors for models (2), (3), and (4) against model (1), the null model.

There are two critical prior settings to consider, the scale setting on the overall effect ($\mu_\theta$ in Rouder, Haaf, Davis-Stober, et al., 2019) and the scale setting on the between-lab heterogeneity ($\sigma_\theta^2$ in Rouder, Haaf, Davis-Stober, et al., 2019). The scale on the overall effect corresponds to the expected size of the overall effect. As Rouder, Haaf, Davis-Stober, et al. (2019), we set this scale to 0.4 since we expect a small-to-medium effect size. The scale of the between-lab variance corresponds to the expected amount of variability in effect size across studies. Again, we kept the value of 0.24 as proposed by Rouder, Haaf, Davis-Stober, et al. (2019).

### 6.2.3 Preregistration and Approach

With these two approaches we are now ready to reanalyse the Many Labs 4 data. Subsequently, we report the results of the Bayesian reanalysis of the key findings reported by R. A. Klein et al. (2019), and the results of the Bayesian reanalysis of the key findings by Chatard et al. (2020). Finally, we provide the results of the multiverse analysis across all possible exclusion criteria.

The analyses, including prior settings, were preregistered on the Open Science Framework (osf.io/ae4wx). However, we decided to deviate from the preregistration by including more constellations of exclusion criteria. Specifically, we originally

[3] This *Oosterwijk* prior has been elicited for a reanalysis of a social psychology study (Gronau et al., 2019), but we believe it is a reasonable prior for many psychological studies more generally.

[4] This *Vohs* prior has been specified by ego depletion experts to analyze ego depletion replication studies (Vohs et al., 2021).

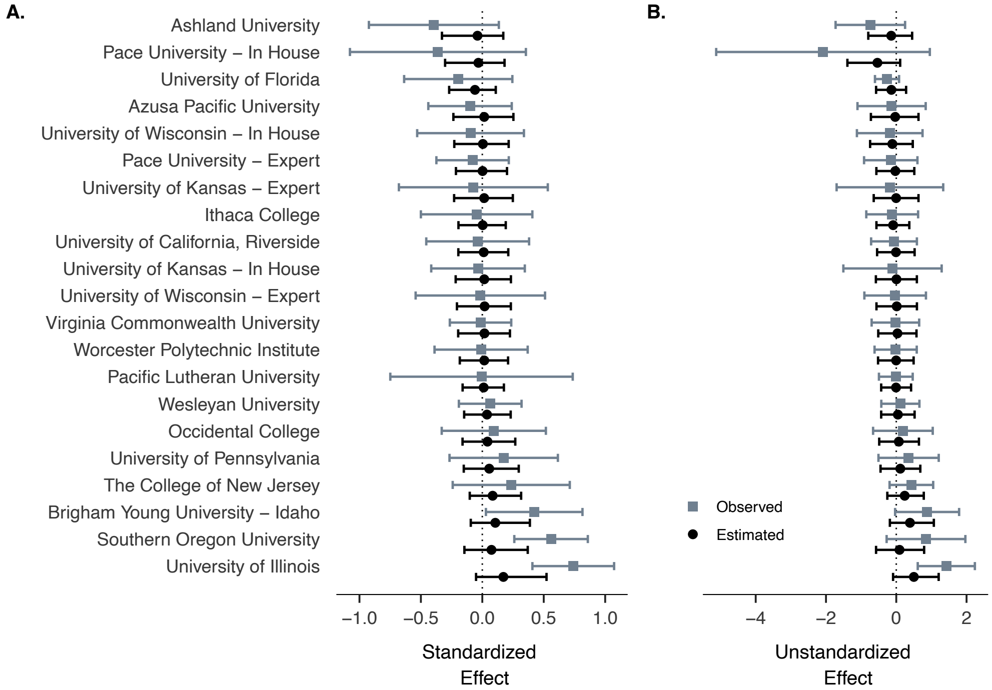


**Figure 6.1:** Forest plot with Bayesian parameter estimates for participant-level exclusion set 1 and no further study-level exclusions. **A.** Bayesian meta-analysis (with two-sided default prior). The grey points represent calculated effect sizes with 95% confidence intervals, the black points represent estimated effect sizes from the random-effects alternative model with 95% credible intervals. **B.** Bayesian hierarchical analysis. The grey points represent unstandardized observed effects for each study with 95% confidence intervals. The black points represent estimated unstandardized effects from the unconstrained model with 95% credible intervals.

planned to only use participant-level exclusion criterion 1 and later decided to include all of them. We believe the changes help to provide a more complete analysis.

The Bayesian model-averaged meta-analyses are conducted using the R-package `metaBMA` (Heck & Gronau, 2017). The Bayesian hierarchical modeling is conducted using the R-package `BayesFactor` (Morey & Rouder, 2021). All R-code is provided at github.com/jstbcs/ml4-reanalysis.

## 6.3 Bayesian Reanalysis of Klein et al.'s Key Findings

### 6.3.1 Model-averaged Meta-analysis of Klein et al.

Figure 6.1A, shows the observed effect size estimates for the first participant-level exclusion criteria without applying any study-level exclusion criteria. The observed effect sizes from each study (grey points) are plotted in increasing order, and the grey bars show the 95% confidence intervals for the effect size estimates. A quick first

**Table 6.2:** Model-averaged Bayes factors for key analyses.

| Inclusion Criteria | | | | Effect $\text{BF}_{01}$ | | | Heterogeneity $\text{BF}_{01}$ |
|---|---|---|---|---|---|---|---|
| Participant-level | N-based | Protocol | Labs | Default | Oosterwijk | Vohs | Default |
| Klein et al. (2019) | | | | | | | |
| All | All | All | 21 | 12.60 | 44.69 | 16.64 | 2.28 |
| White & US-born | All | All | 12 | 7.95 | 16.84 | 7.20 | 2.42 |
| US-identity $> 7$ | All | All | 9 | 4.18 | 4.01 | 2.49 | 1.79 |
| Chatard et al. (2020) | | | | | | | |
| All | N $> 40$ | AA | 7 | 3.82 | 5.84 | 2.75 | 2.54 |
| White & US-born | N $> 40$ | AA | 7 | 1.42 | 0.90 | 0.66 | 2.08 |
| US-identity $> 7$ | N $> 40$ | AA | 7 | 1.45 | 0.73 | 0.62 | 1.89 |

*Note.* All Bayes factors are reported in favor of the null model. AA = Author-Advised.

assessment of Figure 6.1A shows that the confidence intervals of observed effect sizes from 18 of the 21 studies cover zero. The black points in the figure refers to estimated effect sizes from a meta-analytic random-effects alternative model with a two-sided default priors. This model takes the observed study-level variability of effect sizes into account, and therefore estimates less variability of true study effects than the observed effect sizes. For the individual studies, the credible intervals of all estimated effect sizes for all three analyses cover zero.

In order to estimate the overall effect size across studies (Hedges' $g$) we used the same model as was used to estimate the individual-study effects (i.e., a random-effects alternative model with the default prior). For the full sample (participant-level exclusion criterion 1) the overall effect size is estimated as $0.03,\ 95\%\text{CI} = [-0.07, 0.13]$; for participant-level exclusion criterion 2 the overall effect size is estimated as $0.03,\ 95\%\text{CI} = [-0.14, 0.21]$; and for participant-level exclusion criterion 3 the overall effect size is estimated as $0.07,\ 95\%\text{CI} = [-0.21, 0.33]$. The most consistent pattern is that the credible interval widens when the exclusion criterion becomes more restrictive. Overall, these estimates are more consistent with the absence of an effect rather than its presence.

To quantify the absence or presence of an effect we now turn to Bayes factor model comparison. The Bayes factors for the key analyses from R. A. Klein et al. (2019) are shown in the top three rows of Table 6.2. Note that not all studies are included for exclusion criterion 2 because data on ethnicity and country of birth were only collected for some of the labs. Likewise, the American identity was only assessed in the Author-Advised studies, and therefore exclusion criterion 3 leads to the inclusion of only nine studies. The first three Bayes factors in each row are model-averaged Bayes factors referring to evidence against an overall effect. All analyses across participant-level exclusions and prior choices provide evidence against an overall effect, with Bayes factors ranging from 44.69-to-1 to 2.49-to-1 in favor of the null model. Note that the Oosterwijk prior is the most optimistic prior with the least probability density close to zero. Therefore, the Bayes factors are somewhat larger for this prior—the optimistic predictions that follow from the Oosterwijk prior are least consistent with what the data show, which are effect sizes close to zero. The last Bayes factor in each row indicates evidence against heterogeneity of study effects averaged across models with and without an overall effect. These Bayes factors reflect that there is some evidence against study heterogeneity. In sum, the pattern of Bayes factors indicates evidence against an overall mortality salience effect across the three prior settings and the three data sets. These results are in line with the estimation results in Figure 6.1A, and with the overall effect size estimates from a two-sided model.

### 6.3.2 Hierarchical analysis of Klein et al.

Figure 6.1B shows the observed, unstandardized effects and the estimates from the unconstrained multilevel model for the first participant-level exclusion criteria. As can be seen, there is considerable hierarchical shrinkage reducing the variability of estimated effects as compared to observed effects. Effect size estimates from the unconstrained model (similar to Cohen's *d*) are 0.01, 95%CI = $[-0.11, 0.12]$ for participant-level exclusion criterion 1, 0.02, 95%CI = $[-0.17, 0.21]$ for exclusion criterion 2, and 0.05, 95%CI = $[-0.22, 0.32]$ for exclusion criterion 3. Note that posterior means are close to zero, and that all credible intervals cover zero. The estimates are therefore consistent with the absence of an overall effect.

Bayes factors are shown in the first three rows of Table 6.3. $\text{BF}_{0f}$ refers to the Bayes factor between the null model and the unconstrained model; $\text{BF}_{01}$ refers to the Bayes factor between the null model and the common-effect model where the overall effect is positive and there is no variability between study effects; and $\text{BF}_{0+}$ refers to the Bayes factor between the null model and the positive-effects model where study effects may vary but all are consistently positive. All Bayes factors are in comparison to the preferred model, the null model, indicating evidence that none of the studies show an effect. The second best model is the common-effect model where all studies have the same, positive effect, and the Bayes factor between the null model and the common-effect model is between 10.34-to-1 to 2.11-to-1 in favor of the null model depending on the different participant-level exclusion criteria. In sum, this pattern indicates evidence against an overall mortality salience effect (null model), and even if there was an effect (common-effect model) there is no evidence for variability of study effects. These results are consistent across the three data sets, and they are in line with the estimation results shown in Figure 6.1B.

### 6.3.3 Summary of the Reanalysis for Klein et al.

Across both analyses, the meta-analytic approach using Bayesian model-averaging, and the hierarchical approach using participant-level data, we find no evidence for the mortality salience effect. The results are consistent across participant-level exclusion criteria and prior settings. Even though the evidence against an effect is more pronounced when all participants are included in the analysis, this pattern is easily explained by the resolution of the analysis with increasing numbers of observations: The fewer observations, the less evidence in any direction, and the wider the estimated posterior distribution of the overall effect.

**Table 6.3:** Bayes factors for key analyses.

| Inclusion Criteria | | | | | | |
|---|---|---|---|---|---|---|
| Participant-level | N-based | Protocol | Sample Size | $\mathrm{BF}_{0f}$ | $\mathrm{BF}_{01}$ | $\mathrm{BF}_{0+}$ |
| Klein et al. (2019) | | | | | | |
| All | All | All | 2,211 | 33.26 | 10.34 | 8,787.94 |
| White & US-born | All | All | 637 | 19.65 | 5.67 | 123.12 |
| US-Identity > 7 | All | All | 277 | 9.24 | 2.11 | 13.43 |
| Chatard et al. (2020) | | | | | | |
| All | N > 40 | AA | 700 | 13.64 | 2.07 | 11.96 |
| White & US-born | N > 40 | AA | 386 | 6.87 | 0.94 | 2.73 |
| US-identity > 7 | N > 40 | AA | 229 | 4.83 | 0.83 | 1.57 |

*Note.* All Bayes factors are reported in favor of the null model. AA = Author-Advised.

## 6.4 Bayesian Reanalysis of Chatard et al.'s Key Findings

### 6.4.1 Model-averaged Meta-analysis of Chatard et al.

For the reanalysis of the key findings of Chatard et al. (2020) we provide a forest plot of the most exclusive criteria–participant criterion 3, and only author-advised studies with more than 40 participants per cell included–in Figure 6.2. Together with Figure 6.1 Figure 6.2 illustrates the range of included study effects from the most liberal to the most restrictive criteria. Figure 6.2A again shows the results from a meta-analytic random effects model with unconstrained overall effect. Note that all confidence intervals (grey bars) and all credible intervals (black bars) include zero.

We estimated the overall effect size across studies (Hedges' $g$) using the settings from the default prior without constraining the direction of the overall effect. We did so for all data sets using the three participant-level exclusion criteria, only studies that had more than 40 participants per cell collected, and only Author-Advised studies. For participant-level exclusion criterion 1 the overall effect size is estimated as $0.08,\ 95\%\text{CI} = [-0.09, 0.25]$; for exclusion criterion 2 the overall effect size is estimated as $0.16,\ 95\%\text{CI} = [-0.07, 0.40]$; and for exclusion criterion 3 the overall effect size is estimated as $0.18,\ 95\%\text{CI} = [-0.10, 0.47]$. While the point estimates are considerably larger than the ones when all studies are included, the posterior distributions and therefore also the credible intervals are considerably wider due to much smaller sample sizes. In this analysis, only seven studies were included, and only between 700 and 229 participants.

To quantify the absence or presence of an effect we again computed model-averaged Bayes factors. These are shown in the bottom three rows of Table 6.2. The first three Bayes factors in each row are model-averaged Bayes factors referring to evidence against an overall effect using different prior distributions. Here, the pattern is a bit more inconsistent than in the Klein et al. reanalysis, and the outcome depends on a combination of the prior settings and exclusion criteria: Bayes factors (weakly) favor the absence of an effect over its presence for all priors if participant-level exclusion criterion 1 is applied. For the smaller data sets using criteria 2 or 3, the Bayes factors

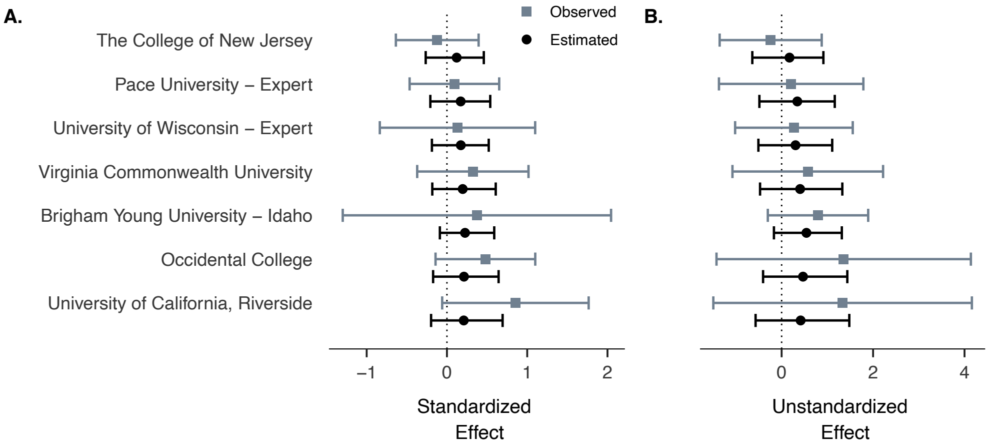


**Figure 6.2:** Forest plot with Bayesian parameter estimates for participant-level exclusion set 3 for studies with more than 40 participants per cell and only author-advised studies included. **A.** Bayesian meta-analysis (with two-sided default prior). The grey points represent calculated effect sizes with 95% confidence intervals, the black points represent estimated effect sizes from the random-effects alternative model with 95% credible intervals. **B.** Bayesian hierarchical analysis. The grey points represent unstandardized observed effects for each study with 95% confidence intervals. The black points represent estimated unstandardized effects from the unconstrained model with 95% credible intervals.

are essentially inconclusive – for the default prior the Bayes factors are still in favor of the null hypothesis but close to 1. For the other two prior settings the Bayes factors are in favor of the presence of an effect but, again, close to 1. The largest Bayes factor in favor of the presence of an effect is with the Vohs prior, and participant-level exclusion setting 3.

The last column in Table 6.2 shows the model-averaged Bayes factor quantifying evidence against heterogeneity of effect sizes across labs. Again, there is weak evidence against heterogeneity. In sum, this pattern is in line with the absence of evidence for or against an overall mortality salience effect.

### 6.4.2 Hierarchical Analysis of Chatard et al.

We also reanalyzed Chatard et al.'s findings with a hierarchical modeling approach. Figure 6.2B shows study estimates from the unconstrained model for the unstandardized effects. As with the standardized effects in panel A, all confidence intervals and credible intervals cover zero.

Effect size estimates from the unconstrained model (similar to Cohen's $d$) of 0.08, 95%CI $= [-0.12, 0.29]$ for participant-level exclusion criterion 1, 0.14, 95%CI $=$ $[-0.11, 0.37]$ for participant-level exclusion criterion 2 and 0.18, 95%CI $= [-0.11, 0.48]$. Note that all credible intervals include zero, and even though the posterior mean increases with more conservative exclusion criteria the width credible interval increases as well implying increasing uncertainty about the effect size. The posterior distribution of the overall effect size is therefore again consistent with the absence of an overall effect.

The pattern of Bayes factors is somewhat less consistent than the estimation results across exclusions. Bayes factors are shown in the last three rows of Table 6.3. The pattern of Bayes factors is, as with the model-averaged analysis, dependent on the participant-level exclusion criterion. Under participant-level exclusion criterion 1 the preferred model is the null model, and it is weakly preferred over the second-best model, the common effect model, by a Bayes factor of $\text{BF}_{01} = 2.07$. For the other two exclusion criteria, the common-effect is preferred over the null model but the Bayes factors are even weaker (1.06 and 1.24 in over the null model). In sum, the pattern for the different data exclusions is in line with the conclusions from the model-averaged analysis: The Bayes factors show the absence of any consistent evidence for or against an effect.

### 6.4.3 Summary of the Reanalysis of Chatard et al.

Both the model-averaged meta-analysis and the hierarchical modeling approach show a similar pattern: Across the three participant-level exclusion criteria and different prior settings, there is only weak and inconsistent evidence for or against an overall mortality salience effect. Here, we advice readers not to overly interpret whether the Bayes factor is 1.5-to-1 for or against the overall effect – none of these Bayes factors are convincing. Instead, all of the analyses in this section point to the conclusion that more data are needed. The exclusion criteria applied here thinned out the data so much – in the final analytic data set only 10% of the initial data is retained – so that no firm conclusion is possible anymore.

## 6.5 Bayesian Multiverse Analysis

To assess the robustness of the previously reported results we conducted a multiverse analysis using the eleven data sets from Table 6.1 (i.e., all rows that are not grey). We conducted a model-averaged meta-analysis and report here the Bayes factors for the presence of an effect against its absence. The analysis is conducted using the three different prior distributions, the default prior, the Oosterwijk prior, and the Vohs prior. The Bayes factors are plotted in Figure 6.3 ($y$-axis). Bayes factors in favor of the mortality salience effect are above the horizontal line, and Bayes factors against the mortality salience effect are below the horizontal line. The $x$-axis refers to the number of participants whose data are included in the analysis. The size of the point reflects the number of studies included in the analysis. The majority of Bayes factors are in line with the absence of the mortality salience effect. Because the Bayes factor depends on the sample size, more evidence against morality salience comes from analyses that are based on more data (i.e., larger number of included participants and studies). Only two constellations of exclusion criteria provide evidence for the mortality salience effect.

To inspect the effect of prior settings one can view the points in Figure 6.3 that are in the same $x$-axis location. Remember that the default prior is the most vague prior and the Oosterwijk prior is more optimistic than the Vohs prior. For the three data

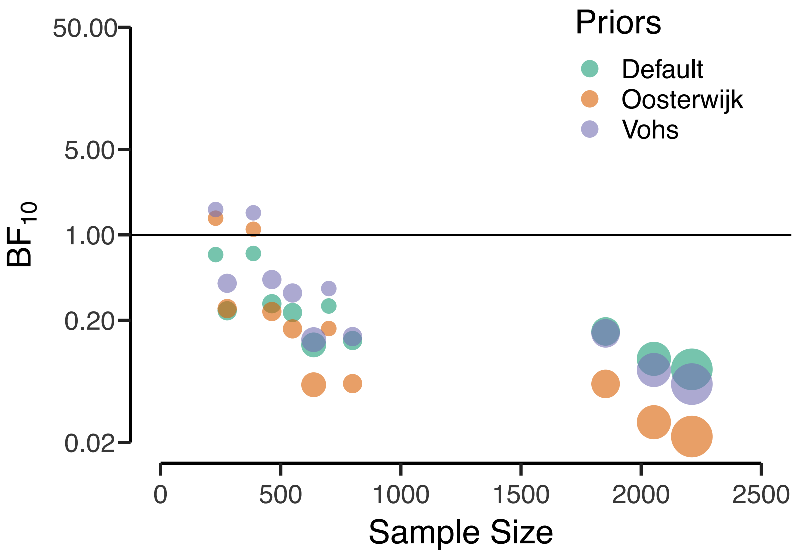


**Figure 6.3:** Results from the multiverse analysis: Bayes factors in favor of a mortality salience effect are above the horizontal line, Bayes factors against the mortality salience effect are below the horizontal line. The color of the points refers to the different priors on the overall effect, the size of the points refers to the number of studies included in the analysis, and the $x$-axis refers to the number of participants the analysis is based on. The majority of analyses provide evidence against the mortality salience effect.

sets with the largest numbers of participants Bayes factors are larger for more optimistic priors because evidence against optimistic and informed models accumulates faster when comparing to a null model. The same logic applies for situations where data are more ambiguous. The smallest data sets show a small positive overall effect, and evidence for this small effect accumulates faster with more optimistic priors than less optimistic ones. Therefore, the Bayes factors are only greater than one (i.e., in favor of an effect) for the Vohs and the Oosterwijk prior. Because the Vohs prior has more density for smaller effect sizes than the Oosterwijk prior, the Bayes factor favors an overall effect most for the Vohs prior.

### 6.5.1 Summary of the Multiverse Analysis

The evidence against the morality salience effect appears fairly robust against choices of exclusion criteria and priors. When conducting a large number of analyses on the same data some of these analyses will almost inevitably lead to some evidence in the opposite direction than the overall results. This is especially the case when the data provide relatively weak evidence (Bayes factors less than 5-to-1 against an effect). Bayes factors close to 1 may signal a lack of resolution of the data and therefore the absence of evidence for or against an effect. When the number of participants is high and many studies are included there is convincing evidence against the mortality salience effect. The four Bayes factors that are weakly in favor of the mortality salience effect are based on two of the smallest data sets and the two more informative prior settings.

## 6.6 Conclusion

We conducted a Bayesian reanalysis of the Many Labs 4 project with varying exclusion criteria, priors, and model choices. In a Bayesian multiverse analysis we calculated a total of 33 model-averaged Bayes factors based on three different prior settings and 11 different data sets resulting from different data exclusion criteria derived from the Many Labs 4 preregistration (R. A. Klein et al., 2019). 29 of the 33 Bayes factors provide evidence against an overall mortality salience effect, ranging between 1.42-to-1 and 44.69-to-1 in favor of the absence of an effect. The remaining four Bayes factors provide only weak evidence for the presence of such an effect, ranging between 1.11-to-1 and 1.61-to-1 in favor of the presence of an effect. Additionally, we do not find evidence for heterogeneity of effects across studies. Even if we do not believe the evidence across 33 Bayesian model comparisons and assume there is an effect, this effect is so small (between $d = 0.03$ and $d = 0.18$) that it renders the entire field of mortality salience studies as uninformative: Most of the studies conducted in the past would have been vastly underpowered, and would require a very specific subgroup of participants.

Our analyses revealed that the evidence is relatively consistent across different exclusion criteria. For the current analysis, we assumed that all exclusion criteria are equally plausible. With this assumption we implicitly assigned an equal weight to all analyses. However, we admit that this may not be the case. Chatard et al. (2020) argue that their chosen criteria are superior when considering theoretical arguments and study planning. With their analysis, they implicitly introduced a weighing where all other exclusion options received a weight of zero. Readers can choose these weights themselves when they consider how to interpret the results reported here.

There are additional issues with selectively subsetting and reanalyzing data sets. A key danger is that for some subsets one always finds results opposite of the conclusions from the analysis of the full data set. On the study level, researchers should therefore first ensure that there is evidence for variability of studies that warrants such subsetting. In the current analysis, we found evidence against study heterogeneity. When interpreting the results we therefore recommend to rely mainly on the estimates from the full data set. Additionally, subsetting the data inevitably reduces the resolution to detect an effect. The critics of the Many Labs 4 project (Chatard et al., 2020) based their main conclusions on analyses with smaller sample sizes. Ironically, while Chatard et al. (2020) argued that sample size should be considered when including studies their exclusion criteria actually reduced the power of the meta-analysis. To tackle this issue—and if there was evidence for study heterogeneity—one could include some of the subsetting criteria as predictor in the meta-analytic model (e.g. author-advised vs. in-house).

In summary, the multiverse analysis conducted here shows a certain convergence of results. Even though the degree of evidence varies, models with no effect of mortality salience are mostly preferred over models with an effect of mortality salience. This result highlights the robustness against choices of priors and exclusion criteria.

The Bayesian multiverse approach provides rich results that go much beyond the original analyses by the Many Labs 4 team. Moreover, multiverse analyses can be executed easily, for example using JASP (JASP Team, 2019). The current analyses were conducted in R, and the code is provided at github.com/jstbcs/ml4-reanalysis. The ease and informativeness of multiverse analyses show that this approach should be more generally used to analyze large-scale studies. The Many Labs idea is that the robustness of empirical phenomena becomes clear when data are collected across several labs. Similarly, the robustness of statistical conclusions becomes clear when data are analyzed using several thoughtfully selected models. A complete assessment of robustness and uncertainty therefore requires both many labs and many models.

# Part III

# The Cross-Cultural Religious Replication Project

# 7

# The Einstein Effect Provides Global Evidence for Scientific Source Credibility Effects and the Influence of Religiosity

People tend to evaluate information from reliable sources more favourably, but it is unclear exactly how perceivers' worldviews interact with this source credibility effect. In a large and diverse cross-cultural sample ($N$ = 10,195 from 24 countries), we presented participants with obscure, meaningless statements attributed to either a spiritual guru or a scientist. We found a robust global source credibility effect for scientific authorities, which we dub 'the Einstein effect': across all 24 countries and all levels of religiosity, scientists held greater authority than spiritual gurus. Additionally, individual religiosity predicted a weaker relative preference for the statement from the scientist vs. the spiritual guru, and was more strongly associated with credibility judgments for the guru than the scientist. Independent data on explicit trust ratings across 143 countries mirrored our experimental findings. These findings suggest that irrespective of one's religious worldview, across cultures science is a powerful and universal heuristic that signals the reliability of information.

## 7.1 Introduction

In a heated debate about the proximity of COVID-19 herd immunity, White House health advisor Dr. Scott Atlas proclaimed "You're supposed to believe the science, and I'm telling you the science" (The White House Press Briefing, 2020). A group of infectious disease experts and former colleagues from Stanford, however, publicly criticized Dr. Atlas, who is a radiologist, for spreading 'falsehoods and misrepresentation of science' through his statements about face masks, social distancing and the

This chapter has been adapted from: Hoogeveen, S., Haaf, J. M., Bulbulia, J. A., Ross, R. M., McKay, R., Altay, S., Bendixen, T., Berniūnas, R., Cheshin, A., Gentili, C., Georgescu, R., Gervais, W. M., Hagel, K., Kavanagh, C. M., Levy, N., Neely, A., Qiu, L., Rabelo, A., Ramsay, J. E., ... van Elk, M. (2022). The Einstein effect provides global evidence for scientific source credibility effects and the influence of religiosity. *Nature Human Behaviour*, *6*(4), 523–535. https://doi.org/10.1038/s41562-021-01273-8.

safety of community transmission (Farr, 2020). In the 2020 pandemic crisis, all eyes turned to scientific experts to provide advice, guidelines and remedies; from COVID-19 alarmists to skeptics, appeal to scientific authority appeared a prevalent strategy on both sides of the political spectrum.[1]

A large body of research has shown that the credibility of a statement is heavily influenced by the perceived credibility of its source (Brinol & Petty, 2009; Chaiken & Maheswaran, 1994; A. J. Harris et al., 2016; McGinnies & Ward, 1980; Petty & Cacioppo, 1986; Pornpitakpan, 2004; C. T. Smith et al., 2013; Sperber et al., 2010). Children and adults are sensitive to the past track record of informants (Birch et al., 2010; Chudek et al., 2012; Clément et al., 2004; P. L. Harris et al., 2018; Jaswal & Neely, 2006; Taylor et al., 1991), evidence of their benevolence toward the recipient of testimony (Fiske & Dupree, 2014; Johnston et al., 2015; Mascaro & Sperber, 2009), as well as how credible the information is on its face (Bloom & Weisberg, 2007; P. L. Harris, 2012). From an evolutionary perspective, deference to credible authorities such as teachers, doctors, and scientists is an adaptive strategy that enables effective cultural learning and knowledge transmission (Hahn et al., 2016; Henrich, 2015; Henrich & Gil-White, 2001; D. D. Johnson, 2020; D. D. Johnson & Fowler, 2011; Mercier, 2020; Sperber, 1997). Indeed, if the source is considered a trusted expert, people are willing to believe claims from that source without fully understanding them. We dub this 'the Einstein effect'; people simply accept that $E = mc^2$ and that antibiotics can help cure pneumonia because credible authorities such as Einstein and their doctor say so, without actually understanding what these statements truly entail.

Knowing that a statement originates from an epistemic authority may thus increase the likelihood of opaque messages being interpreted as meaningful and profound. According to Sperber (2010), in some cases, incomprehensible statements from credible sources may be appreciated not just in spite of but *by virtue of* their incomprehensibility, as exemplified by the speech of spiritual or intellectual gurus (the "Guru effect"). Here, we investigate to what extent different epistemic authorities affect the perceived value of nonsensical information. To this end, we contrasted judgements of gobbledegook spoken by a spiritual leader with gobbledegook spoken by a scientist. In addition, we assessed whether the source effect is predicted by individual religiosity and varies cross-culturally, as a proxy for how scientists and spiritual authorities function as "gurus" for different individuals and within different cultural contexts.

Although source credibility effects have typically been investigated for persuasion in marketing and communication, both science and spirituality may present particularly suitable contexts for inducing strong source effects. Scientists are generally considered competent and benevolent sources (Funk, 2020; Krause et al., 2019) and scientific information is often difficult and counterintuitive (McCloskey et al., 1983; Reynolds et al., 2010; E. U. Weber & Stern, 2011). The combination of a credible authority and intangible information can increase the probability of obscure scientific information being accepted, by enhancing perceivers' reliance on the source (Chaiken & Maheswaran, 1994; Mercier, 2016; Petty & Cacioppo, 1986). Even indirect context cues, such as those emphasizing the scientific nature of a piece of information can increase the probability that (dubious) information is believed (A. M. Evans et al.,

[1] Please see the Appendix 7.C for a short commentary on how the present work might relate to the COVID-19 situation.

2020). Some experimental evidence, for instance, suggests that irrelevant neuroscience information (Fernandez-Duque et al., 2014; McCabe & Castel, 2008; Weisberg et al., 2008) or nonsense mathematical equations (Eriksson, 2012) can boost the perceived quality of presented claims, though note that replication studies suggest that mere brain images may not suffice (Gruber & Dickerson, 2012; Michael et al., 2013). Notably, these effects were only present among nonexperts (i.e., people with little formal neuroscientific or mathematical training). This distinction suggests that the appeal of "sciencey" information may be particularly strong when analytical assessment fails and one can only rely on secondary credibility cues.

Similar to the anticipated complexity of scientific information, prior beliefs about religious or spiritual texts instigate expectations that the information presented will be obscure. Supernatural explanations often appeal to phenomena that operate outside of the natural world and to experiences deemed ineffable, mysterious and exempt from empirical validation (Boyer, 2001; Friesen et al., 2015; K. A. Johnson et al., 2019; Legare et al., 2012; Liquin et al., 2020; Van Leeuwen, 2014). Some scholars have argued that incomprehensible theological language and irrational beliefs may serve as a costly signal towards the religious ingroup, signalling quality by hard-to-fake moral commitment, intellectual capacity and epistemological investment (Irons, 2008; Mahoney, 2008). However, irrespective of content biases, the evaluation of spiritual or theological obscurity critically depends on one's personal beliefs about the credibility of spiritual gurus or religious authorities.

Various lines of evidence suggest that perceived credibility of both content and source indeed depends on individual difference factors such as the perceiver's (political) ideology and worldview (Brandt & Crawford, 2020; Gauchat, 2011, 2012; Lachapelle et al., 2014). In the absence of the means to rationally evaluate a claim and reliable source information, people likely infer credibility based on beliefs about the group to which the source belongs (e.g., 'conservatives', 'scientists'). In this process, similarities between one's own worldview and that of the source's group may serve as a proxy for being a benevolent and reliable source (Hahn et al., 2016; Levy, 2019). In a religious context, Christians were found to be more affected by an intercessory prayer when supposedly performed by a (charismatic) Christian than a non-Christian (Schjoedt et al., 2011) and to require less evidence for religious claims (e.g., efficacy of prayer to cure illness) than for scientific claims (e.g., efficacy of medication; Lobato et al., 2019; McPhetres and Zuckerman, 2017). These differences were not present among secular individuals. Furthermore, evangelical Christians were more likely to accept statements opposing their personal views when attributed to an ingroup religious leader versus an outgroup religious leader (Robinson, 2010). This effect was moderated by the amount of contact participants had with the specific group the religious leader belonged to, which highlights the importance of the person-source fit for message acceptance.

To account for these effects, alongside traditional dual-process models of persuasion (Chaiken & Maheswaran, 1994; Munro & Ditto, 1997; Petty & Cacioppo, 1986; Tversky & Kahneman, 1974), various authors have recently proposed a Bayesian framework in which subjective beliefs about the source (e.g., trustworthiness) and one's worldviews contribute to belief updating in response to new information following Bayesian principles (Cook & Lewandowsky, 2016; Hahn et al., 2009; A. J. Harris et al., 2016; Jern et al., 2014). By including background beliefs, these Bayesian networks describe how a differential weighing of evidence and even divergent updating (belief polarization) can be considered rational and normative. This may explain, for instance, how strong religious believers can become more convinced of their beliefs in the face of disconfirmatory evidence, especially when their faith is being challenged (Batson, 1975; Jern et al., 2014). Similarly, strong conservatives who distrust science may become less convinced of human-caused global warming when presented with scientific consensus information (Cook & Lewandowsky, 2016). In other words, laypeople may apply their own 'power priors' (Ibrahim & Chen, 2000) to calibrate evidence from different sources, whose trustworthiness is subjectively determined, partly by their broader worldview.

In sum, whereas previous studies have established source credibility effects in a wide array of domains, as-of-yet little is known about whether and to what extent people's worldview is predictive of the relative credibility evaluation of information from scientific and spiritual sources. In the present study, we presented participants ($N = 10,195$, from 24 countries) with meaningless verbiage (henceforth, "gobbledegook"; also referred to in the literature as "pseudo-profound bullshit"; Pennycook et al., 2015) randomly credited to either a spiritual authority or a scientific authority. We assessed (1) whether trusting scientific experts over spiritual leaders is a general heuristic (i.e., the Einstein effect), and (2) to what extent perceivers' religiosity predicts the relative confidence in the truth of the gobbledegook statements from both sources. Note that we chose a "spiritual guru" authority frame, instead of "religious leader," because we wanted to avoid selecting an authority specific to any particular religion, to keep the study consistent across countries. While religiosity and spirituality are overlapping but not interchangeable constructs (Paloutzian & Park, 2014; Zinnbauer et al., 1997), self-reported religiosity has been positively associated with belief in spiritual phenomena such as fate, spiritual energy, and a connected universe (Lindeman et al., 2019; McClintock et al., 2016; M. S. Wilson et al., 2013, though not unequivocally; Rice, 2003). Consequently, we expected religiosity to be associated with increased receptivity to gobbledegook from a spiritual authority.

All confirmatory hypotheses and included measures were preregistered on the Open Science Framework (see osf.io/faj2z/). This link contains the original preregistration file. The registered component (including additional sub-projects) can be found at osf.io/xg8y5/files. In addition, for exploratory purposes, we included response time measures and a memory test to obtain insight into the cognitive processes underlying the source credibility effect (these measures were anticipated in the preregistration, but no concrete hypotheses were formulated). In order to further validate the findings from our experimental paradigm, we also analysed a large dataset from 117,191 individuals across 143 countries (including the same countries included in our study) that contains explicit trust ratings of scientists and traditional healers, as well as participant religiosity (Gallup, 2019).

## 7.2 Methods

### 7.2.1 Participants

In total, 10,535 participants completed the online experiment. Of these, 340 participants (3.23%) were excluded because they failed the attention check (but see Table 7.3 for equivalent results when data all participants are included), leaving an analytic sample of $N = 10,195$ from 24 countries (see Table 7.1 for descriptive statistics per country). Participants were recruited from university student samples, from personal networks, and from 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 a financial remuneration, the possibility for a reward through a raffle, course credits, or no compensation. There were no a priori exclusion criteria; everyone over 18 years old could participate. Participants were forced to answer all multiple choice questions, hence there was no missing data (except for 36 people who did not provide a valid age). The countries were convenience-sampled (i.e., through personal networks), but were selected to cover all 6 continents and include different ethnic majorities and religious majorities (Christian, Muslim, Hindu, Jewish, Eastern religions, as well as highly secular societies). Table 7.1 displays the method of recruitment and compensation per country.

The study was approved by the local ethics committee at the Psychology Department of the University of Amsterdam (Project #2018-SP-9713). Additional approval was obtained from local IRBs at the Adolfo Ibáñez University (Chile), the Babes-Bolyai University (Romania), the James Cook University (Singapore), Royal Holloway, University of London (UK), and the University of Connecticut (US).

### 7.2.2 Sampling Plan

We preregistered a target sample size of $n = 400$ per country and 20-25 target countries. The preregistered sample size and composition allowed us to look at overall effects, effects within countries, and between countries. As we applied a Bayesian statistical framework, we needed a minimum of 20 countries to have sufficient data for accurate estimation in cross-country comparisons (Hox et al., 2012). However, our main interest were overall effects - rather than effects for individual countries. With approximately 8,800 participants, we would have sufficient data to reliably estimate overall effects, especially since the source effect is within-subjects. Data collection was terminated by November 30th, 2019. The data from ten participants who completed the survey after this termination date were retained in the dataset.

### 7.2.3 Materials

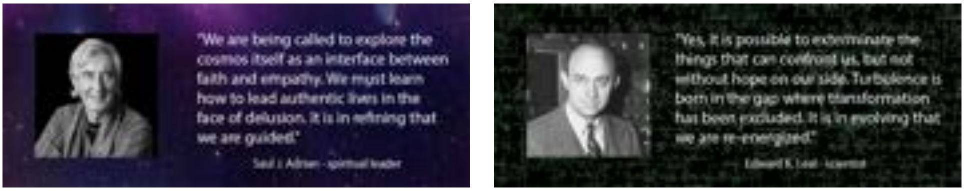
The study was part of a larger project on cross-cultural effects related to religiosity (see the online Appendix for details about the project). The full translated survey for each included country can be found at osf.io/kywjs/. The relevant variables for the current study were individual religiosity, the manipulated source of authority, and the ratings of the statements.

Participant religiosity was measured using established items taken from the World Values Survey (World Values Survey, 2010), covering religious behaviours (institutionalized such as church attendance and private such as prayer/mediation), beliefs, identification, values, and denomination (see the online Appendix for the exact items). Besides having high face-validity, these measures have been applied cross-culturally in other studies (Lindeman et al., 2015; Lun & Bond, 2013; Stavrova, 2015). A Bayesian reliability analysis using the `Bayesrel` package (Pfadt & van den Bergh, 2020) indicated good internal consistency of the religiosity measure, McDonald omega

**Table 7.1:** Descriptives Statistics per Country

| | *N* | Age (SD) | Women (%) | Religiosity | Sample | Compensation |
|---|---|---|---|---|---|---|
| Australia | 463 | 48.3 (16.0) | 48.4 | 0.52 | online panel | money |
| Belgium | 320 | 34.6 (13.1) | 55.6 | 0.24 | mixed | raffle |
| Brazil | 402 | 28.8 (10.4) | 73.1 | 0.51 | mixed | none; credits |
| Canada | 351 | 33.2 (10.5) | 52.4 | 0.28 | online panel | money |
| Chile | 308 | 30.8 (9.9) | 59.1 | 0.33 | mixed | raffle |
| China | 390 | 32.1 (8.4) | 55.9 | 0.32 | online panel | money |
| Croatia | 309 | 28.0 (6.9) | 78.3 | 0.41 | mixed | raffle |
| Denmark | 415 | 27.9 (10.3) | 71.3 | 0.26 | mixed | raffle |
| France | 405 | 40.6 (12.8) | 64.2 | 0.29 | online panel | money |
| Germany | 1,287 | 27.5 (9.0) | 62.2 | 0.32 | mixed | raffle |
| India | 394 | 30.4 (6.5) | 36.3 | 0.73 | online panel | money |
| Ireland | 434 | 42.6 (15.0) | 51.8 | 0.48 | online panel | money |
| Israel | 501 | 27.9 (10.1) | 73.5 | 0.37 | students | credits |
| Italy | 342 | 27.2 (8.2) | 50.9 | 0.26 | mixed | none; money |
| Japan | 424 | 40.6 (10.0) | 43.9 | 0.29 | online panel | money |
| Lithuania | 291 | 24.1 (7.0) | 83.2 | 0.35 | students | none |
| Morocco | 329 | 32.1 (11.8) | 16.1 | 0.70 | online panel | money |
| Netherlands | 482 | 57.6 (14.7) | 25.3 | 0.28 | online panel | money |
| Romania | 539 | 24.4 (7.4) | 85.2 | 0.55 | mixed | raffle |
| Singapore | 308 | 22.2 (3.4) | 62.0 | 0.45 | students | credits |
| Spain | 337 | 41.9 (13.9) | 31.2 | 0.21 | online panel | money |
| Turkey | 362 | 39.2 (11.1) | 24.6 | 0.33 | online panel | money |
| UK | 400 | 36.2 (12.7) | 65.8 | 0.23 | online panel | money |
| US | 402 | 35.8 (14.4) | 51.0 | 0.45 | mixed | none; money |
| Total | 10,195 | 33.8 (13.8) | 55.9 | 0.38 | – | – |

*Note.* Religiosity refers to the self-reported level of individual religiosity, transformed on a 0-1 scale. Sample indicates the composition of the sample based on the method of recruitment per site.


"We are being called to explore the cosmos itself as an interface between faith and empathy. We must learn how to lead authentic lives in the face of delusion. It is in refining that we are guided."
Saul J. Adrian - spiritual leader
"Yes, it is possible to exterminate the things that can confront us, but not without hope on our side. Turbulence is born in the gap where transformation has been excluded. It is in evolving that we are re-energized."
Edward K. Leal - scientist

**(a)** Spiritual authority – Statement A

**(b)** Scientific authority – Statement B

**Figure 7.1:** Example stimuli used in the survey. The statements were generated using the New-Age bullshit generator (http://sebpearce.com/bullshit/) and translated into the language the study was conducted in. The statements were counterbalanced between sources across participants.

= 0.930 [0.927, 0.931]. The religious membership item was removed from the scale, as this item was only moderately correlated with the other items (item-rest correlation = 0.608, all others > 0.706) and dropping it improved the reliability to omega = 0.939 [0.938, 0.941]. The remaining seven individual religiosity items were transformed on a 0-1 scale (to make each item contribute equally to the scale), tallied to create a religiosity score per participant, and grand-mean standardized for the analyses.

The experimental stimuli consisted of two gobbledegook statements that were attributed to a spiritual guru and to a scientific authority (within-subjects). We created two versions of the statement, manipulating (1) the background of the frame: an opaque new-age purple galaxy background vs. an opaque dark green chalkboard with physics equations, (2) the accompanying gray-scale photo of the alleged source: a man in robes (photo of José Argüelles) vs. a man in an old-fashioned suit (photo of Enrico Fermi), and (3) the reported profession: spiritual leader vs. scientist. Additionally, in the introductory text, the source was further announced as "Saul J. Adrian - a spiritual authority in world religions" vs. "Edward K. Leal - a scientific authority in the field of particle physics", names counter-balanced. The names were fictitious and the photos were taken from Wikipedia with re-use permission. The two versions of the text were three-sentence, 37/38 word statements. We generated the statements using the New-Age bullshit generator (http://sebpearce.com/bullshit/), that combines new-age buzzwords in a syntactically correct structure resulting in meaningless, but pseudo-profound sounding texts (Pennycook et al., 2015). The two versions of the text were counterbalanced between sources. Participants were randomly assigned to the scientific-spiritual or the spiritual-scientific ordered condition. The stimuli in each language are provided at osf.io/qsyvw/.

The main outcome variable pertained to judgments of importance and credibility of gobbledegook, measured on a 7-point Likert scale from *not at all important / not at all credible* to *extremely important / extremely credible*, respectively. A multiple choice recognition item for the source that expressed the statement was included as a manipulation check. In our preregistration, we did not specify that we would exclude participants based on incorrect recall of the source of the statement. We therefore kept all observations in the data set for the main analyses and additionally ran the models without the observations for which the source was not recalled correctly. The results of this robustness check are provided in Table 7.3. For exploratory purposes, we also measured reading and processing time for the statement, as well as depth of processing. The latter was operationalized as the number of items correctly identified as having appeared in the statement. Participants were presented with a list of 10 words, including 5 targets and 5 distractors, and were asked to select the words that they recognised from the statement.

### 7.2.4 Procedure

Participants received a link to the Qualtrics survey, either by email, social media or through an online platform. After reading the instructions and providing informed consent, they first completed items for a separate study about religiosity and trustworthiness. Next, they were presented with the first statement and source stimulus, rated its importance and credibility, completed the manipulation check to validate that they registered the source, and completed the word recall item. These elements were then repeated for the second statement. After that, participants completed items about body-mind dualism. Finally, they provided demographics, a quality of life scale, the religiosity items and were given the opportunity to provide comments. It took about 10 minutes to complete the entire survey (median completion time was 11.4 minutes).

### 7.2.5 Data Analysis

We used the R package `BayesFactor` (Morey & Rouder, 2018) to estimate and test the multilevel Bayesian regression models (Haaf & Rouder, 2017; Rouder, Haaf, Davis-Stober, et al., 2019). The multilevel Bayesian modelling approach allows us to systematically evaluate the evidence in the data under different models: (i) across all countries the effect is truly null; (ii) all countries share a common nonzero effect; (iii) countries differ, but all effects are in the same (predicted) direction; and (iv) in some countries the effect is positive whereas in others the effect is negative. The models differ in the extent to which they constrain their predictions, from the most constrained (i) to completely unconstrained (iv). We refer to these models as the null model, the common effect model, the positive effects model, and the unconstrained model, respectively. Note that while the predictions from model (iii) are less constrained than those from model (ii), it is more difficult to obtain evidence for small effects under the latter model because it assumes that the effect is present in every country, rather than only in the aggregate sample. When applied to our hypothesis for the source effect, evidence for (i) would indicate that people from these 24 countries do not differentially evaluate credibility of claims from a guru or a scientist, evidence for (ii) would indicate that on average people from these 24 countries consider claims from a scientist more credible than from a guru (or vice versa) with little between-country variability in the size of the effect, evidence for (iii) would indicate that in all of the 24 countries, people consider claims from a scientist more credible than from a guru (or vice versa), but there is cultural variation in the size of this effect, and evidence for (iv) would indicate that in some countries people consider claims from a scientist more credible than from a guru, and in other countries people consider claims from a guru more credible than from a scientist, indicating cultural variation in the direction (and size) of the effect. We used the interpretation categories for Bayes factors proposed by Lee and Wagenmakers (2013), based on the original labels specified by Jeffreys (1939).

For the main effect of source ($\mathcal{H}_1$), we specified the following unconstrained model. Let $Y_{ijk}$ be the credibility rating for the $i$th participant, $i = 1, ..., N$, in the $j$th country, $j = 1, ..., 24$, for the $k$th condition, $k = 1, 2$. Then:

$$Y_{ijk} \sim N(\mu + \alpha_j + v_i\beta + r_i\delta_j + x_k\gamma_j, \sigma^2).$$

Here, the term $\mu + \alpha_j$ serve as the baseline credibility intercepts with $\mu$ being the grand mean and $\alpha_j$ the $j$th country's deviation from the grand mean. The $\beta$ term reflects the fixed effect of the level of education covariate. $\delta_j$ is the $j$th country's main effect of religiosity on credibility ratings. The crucial parameter here is $\gamma_j$ which is the source effect for the $j$th country. In the common effects model, we will replace $\gamma_i$ with $\gamma$. The variable $x_k = -0.5, 0.5$ if $k = 1, 2$, respectively, where $k = 1$ indicates the scientist condition and the $k = 2$ indicates the guru condition. The variable $v_i$ is the standardized participant-level education covariate. The variable $r_i$ is the standardized religiosity score for each participant. Finally, $\sigma^2$ is the variance in credibility ratings across participants.

To test the source-by-religiosity interaction for hypothesis 2, the model from (1) is extended by including an interaction term:

$$Y_{ijk} \sim N(\mu + \alpha_j + v_i\beta + r_i\delta_j + x_k\gamma_j + r_i x_k\theta_j, \sigma^2),$$

where $\theta_j$ is the parameter of interest, the religiosity*source interaction effect, with $r_i x_k$ as the product of the experimental condition and the standardized individual religiosity score. The parameter estimates as reported in the results section are based on the full model from (2).

In order to systematically investigate which third variables should and should not be included in the statistical model, we used *directed acyclic graphs* (DAGs Pearl, 1995) to visually represent the causal relations between the variables in our data (McElreath, 2020; Pearl, 2019; Rohrer, 2018). In short, this method entails specifying directed relations (arrows) between different constructs and measures (nodes) in a given design, that allow one to intuitively reflect causal structures and determine which third variables should be accounted for and which should be ignored in the statistical model. Based on DAGs created in the R package `ggdag` (M. Barrett, 2021), both *country* and *level of education* were identified as potential confounding factors that warranted inclusion, as they may affect both religiosity (Albrecht & Heaton, 1984; Schwadel, 2016) and overall credibility assessments (e.g., due to skepticism). Country was therefore added as a clustering factor, while level of education was added as a fixed covariate in all models. We also ran the models while including all participant-level variables related to the primary measures, i.e., gender (Miller & Hoffmann, 1995), age (Argue et al., 1999), SES (Pyle, 2006; C. Smith & Faris, 2005), statement version (A or B), and presentation order (guru–scientist or scientist–guru). Note that including these covariates improved the model fit, but the qualitative results remain the same regardless of the (set of) covariates. See the figures in the online Appendix (https://osf.io/9smk5/) for details on the causal graphs and Table 7.3 for the primary results without any and with all covariates.

#### 7.2.5.1 Prior Settings

The BayesFactor package applies the default priors for ANOVA and regression designs (Rouder & Morey, 2012; Rouder et al., 2012), in which the researcher can determine the scale settings for each individual predictor in the model. We used the settings for the critical priors in the multilevel models as proposed by Rouder, Haaf, Davis-Stober, et al. (2019), concerning the scale settings on $\mu_\gamma, \mu_\theta$ and $\sigma^2_\gamma, \sigma^2_\theta$. The scale on $\mu_\gamma, \mu_\theta$ reflects the expected size of the overall source effect and source-by-religiosity effect, respectively, and is set to 0.4 (small-medium effect). The scale of $\sigma^2_\gamma, \sigma^2_\theta$ reflects the expected amount of variability in these effects across countries. This scale is set to 60% of the overall effect, resulting in a value of 0.24. The prior scale for the overall between-countries variance was set to 1. We used 31,000 iterations for the Markov chain Monte Carlo sampling and discarded the first 1,000 iterations ("burn-in").

### 7.2.6 Deviations from Preregistration

We deviated from the preregistration in the following ways. First, in our preregistration, we formulated a hypothesis about the interaction between source and perceived cultural norms of religiosity in one's country. However, in retrospect, we realized this hypothesis lacked theoretical justification and the proposed analysis was methodologically suboptimal (see the online Appendix for details on this analysis).

Second, as a stopping rule, we preregistered that data collection would be terminated (a) when the target of $n = 400$ per country was reached, or (b) by September 30th, 2019. However, due to unforeseen delays in construction of the materials and recruitment, this deadline was extended until November 30th, 2019. We did not download or inspect the data until after November 30th.

Third, we preregistered to only include countries where usable data from at least 300 participants were collected (i.e., complete data from attentive participants). However, we decided to keep the $n = 291$ participants from Lithuania in the final sample, as the hierarchical models account for uncertainty in estimates from countries with smaller samples and removing these data will actually reduce the overall precision of the estimates. Moreover, it would simply be unfortunate to remove all data from a highly understudied country.

Fourth, we preregistered that we would use the R package brms (Bürkner, 2017) to analyse the data and estimate model parameters. However, we ended up using the BayesFactor package (Morey & Rouder, 2018). This method is arguably more suitable for model comparison and calculating Bayes factors in particular. However, we also ran the models as preregistered and report these results in the online Appendix.

Fifth, we added level of education as a participant-level covariate to the models, which improved the model fits. Note that adjustments 3-5 did not qualitatively change any of the results (see Table 7.3 and the online Appendix).

## 7.3 Results

The two dependent variables that were measured (i.e., *importance* of the message and *credibility* of the message) were highly correlated for both the scientific source (Spearman's $\rho = 0.772$, 95% credible interval $[0.764, 0.779]$) and for the spiritual

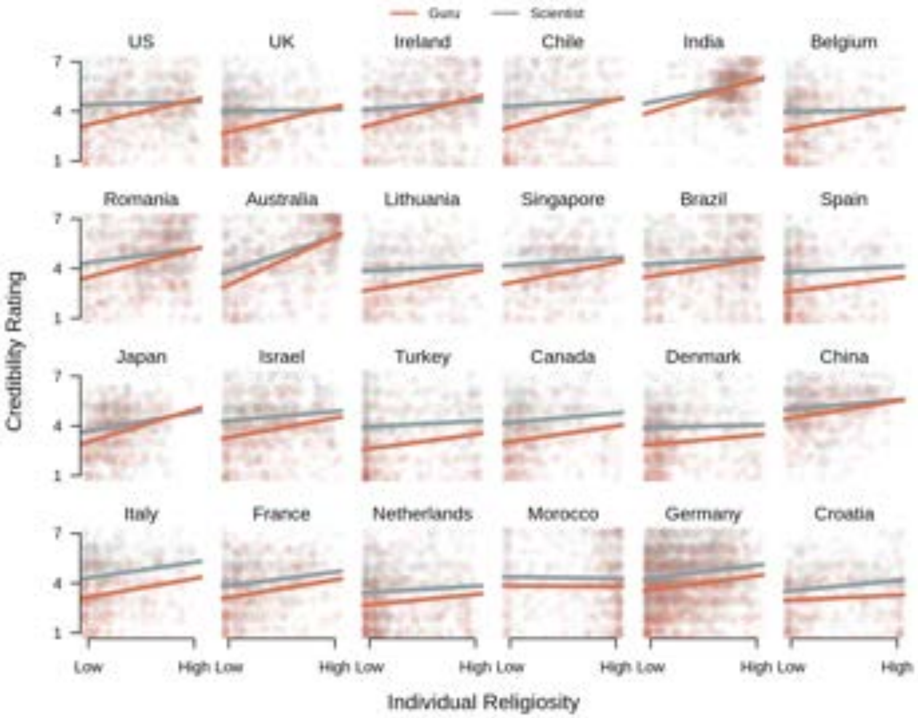


**Figure 7.2:** Observed relation between religiosity and credibility ratings per source, for each country. Countries are ordered by size of the source-by-religiosity interaction (from left to right, top to bottom). Red lines denote ratings for the spiritual guru and grey lines denote ratings for the scientist. Data points are jittered to enhance visibility. Credibility was measured on a 7-point Likert scale.

source (Spearman's $\rho = 0.827$, 95% credible interval $[0.822, 0.833]$; see also the online Appendix; van Doorn, Ly, et al., 2020). As the pattern of results was equal across the dependent variables, we decided to only describe the findings for *credibility* in detail (see Table 7.3 for the results for *importance*).

### 7.3.1 Effect of Source on Credibility

First, we assessed the extent to which the perceived credibility of a gobbledegook statement is affected by its source (i.e., a scientist vs. a spiritual guru). Note, our initial hypothesis was that there would be no main effect of source, that is, we expected evidence for the null-model. However, based on visual inspection of the data (see Figure 1) a main effect of source seems evident. To quantify the evidence for the effect of source, we compared between the null model without an effect of condition (i.e., the scientist and spiritual guru are judged equally credible), the model with a *common positive effect* of condition across countries (i.e., the scientist is judged more credible than the guru, to an equal degree in every country), the model with a *varying positive effect* of source (i.e., the scientist is judged more credible than the guru, but

**Table 7.2:** Bayes factor model comparisons to test $\mathcal{H}_1$ and $\mathcal{H}_2$

| Model | | Bayes factor | $p(\mathcal{M})$ |
|---|---|---|---|
| Hypothesis 1: Source effect | | | |
| $\mathcal{M}_0$ | Country$_u$ + Religiosity$_u$ | 1-to-$10^{228}$ | $< .01$ |
| $\mathcal{M}_1$ | Country$_u$ + Religiosity$_u$ + Source$_1$ | 1-to-$10^{17}$ | $< .01$ |
| $\mathcal{M}_+$ | Country$_u$ + Religiosity$_u$ + Source$_+$ | $*$ | .92 |
| $\mathcal{M}_u$ | Country$_u$ + Religiosity$_u$ + Source$_u$ | 1-to-12.30 | .08 |
| Hypothesis 2: Source-by-Religiosity Effect | | | |
| $\mathcal{M}_0$ | Country$_u$ + Religiosity$_u$ + Source$_u$ | 1-to-$10^{15}$ | $< .01$ |
| $\mathcal{M}_1$ | Country$_u$ + Religiosity$_u$ + Source$_u$ + Source*Religiosity$_1$ | $*$ | .50 |
| $\mathcal{M}_+$ | Country$_u$ + Religiosity$_u$ + Source$_u$ + Source*Religiosity$_+$ | 1-to-1.28 | .39 |
| $\mathcal{M}_u$ | Country$_u$ + Religiosity$_u$ + Source$_u$ + Source*Religiosity$_u$ | 1-to-4.60 | .11 |

*Note.* Asterisks mark the preferred model for each hypothesis. The remaining values are the Bayes factors for the respective model vs. the preferred model. Subscripts reflect parameter constraints; $_u$ indicates an unconstrained effect, $_1$ indicates a common (positive/negative) effect, $_+$ indicates a varying positive/negative effect. $p(\mathcal{M})$ gives the posterior model probability per hypothesis. All models include the covariate level of education.

to varying degrees across countries), and the *unconstrained* model that allows the source effect to be varying from both positive to negative (i.e., in some countries, the scientist is considered more credible than the guru, in other countries, the guru is considered more credible than the scientist).

The Bayes factor model-comparison summarized in Table 7.2 shows that the data provide most evidence for the *positive effects model*, which assumes a varying but consistently positive effect across countries. The source effect is favoured $1.1 \times 10^{210}$-to-1 over the null-model, which indicates strong evidence that the meaningless statement from the scientist is considered more credible than the meaningless statement from the guru. The positive effects model strongly outperforms the common effect model ($\text{BF}_{+1} = 8.9 \times 10^{17}$; explained variance (Bayesian $R^2$) is 17.9%, 95% credible interval [17.0%, 18.7%]). The mean and 95% credible interval of the unstandardized size of the source effect in the full model is 0.70 [0.60, 0.79] on a 7-point Likert scale and the standard deviation between countries is 0.16. Also note that as shown in Figure 1 the within-country individual differences in credibility ratings are large, indicating that most of the variance is located at the lower level (i.e., the individual level). The intraclass correlation coefficients (ICCs) quantifying the proportion of variance explained by the country clustering, as well as the total explained variance by the included effects for all models (Bayesian $R^2$) are reported in Appendix 7.A. There, we also report MCMC diagnostics to verify the adequacy of the Bayesian models, as well as the estimates for the intercepts, source effect, and the source-by-religiosity interaction effect for each country.

### 7.3.2 Interaction Between Source and Religiosity on Credibility

The *source-by-religiosity interaction effect* assesses to what extent the effect of source depends on raters' own religious background (religiosity was globally standardised).

Our hypothesis states that for low religious individuals, credibility ratings should be higher for gobbledegook from a scientific source than for gobbledegook from a spiritual guru. For highly religious individuals, the reversed effect is expected, i.e., higher credibility ratings for gobbledegook ascribed to a guru than for gobbledegook ascribed to a scientist. The interaction term was therefore constrained to be *negative*, in the sense that the coefficient of the source effect becomes smaller (or negative) with increased religiosity. Note that although the interaction term was constrained to have a negative sign, for consistency, we still refer to the model as the positive effects model.

For hypothesis 2, the model comparison summarized in Table 7.2 shows that the data provide most evidence for the *common source-by-religiosity interaction model*, which assumes a consistent interaction effect across countries, $\text{BF}_{10} = 0.99 \times 10^{15}$ ($R^2 = 18.1\%$ [17.2%, 19.0%]). The data are uninformative for distinguishing between the common interaction and the varying positive interaction model ($\text{BF}_{1p} = 1.28$), indicating that both are equally plausible. While we cannot conclude whether or not the size of the interaction effect differs substantially between countries, both models provide strong evidence for a source-by-religiosity effect across all countries. The mean of the unstandardized source-by-religiosity interaction effect is -0.21 [-0.29, -0.14] and the standard deviation between countries is 0.09 on the 7-point Likert scale. As becomes evident from Figure 2d, the interaction entails that the relative preference in credibility for statements from the scientist versus the spiritual guru decreases with higher religiosity. This effect is further unpacked in Figure 2c, which shows that in every country, except for Croatia, religiosity is more predictive of credibility ratings for statements from the guru than for statements from the scientist.

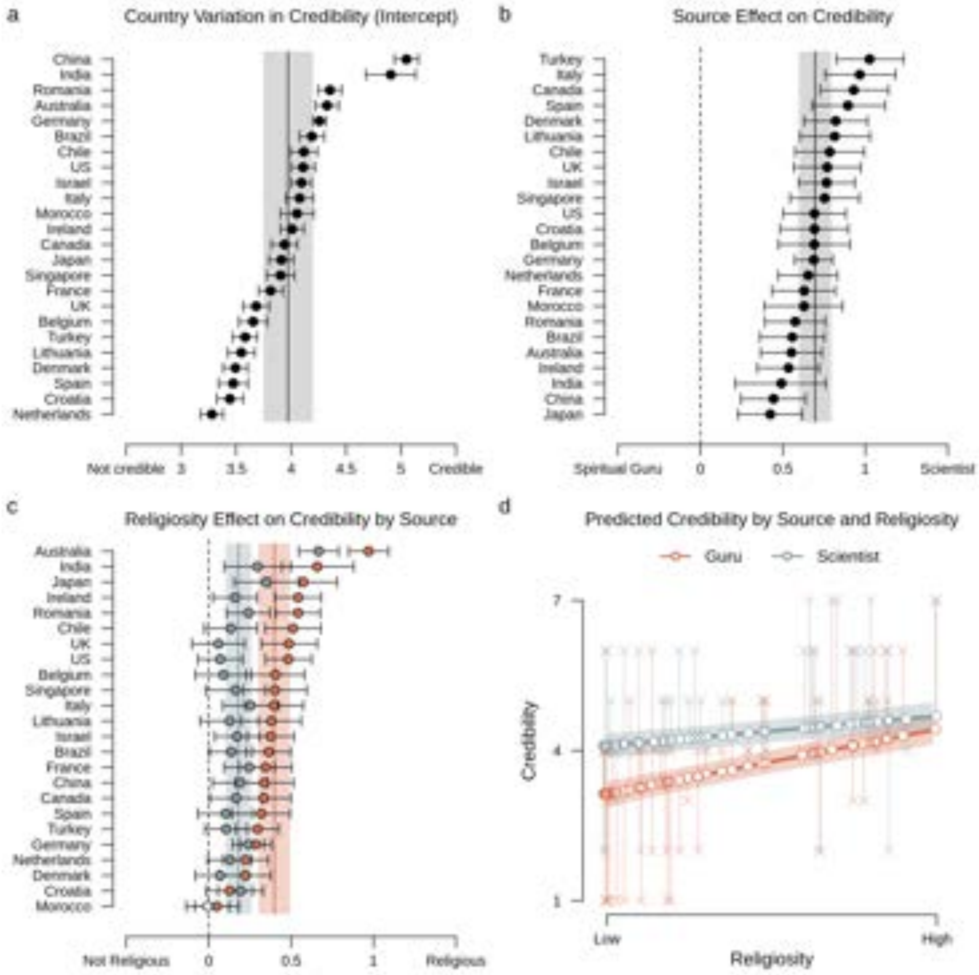


**Figure 7.3:** Summary of the multilevel-model (unconstrained) estimates per country and predicted overall effects. It is apparent that there is substantial variation across the 24 countries in (**a**) overall credibility judgments (i.e., intercept) and (**b**) the effect of scientific vs. spiritual source. Panel **c** shows that individual religiosity has a stronger effect on credibility judgments for the spiritual guru (red circles) than for the scientist (grey circles). The estimates are ordered from largest to smallest, and the open circles denote negatively valued effects. The errorbars give the 95% credible interval for each country. The vertical lines denote the overall estimated effect with the 95% credible interval in the shaded bands. The dashed lines indicates zero. Panel **d** displays the predicted credibility as a function of source and individual religiosity, showing that the difference in credibility ratings for the scientist (grey lines) vs. the guru (red lines) is less pronounced for high religiosity individuals than low religiosity individuals. The shaded bands reflects the 95% credible intervals, the x's reflect the observed values for 2 randomly sampled participants per country, and the circles reflect the corresponding estimated values. The x's and circles are jittered to enhance visibility.

### 7.3.3 Exploratory Analyses

In an exploratory fashion, we assessed to what extent the source manipulation influenced the effort participants put into processing the statements. To this end, we looked at (1) response time for the evaluation of each statement as a proxy for processing time of the message, and (2) memory performance of words presented in the statements as a proxy for encoding quality. For these exploratory models, we only assessed evidence for a common effect, as visual inspection of the data suggested no or only very small and homogeneous effects (see Figure 3).

#### 7.3.3.1 Processing Time

For processing time the data indicate a common effect of source: participants spent more time processing the statement of the scientist (median RT = 28.30 seconds) than that of the guru (median RT = 27.0 seconds; $\text{BF}_{10}$ = 8,050.48). Processing times were log-transformed for the analysis, to account for the positive skew that is typically observed in response time data. However, the standardized effect size is very small: 0.058 [0.023, 0.087]. There was strong evidence against an interaction between source and religiosity ratings on processing time: religiosity is not predictive of the difference in processing time for the scientist vs. the guru ($\text{BF}_{10}$ = 0.03, $\text{BF}_{01}$ = 30.78).

7

#### 7.3.3.2 Memory Performance

After the rating question, participants were presented with a recall item that required them to indicate which words they recognized from the statement. The list consisted of 5 target (included in the statement) and 5 distractor words (not in the statement) for each source. An $F_1$ score was calculated per person per source, which gives the harmonic mean of the precision (proportion true positives of all selected words) and recall (proportion true positives of all presented target words). $F_1$ ranges between 0 and 1, with 1 being perfect performance.

The analysis indicated anecdotal evidence against a common effect of source on memory performance: participants did not perform better on recognising words from the statement by the scientist than by the guru ($\text{BF}_{10}$ = 0.53; $\text{BF}_{01}$ = 1.90; standardized estimate = 0.014 [0.001, 0.035]). Finally, there was moderate evidence against an interaction, $\text{BF}_{10}$ = 0.31, $\text{BF}_{01}$ = 3.27.

As a sanity check, we showed that there is an extremely strong effect of processing time on memory performance; participants who spent more time processing the statement, also performed better on the memory task ($\text{BF}_{10} = \infty$).

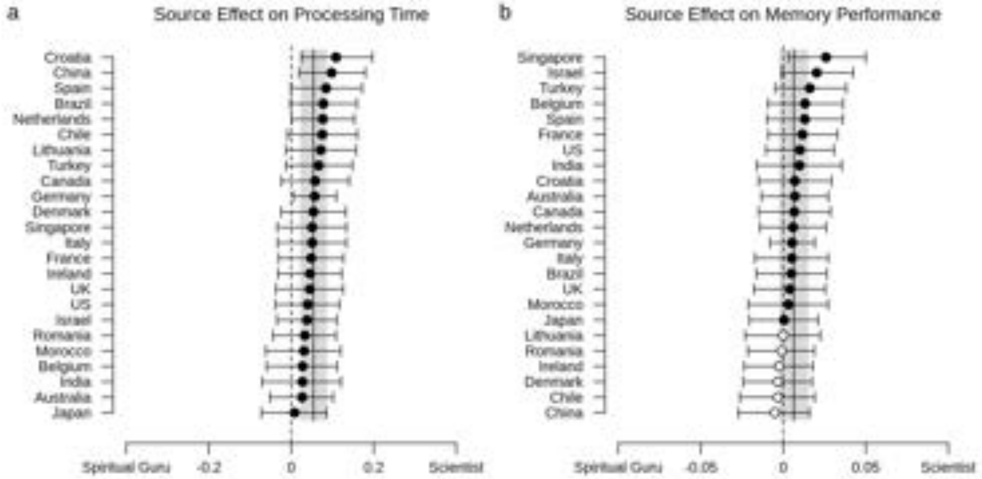


**Figure 7.4:** Multilevel-model (unconstrained) estimates for the source effect (**a**) on (log-transformed) processing time and (**b**) on memory performance (range 0–1). The estimates are ordered from largest to smallest, and the open circles denote negatively valued effects. The errorbars give the 95% credible interval for each country. The vertical lines denote the overall estimated effect with the 95% credible interval in the shaded bands. The dashed lines indicates zero.

### 7.3.4 Validation Using Previously Collected Trust Ratings

In addition to the experimental data collected in this study, we also examined an existing dataset that includes surveyed trust ratings for scientists and traditional healers for 117,191 participants across 143 countries. Note that the analysis on this dataset was not preregistered. Analysis of these data corroborated the results from our experimental manipulations; on average scientists are considered more trustworthy than traditional healers, standardized estimate = 0.30 [0.06, 0.58] (for comparison: the standardized estimate for the experimental source effect on credibility is 0.41 [0.22, 0.49]). While the positive effects model strongly outperforms both the null model and the common effect model ($\mathrm{BF}_{+0}$, $\mathrm{BF}_{+1} > 10^{308}$; $R^2$ for the positive effects model $= 28.1\%$ $[27.8\%, 28.3\%]$), the analysis indicates most evidence for the unconstrained model $\mathcal{M}_u$, which indicates that scientists are not explicitly trusted more than traditional healers in all of the 143 countries, $\mathrm{BF}_{u+} = 320.76$. Nonetheless, as displayed in Figure 4a, only in 3 out of the 143 countries the mean of the estimated source effect is negative, while the overall effect is clearly positive.

We also investigated the fit-effect in this dataset, by including an interaction term between authority (scientists vs. traditional healers) and religiosity (religious vs. not religious). Because in 41 countries all of the participants indicated that they were religious, we could not reliably estimate varying effects for the authority-by-religiosity interaction. There was, however, strong evidence for an overall interaction between authority and religiosity, $\mathrm{BF}_{10} = 6.3 \times 10^{14}$, $R^2 = 28.1\%$ $[27.8\%, 28.4\%]$ standardized estimate = -0.09 [-0.14, -0.02] (for comparison: the standardized estimate for the experimental source-by-religiosity effect on credibility is -0.12 [-0.16, -0.08]). The

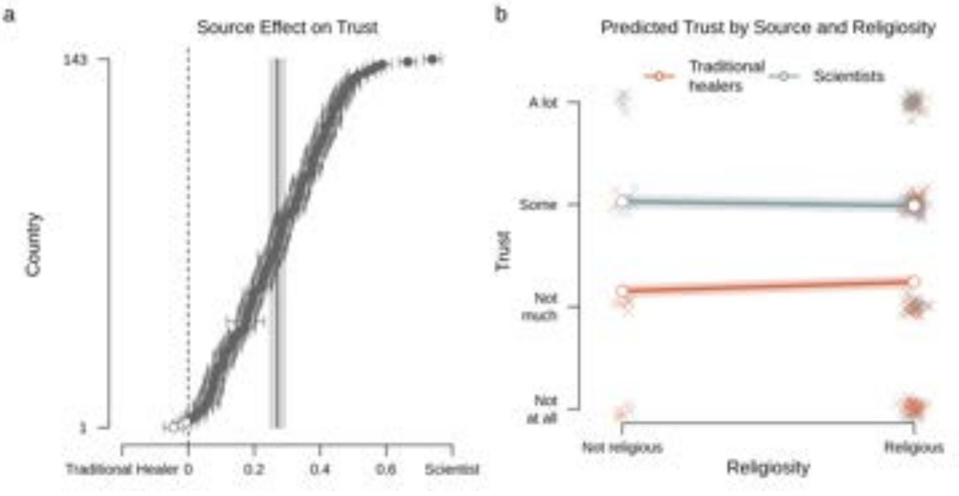


**Figure 7.5:** Multilevel-model (unconstrained) estimates and predicted overall effects for explicit trust ratings. Panel **a** displays the source effect on trust ratings for each of the 143 countries, showing that in all but 3 countries, scientists are trusted more than traditional healers. The estimates are ordered from largest to smallest, and the open circles denote negatively valued effects. The errorbars give the 95% credible interval for each country. The vertical lines denote the overall estimated effect with the 95% credible interval in the shaded bands. The dashed lines indicates zero. Panel **b** displays the predicted trust rating as a function of source and individual religiosity, showing that religious individuals trust scientists slightly less and traditional healers more compared to non-religious individuals. The shaded bands reflects the 95% credible intervals, the x's reflect the observed values for 2 randomly sampled participants per country, and the circles reflect the estimated values per condition. The x's are jittered to enhance visibility.

pattern of the interaction is the same as for the experimental credibility data: the relative difference between trust in scientists vs. traditional healers is smaller for religious individuals than for non-religious individuals. Interestingly, while the experimental study found that religiosity was associated with increased credibility ratings for both sources, albeit to a smaller extent for the scientist (see Figure 2c), the trust data show a positive effect of religiosity on trust for traditional healers (standardized estimate = 0.03 [0.02, 0.04]), yet a negative effect of religiosity on trust for scientists (standardized estimate = -0.01 [-0.02, -0.01]). See Appendix 7.B for an additional exploratory analysis on the country-level correlation in the source effect between the primary experimental dataset and secondary validation dataset on trust.

### 7.3.5 Robustness and Additional Checks

We conducted 8 additional analyses that the results should be robust against, including all specifications mentioned in the preregistration:

1. Excluding observations for which participants did not correctly recall the source

of the statement ($n_{obs} = 1616$ [7.95%]);

2. Excluding data from Lithuania because $n < 300$ (as preregistered);
3. Using a different, less informed prior setting for $r$ scale; $r = \frac{\sqrt{2}}{2} \approx 0.707$, corresponding to a 'wide' prior scale provided in the `BayesFactor` package (Morey & Rouder, 2018);
4. Using the *importance* rating instead of the *credibility* rating as the outcome variable.
5. Applying a between-subjects design by only taking the first observation per participant.
6. Including all participants, including those who failed the attention check.
7. Running the analyses without adding any predictors as covariates;
8. Running the analyses including all covariates that might affect either the independent variable (religiosity) or the dependent variable (credibility ratings): statement version (A or B), presentation order (guru–scientist or scientist–guru), participant age (in decades), participant gender, level of education, and perceived socio-economic status (SES).

The results of these robustness analyses are given in Table 7.3 and corroborate the conclusions from the main analyses: the data indicate (a) a source effect that varies between countries but is consistently positive (scientist > guru), and (b) a positive source-by-religiosity interaction effect (either a common or varying effect).

**Table 7.3:** Bayes factor of different models for robustness checks

| Robustness Set | $N_{obs}$ | Estimate [95%CI] | $\text{BF}_{10}$ | $\text{BF}_{+1}$ | Preferred |
|---|---|---|---|---|---|
| Source effect | | | | | |
| Main analysis | 20,318 | 0.70 [0.60, 0.79] | $10^{210}$ | $10^{17}$ | $\mathcal{M}_+$ |
| Excluding source incorrect | 18,702 | 0.78 [0.69, 0.88] | $10^{249}$ | $10^{15}$ | $\mathcal{M}_+$ |
| Excluding Lithuania ($n < 300$) | 19,736 | 0.69 [0.59, 0.79] | $10^{200}$ | $10^{17}$ | $\mathcal{M}_+$ |
| Default prior settings | 20,318 | 0.70 [0.56, 0.84] | $10^{210}$ | $10^{15}$ | $\mathcal{M}_+$ |
| Importance as outcome variable | 20,318 | 0.53 [0.43, 0.63] | $10^{113}$ | $10^{11}$ | $\mathcal{M}_+$ |
| Between-subjects design | 10,159 | 0.83 [0.68, 0.98] | $10^{145}$ | $10^{20}$ | $\mathcal{M}_+$ |
| Including all subjects | 20,980 | 0.69 [0.59, 0.78] | $10^{210}$ | $10^{20}$ | $\mathcal{M}_+$ |
| No covariates | 20,318 | 0.70 [0.60, 0.79] | $10^{199}$ | $10^{17}$ | $\mathcal{M}_+$ |
| All covariates | 20,318 | 0.70 [0.60, 0.79] | $10^{211}$ | $10^{17}$ | $\mathcal{M}_+$ |
| Fit Effect (Source*Religiosity) | | | | | |
| Main analysis | 20,318 | -0.21 [-0.29, -0.14] | $10^{15}$ | 0.78 | $\mathcal{M}_1$ |
| Excluding source incorrect | 18,702 | -0.23 [-0.32, -0.15] | $10^{17}$ | 4.85 | $\mathcal{M}_+$ |
| Excluding Lithuania ($n < 300$) | 19,736 | -0.21 [-0.29, -0.13] | $10^{14}$ | 0.90 | $\mathcal{M}_1$ |
| Default prior settings | 20,318 | -0.21 [-0.34, -0.09] | $10^{13}$ | $10^{-6}$ | $\mathcal{M}_1$ |
| Importance as outcome variable | 20,318 | -0.18 [-0.26, -0.10] | $10^{9}$ | 0.02 | $\mathcal{M}_1$ |
| Between-subjects design | 10,159 | -0.22 [-0.33, -0.12] | $10^{7}$ | 4.67 | $\mathcal{M}_u$ |
| Including all subjects | 20,980 | -0.22 [-0.29, -0.14] | $10^{15}$ | 0.56 | $\mathcal{M}_1$ |
| No covariates | 20,318 | -0.22 [-0.29, -0.14] | $10^{14}$ | 0.77 | $\mathcal{M}_1$ |
| All covariates | 20,318 | -0.21 [-0.29, -0.13] | $10^{16}$ | 0.09 | $\mathcal{M}_1$ |

*Note.* Across all eight sets of robustness checks, the results are qualitatively equal to those of the main analyses (column 1); the data indicate (a) a strong source effect that varies between countries but is consistently positive (scientist > guru), (b) a source-by-religiosity interaction effect (either a common or varying effect). Subscripts reflect parameter constraints; $_0$ indicates the null model, $_+$ indicates a varying positive effect, and $_1$ indicates a common effect. Preferred refers to the best predicting model based on the data.

## 7.4 Discussion

In the current cross-cultural study, we used a straightforward manipulation and measurement of source credibility effects at the individual level. We found a robust source effect on credibility judgments of meaningless statements ascribed to different authority figures; across all 24 countries and all levels of religiosity, gobbledegook from a scientist was considered more credible than the same gobbledegook from a spiritual guru. In addition to this robust overall Einstein effect, participants' background beliefs predicted the credibility evaluations; individuals scoring low on religiosity considered the statement from the guru less credible than the statement from the scientist, while this difference was less pronounced for highly religious individuals. These patterns were consistent with explicit trust data collected for over 100,000 individuals from 143 countries: across 140 out of 143 of these countries, people indicated greater trust in scientists than in traditional healers, with a larger difference for non-religious compared to religious individuals. Robustness analyses for the experimental study indicated that the effects were robust against different data inclusion criteria (e.g., attention checks) and analytic choices (e.g., selection of covariates, dependent variable, prior settings). Moreover, the effects also compellingly emerged when analysed as a between-subjects design (see Table 7.3), suggesting that they are not simply explained by social desirability or participants responding in line with their guess of the research hypothesis (also note that recent empirical work indicates that online survey experiments are generally robust to experimenter demand effects; Mummolo and Peterson, 2019). Results of exploratory reaction time analyses suggest that in addition to giving more positive evaluations, people may actually put more effort into processing information from credible sources (though they did not recall it better). In particular, participants spent more time and may have tried relatively harder to decipher the gobbledegook from the scientist, whereas prior scepticism may have steered some to immediately dismiss the information from the guru as nonsense.

The pattern of results suggests that variability in the source effect between individuals and countries is more strongly driven by differences in credibility of the spiritual authority than the scientific authority. Based on the literature one could consider various plausible hypotheses explaining cross-cultural variation in the source effects, for instance in terms of cultural religiosity, vertically vs. horizontally structured societies, general trust in authorities, and specific trust patterns toward religious and secular authorities (Gervais et al., 2018; Inglehart, 2006; Mitkidis et al., 2015; Singelis et al., 1995; Stavrova, 2015; World Values Survey, 2010). However, while our analysis indicated quantitative differences in the size of the source effect between countries (i.e., varying positive effects), we did not find qualitative differences (i.e., changes in the direction or presence of the effect). Descriptively, the weakest source effects (i.e., smallest difference between the scientific and the spiritual source) are observed in Asian countries (Japan, China, India), possibly because the spiritual guru as presented in the survey more closely fits Eastern belief systems than Abrahamic faith traditions. However, this explanation remains speculative and we are hesitant to over-interpret the cross-national variability both in the overall credibility judgments and the effect of source. While we included main effects of age, gender, level of education and socio-economic status in the analyses, the different sampling strategies that were applied between countries also calls for caution in making inferences based on direct comparisons.

Our findings could reflect a universal gullibility with regard to gobbledegook statements: only a small minority of participants, regardless of their national or religious background, displayed candid scepticism towards the nonsense statements, and 76% of participants rated the scientist's gobbledegook at or above the midpoint of the credibility scale (vs. 55% for the guru). However, the notion of a general gullibility underlying the observed effects is not entirely supported by the data. The median response was the midpoint of the credibility scale. Participants may have primarily used the midpoint of the scale to indicate that they were uncertain about whether or not the claim was credible, i.e., to refrain from passing judgment at all (Krosnick, 1991; Raaijmakers et al., 2000; Sturgis et al., 2014). This response might appear as a lack in motivation to critically reflect on the information that was presented; at the same time, saving one's cognitive resources can also be considered 'strategic'. First, as with most psychology experiments, our study was a zero-stakes task with no incentive for accuracy, which may have lowered effort and biased responses toward the midpoint. Second, when analytic reasoning about the plausibility of a presented claim does not yield any conclusion, the most rational thing to do may be either suspending judgment (selecting the neutral midpoint of the rating scale) or calibrating judgment to prior beliefs about the source of the claim. If one considers the group to which the source belongs generally competent and benevolent, it makes sense to give a positive judgment of their difficult-to-evaluate claim. After all, credible experts often acquired credentials based on their reputation of discovering phenomena that seem implausible at first glance (Levy, 2019). For instance, the premises of using vaccines ('inserting a virus prevents disease') or facts about climate change ('humans are changing the weather') are intuitively dubious, yet reputable scientists have convinced many laypeople of their truth.

In this study, we intentionally selected authorities that are generally considered benevolent (Funk, 2020; Krause et al., 2019) and we generated statements that are nearly impossible to (in)validate and that bear no relation to controversial or politicized scientific topics about which people may have strong prior attitudes (such as efficacy of vaccinations, climate change etc.). By using ambiguous claims without any specific ideological content, we tried to isolate the worldview effect regarding the source from any worldview effect related to the content of the claims. At the same time, we aimed to maximize the efficacy of our manipulation, by varying the names, photographs, and visual contexts (chalkboard vs. stars) in addition to the authorities' profession. This approach makes it more difficult to single out which specific factor contributes to the source effect (e.g., the observed effects might be partly driven by the authorities' appearance rather than their domain of expertise). Relatedly, some participants might have recognized the depicted men (Enrico Fermi and José Argüelles), although we consider it unlikely that many did. As we did not ask whether participants recognized any of the depicted sources, we tried to indirectly and retrospectively assess recognition by scanning the open text items at the end of the survey (comments and awareness item) for any mentioning of either 'Enrico', 'Fermi', 'José', or 'Argüelles' (ignoring capitalization or diacritical marks). Only one (Spanish) participant mentioned recognizing both of the sources. While this obviously does not prove no other participants might have known the depicted sources, it seems unlikely that this was the case for a large proportion of participants. On the other hand, the multifaceted nature of the manipulation also increases its ecological validity; our stimuli resemble popular internet memes and real-life instances of source credibility also involve a combination of different features (e.g., authorities typically look the part in public and appear in congruous contexts). Furthermore, a recent study showed that the mere mentioning of a famous source such as Aristotle or the Dalai Lama enhanced profundity ratings for pseudo-profound nonsense relative to unauthored versions, suggesting that even the mere name of an authority may suffice to induce source effects (Gligorić & Vilotijević, 2020).

The effects observed in our experimental data and the associations identified in the existing trust data were highly comparable, suggesting that by using our source credibility manipulation we tapped into participants' attitudes about scientific and religious authorities. A noteworthy divergence, however, is that whereas our data showed a small positive relation between religiosity and credibility ratings for gobbledegook from the scientist, the trust data demonstrated a small but negative association between religiosity and trust in scientists. The finding that religious people are generally less trusting towards science has often been reported in the literature (Farias et al., 2013; Gauchat, 2012; McPhetres & Zuckerman, 2018; O'Brien & Noy, 2018). However, recent studies suggest that the negative relation between religiosity and trust in science might be US-specific and be weak or absent in other countries (Cacciatore et al., 2018; McPhetres et al., 2020; Rutjens et al., 2021; Rutjens & van der Lee, 2020). Additionally, although trust is likely closely linked to credibility, explicit trust assessments and credibility ratings of specific statements may diverge, perhaps particularly for the kind of obscure statements used in the current study. That is, the gobbledegook statements may still have resonated better with religious individuals than non-religious individuals, resulting in the main effect of religiosity on credibility ratings. This main effect may be driven by a tendency for intuitive reasoning, which has been related to religiosity (Gervais et al., 2018; Pennycook et al., 2012; Pennycook et al., 2016) and receptivity of pseudo-profound and pseudo-scientific nonsense (A. M. Evans et al., 2020; Pennycook et al., 2015). It could thus be that mistrust in science only partially dampens the allure of well-sounding science-related gobbledegook for intuitive reasoners (A. M. Evans et al., 2020).

Notably, our study showed that across 24 countries even those who are highly religious are prone to a scientific source credibility bias, what we have deemed the Einstein effect. Looking ahead, there are at least six compelling horizons for future research to address the generalizability and underlying causes of the Einstein effect. First, whether scientific education diminishes the appeal of scientific authority outside its immediate domain remains unclear. Although those who place faith in science are prone to Einstein effects (Eriksson, 2012; Fernandez-Duque et al., 2014; Macdonald et al., 2017; Mayo, 2019), strong scepticism is normative within the practice of science – as anyone who has experienced peer-review will attest. Although it is 150 years after Charles Peirce famously argued for fixing beliefs from the "method of science" in favour the "method of authority" the role of appeals to scientific authority among scientists remains unclear (Peirce, 1992). Second, future researchers might investigate whether political partisanship predicts differences in scientific-source credibility. Although political commitments may share common psychological features with religious commitments (Grafman et al., 2020; Graham et al., 2009; M. K. Johnson et al., 2011; Malka et al., 2012), the rise of anti-science populist ideologies might diminish or reverse Einstein effects among political partisans. In contrast, individual differences in deference to science (Howell et al., 2020) may predict enhanced Einstein effects, although a recent study failed to find this pattern for faith in science (van der Miesen et al., 2022). Third, the historical origins of scientific source credibility across different cultures remain unclear. If we were to wind back the clock a century to Einstein's era, would we also observe preferential source-credibility for scientific authority over spiritual authority? Fourth, the proximate and sustaining social and technological causes of scientific source credibility are not addressed in our study, and remain ripe for investigations. Is scientific source credibility an artefact of global information networks, country-wide science education, or the sequestering of religious authority to the private domain? Fifth, although our study covers 24 countries worldwide, we cannot claim universality for our findings. Indeed, investigating source credibility in cultures where spiritual authority dominates may help to clarify the mechanistic questions that our study raises but does not address. Sixth, future work may extend the current work and investigate how the Einstein effect is affected by content cues (e.g., the use of jargon, argument coherence, disclosure of uncertainty; Corner and Hahn, 2009) and personal attitudes towards the topic (Kahan et al., 2011; Kruglanski et al., 2005; Scurich & Shniderman, 2014).

In conclusion, our results strongly suggest that scientific authority is generally considered a reliable source for truth, more so than spiritual authority. Indeed, there are ample examples demonstrating that science serves as an important cue for credibility; the cover of Donald Trump's niece's family history book is adorned by "Mary L. Trump, PhD"; advertisements for cosmetic products often claim to be "clinically proven" and "recommended by dermatologists", and even the tobacco industry used to appeal to science (e.g., "more doctors smoke Camels than any other cigarette"). By systematically quantifying the difference between acceptance of statements by a scientific and spiritual authority in a global sample, this work addresses the fundamental question of how people trust what others say about the world.

## Appendix 7.A Additional Model Statistics

For each of the models included in the analyses, we calculated the intraclass correlation (ICC; proportion of the total variance that is accounted for by the clustering) and the explained variance (Bayesian $R^2$; proportion of the total variance that is accounted for by the effects). Explained variance was assessed using the `bayes_R2` function from the `rstantools` package (Gabry et al., 2020), based on the method described by Gelman et al. (2019). Explained variance is given separately for general $R^2$ (all common and varying effects included in the respective model) and for the marginal $R^2$ (the common effects only). The means and 95% credible intervals for each of the relevant models described in the main text are given in Supplementary Table 7.4.

### 7.A.1 MCMC diagnostics

To investigate convergence of the MCMC chains, we calculated split-$\hat{R}$ (Gelman et al., 2014) based on the rank-based method described in Vehtari et al. (2021). The smallest and largest $\hat{R}$ values were 0.99997 and 1.00040, respectively, indicating good within-chain convergence. The traceplots for these smallest and largest $\hat{R}$ values are shown in Supplementary Figure 7.6a and b.

The number of effective samples ($\hat{N}_{eff}$) was calculated per parameter to assess to what extent autocorrelation in the chains reduces the certainty of the posterior estimates (Geyer, 2011). Ideally, $\hat{N}_{eff}$ is as large as possible (Vehtari et al., 2021). The $\hat{N}_{eff}$ for each of the 107 estimated parameters is displayed in Supplementary Figure 7.6c. Note that $\hat{N}_{eff}$ can be larger than the the total number of iterations (in this case: $N = 30,000$) when the samples are anti-correlated or antithetical (Carpenter, 2018). The smallest $\hat{N}_{eff} = 24,210.67$ for the varying slope of the source-by-religiosity interaction for Croatia. For many parameters, $\hat{N}_{eff}$ is equal to the number of iterations or even higher. We therefore concluded that the effective sample size is sufficient for valid interpretation of the estimates and inference.

## Appendix 7.B Country Comparisons Across Datasets

To explore the country-level patterns in the source effect between both datasets, we assessed the correlation between the experimental source credibility effect in the primary dataset and the contrast of the trust ratings for scientists and traditional healers in the validation dataset per country. The raw observed relation as well as the relation between the modeled source effects are depicted in Supplementary Figure 7.7a and b. The plots do not suggest a strong correlation between source effects, which is corroborated by the evidence for the correlation: $\text{BF}_{+0} = 1.06$; $\text{BF}_{+0} = 0.97$ for the observed and estimated source effects, respectively. These Bayes factors imply *absence of evidence*, meaning that we cannot conclude whether or not the country-level source effects are related between the two datasets. The 95% credible intervals further support the uncertainty of the correlation: $\rho_{obs} = 0.17$ [-0.22,0.52]; $\rho_{est} = 0.15$ [-0.22,0.50]. We note however, that in addition to the uncertainty related to the small number of observations[2], caution is also warranted due to the difference in included

[2] These were the 24 countries from the main dataset minus China, for which no religiosity data was available in the validation dataset.

**Table 7.4:** Explained variance and intraclass correlation for all relevant models.

| | $R^2$ | | Marginal $R^2$ | | Intraclass correlation | |
|---|---|---|---|---|---|---|
| | Mean | 95% CI | Mean | 95% CI | Mean | 95% CI |
| Common Effect Models | | | | | | |
| Source Effect | 0.173 | [0.165, 0.182] | 0.076 | [0.060, 0.094] | 0.125 | [0.079, 0.198] |
| Source-by-Religiosity | 0.181 | [0.172, 0.190] | 0.081 | [0.062, 0.102] | 0.142 | [0.095, 0.213] |
| Processing Time | 0.107 | [0.099, 0.114] | 0.015 | [0.012, 0.020] | 0.147 | [0.091, 0.235] |
| Memory Performance | 0.098 | [0.090, 0.105] | 0.004 | [0.002, 0.006] | 0.128 | [0.078, 0.207] |
| Source Effect Trust | 0.229 | [0.226, 0.232] | 0.141 | [0.139, 0.143] | 0.110 | [0.089, 0.134] |
| Source-by-Religiosity Trust | 0.281 | [0.278, 0.284] | 0.133 | [0.110, 0.157] | 0.293 | [0.258, 0.332] |
| Varying Effects Models | | | | | | |
| Source Effect | 0.179 | [0.170, 0.187] | 0.077 | [0.058, 0.099] | 0.150 | [0.103, 0.220] |
| Source-by-Religiosity | 0.182 | [0.174, 0.191] | 0.082 | [0.064, 0.101] | 0.141 | [0.095, 0.212] |
| Processing Time | 0.108 | [0.100, 0.115] | 0.015 | [0.011, 0.020] | 0.152 | [0.097, 0.238] |
| Memory Performance | 0.099 | [0.091, 0.106] | 0.004 | [0.002, 0.006] | 0.134 | [0.085, 0.210] |
| Source Effect Trust | 0.281 | [0.278, 0.283] | 0.133 | [0.110, 0.157] | 0.296 | [0.261, 0.334] |

*Note.* Explained variance, split into general explained variance and marginal explained variance (fixed effects only), and intraclass correlations. The 95% CI gives the lower and upper bound of the credible interval. Note that there was no varying effect of the source-by-religiosity interaction for the trust model (validation dataset).

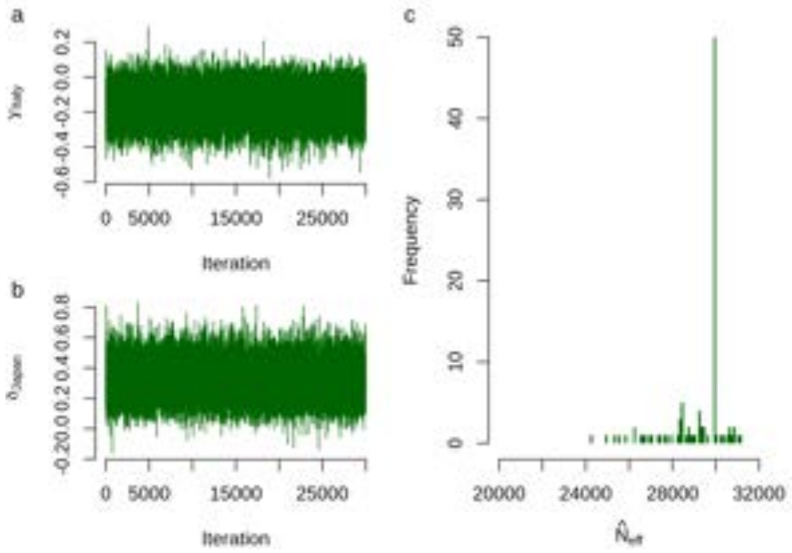


**Figure 7.6:** MCMC diagnostics. **a.** Chains for parameters with the smallest (varying slope for source effect in Italy) and **b.** largest (varying slope for the religiosity effect in Japan) $\hat{R}$ values. **c.** Number of effective samples for each parameter in the full model.

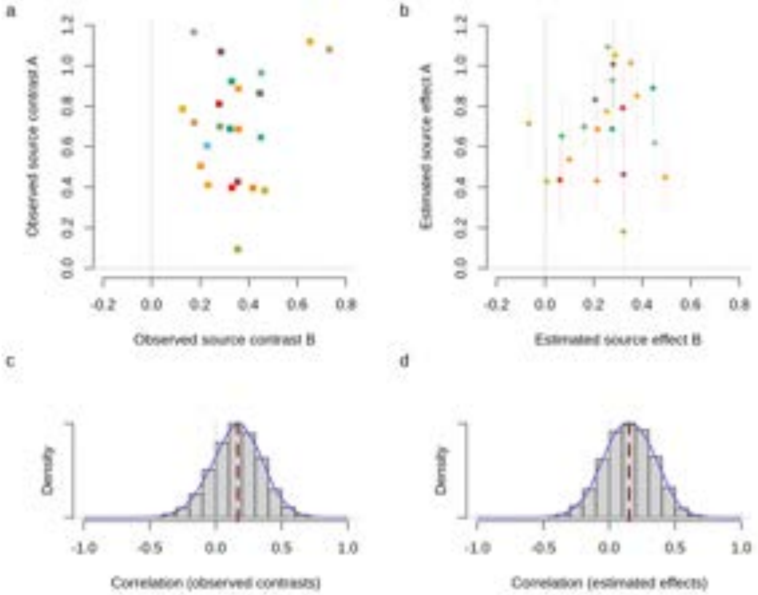


**Figure 7.7:** Correlation between the source effect in the new experimental dataset (set A) and the validation survey data on trust (set B). Panel **a** shows the relationship between the observed contrast effects (scientist minus guru) in both datasets. Each square represents a country. Panel **b** shows the country-level estimates (medians) of the source effect in the experimental dataset and the validation dataset. Each dot represents a country. The horizontal and vertical lines denote the 95% credible intervals. Panels **c** and **d** display the posterior distribution of the correlation coefficient $\rho$ using the observed contrasts and estimated effects, respectively. The vertical dashed line reflects the median value for $\rho$.

samples and exact items (credibility of specific nonsense statements vs. explicit trust in authorities) between datasets.

## Appendix 7.C A Note on Scientific Credibility and COVID-19

In the main paper, we included the case of COVID-19 only as a timely example to introduce our general topic, but we do not further elaborate on trust and credibility of authorities related to COVID-19 specifically. That is, we believe that our findings bear a broader and more general relevance for understanding source credibility-effects, that go beyond the current situation. Many others have investigated the perception of experts in relation to COVID-19 specifically in great detail, see for instance (Agley, 2020; Battiston et al., 2020; Funk, Kennedy, et al., 2020; Kreps & Kriner, 2020; Open Knowledge Foundation, 2020; Sibley et al., 2020; Wissenschaft im Dialog, 2020). While we do not discuss COVID-19 at length in the main paper, we quickly reflect here on the potential implications of these findings, using the Netherlands as an illustration.

The pattern found in the studies referred to above is somewhat mixed, yet most data seem to suggest that trust in science/scientists has either remained the same or even increased during the pandemic. In the Netherlands for instance, the majority of the general public also still places more trust in the Outbreak Management Team (OMT; a team of experts convened to advise the government on policy in the event of an outbreak of infectious disease) and RIVM (Dutch equivalent of the CDC) than Maurice de Hond or Willem Engel (Dutch public figures and self-declared COVID-19 experts). This is for instance indirectly indicated by increased vaccination willingness over the last months (about 80% in NL). Moreover, the public still mostly relies on information regarding vaccination provided by vaccination centers (60.6%), the RIVM website (48.1%) and GPs (39.6%), to a stronger extent than that provided by the media (34.8%), trusted celebrities (2.5%) or social media (2%; see www.rivm.nl/gedragsonderzoek/maatregelen-welbevinden/vaccinatiebereidheid). So while there are certainly individual differences in the perception of who is considered an expert, it seems that, on average, scientific expertise is still considered the most trustworthy source of information compared to other sources in relation to COVID-19 - and perhaps more generally as our study suggests.

# 8

# Mind-Body Dualism and Religion: An Investigation Across 24 Countries

People from a variety of cultures and ages reason dualistically, as they consider mental states (e.g., love, knowledge) more likely to continue after biological death than bodily states (e.g., hunger, hearing). It is unclear, however, to what extent the tendency for mind-body dualism is natural and intuitive. Using a large sample ($N$ = 10,195 participants) from 24 different countries, we replicated previous findings that people universally tend to reason dualistically. In addition, individual religiosity was predictive of more overall continuity beliefs and stronger mind-body dualism and a framing manipulation emphasizing a religious conception of death increased continuity judgments, though not mind-body dualism. At the same time, the modal response across the majority of countries and the aggregated sample was complete cessation of all states, and explicit afterlife beliefs were more prevalent than implicit afterlife beliefs. Based on these data, an intuitive materialism account, assuming a default conception that all mental activity ends at physical death, yet allowing for culturally acquired explicit afterlife beliefs, appears more plausible than an intuitive dualism account.

## 8.1 Introduction

The relationship between the human mind and body has intrigued philosophers and theologians for centuries. Over the last several decades, social scientists have joined the discussion by exploring laypeople's conceptualization of the mind. Many people nowadays endorse the neuroscientific view of the mind as a product of the physical brain (Berent & Platt, 2021; Riekki et al., 2013; Valtonen et al., 2021). At the same time, substantial research has shown that people from a variety of cultures and age

This chapter has been adapted from: Hoogeveen, S., Altay, S., Bendixen, T., Berniūnas, R., Bulbulia, J. A., Cheshin, A., Gentili, C., Georgescu, R., Haaf, J. M., Hagel, K., Kavanagh, C. M., Levy, N., McKay, R., Neely, A., Qiu, L., Rabelo, A., Ramsay, J. E., Ross, R. M., Turpin, H., … van Elk, M. (2022). *Does she still love and feel hungry? Afterlife beliefs, mind-body dualism, and religion across 24 countries* (Manuscript in preparation).

groups reason dualistically, as they treat psychological and biological states differently (Astuti & Harris, 2008; Bering, 2002, 2006; Bering & Bjorklund, 2004; Bloom, 2007; Chudek et al., 2013; Cohen et al., 2011; Forstmann & Burgmer, 2015; P. L. Harris & Giménez, 2005; Huang et al., 2013).

Philosophical thought experiments about body duplication and transfer have been used to probe how people think about the mind and the body (Parfit, 1984). This research has shown, among other things, that bodily features (e.g., having a scar) are more likely to be judged as transferable to a duplicate than mental states (e.g., remembering one's relatives; Forstmann and Burgmer, 2015; Hood et al., 2012). By contrast, in scenarios involving mind-switching (in which the mind of person A is transferred to the body of person B; Cohen and Barrett, 2008; Hood et al., 2012), pre-life settings (i.e., the existence of states prior to biological conception; Emmons and Kelemen, 2014), and afterlife settings (i.e., the continuation of states after biological death) mental states are rated as more likely to transfer than body-related states. More specifically, many people seem to implicitly believe that while states often classified as bodily such as hunger cease at death, high-level mental states such as love do not.[1] In the current chapter, we distinguish between implicit and explicit afterlife beliefs, in the sense that the implicit measures do not directly assess people's beliefs about the existence of an afterlife, a soul or disembodied spirits. As shown by the prevalent coexistence of seemingly incompatible beliefs about the nature of death (Astuti & Harris, 2008; P. L. Harris & Giménez, 2005; Legare et al., 2012; Watson-Jones et al., 2017), these implicit beliefs may exist even among people who explicitly reject an afterlife, such as atheists, or vice versa.

Mind-body dualism is typically measured by asking about the continuity of various processes of a deceased individual (Bering, 2002; P. L. Harris & Giménez, 2005). Participants read a vignette about a person who has recently died and are asked to what extent they think that person can, for instance, still experience hunger and pain (physical) or feel love and have memories (mental) after they have died (see Figure 8.1 for an example). Mind-body dualism is then operationalized as the relative difference in continuity judgments for mental and physical states. In the current study, we investigated the universality of laypeople's mind-body dualism, and its relation to religiosity. That is, we set out to conceptually replicate the main finding that mental states are more likely to be judged to continue after death than bodily states in a large cross-cultural sample ($N$ = 10,195 participants from 24 countries). In addition, we assessed to what extent individual religiosity of the rater and a contextual emphasis on religion influence both the tendency to make overall continuity judgments and to reason dualistically (i.e., to make more continuity judgements for mental states relative to biological states).

Several theories have been proposed to explain folk dualistic reasoning and the link with explicit afterlife beliefs. A basic premise of mind-body dualism is what H. C. Barrett et al. (2021) call the *parallel systems* account. According to this account,

[1] We note that hunger is arguably also a mental state. In the mind-body dualism paradigm, a distinction is typically made between states that are strongly body-dependent such as hunger, feeling pain, seeing or hearing and high-level mental states that are less closely linked to the physical body such as love, knowledge, desire. Throughout the literature various terms have been used to denote the body-dependent states, e.g., psychobiological, physiological, physical, body-related. Here, we use the term 'bodily states' to clearly contrast these body-dependent states with the high-order mental states, that are simply referred to as 'mental states'.

people can think about organisms in two distinct ways: as agents and as physical objects. By default, agency is attributed to living people and animals, but not to deceased ones (e.g., H. C. Barrett & Behne, 2005; H. M. Gray et al., 2007). The shift in focus between agency and body allows for the possibility to distinguish between a mind and a body, and reason about people's emotions and physical movements, respectively.

Some authors, however, have taken the mind-body distinction one step further (e.g., Bering, 2002; Bloom, 2005). Specifically, it is argued that humans are not only capable of distinguishing between a mind and a body, but that doing so is the cognitive default: dualistic beliefs about the mind and body as being separate entities are natural and innate. This has led to the influential 'intuitive dualism' approach. In other words, folk dualism might be a universal default that can be *unlearned* through formal education, rather than that it is *learned* through cultural narratives. Bloom (2005), for instance, argued that humans are intuitive Cartesian substance dualists: we intuit that the mind is separate from the body, that the mind is the sole source of our identity and that the body is no more than a vehicle for the mind. Bering (2002, 2006) introduced the 'simulation constraint hypothesis' as a potential causal mechanism for folk dualism: because it is impossible to imagine what it's like to be dead –specifically what it is like to be devoid of emotions and cognitions– we (mistakenly) assume that dead individuals still possess mental capacities.

However, other scholars have challenged the naturalness of mind-body dualism. They argue that the 'intentional stance' and 'offline social reasoning' provide more parsimonious accounts to explain empirical patterns of dualistic reasoning showing that mental states are judged as more likely to continue after death than physical states (Dennett, 2006; Hodge, 2008, 2011b). People use intentionality to reason about other individuals, including deceased ones and because of the focus on intentions and social relations, mental states are more likely to be thought of as continuing compared to physical states. This does not assume, however, that we intuitively view humans as disembodied minds. Relatedly, H. C. Barrett et al. (2021) and Barlev and Shtulman (2021) argued that the empirical patterns are in fact at odds with an intuitive dualism account: across most studies, the modal response of continuity judgments is cessation rather than continuation, even for high-level mental states such as love. Instead, an *intuitive materialism* account might be more appropriate, in which the default is to view death in biological terms upon which all mental activity ends. According to H. C. Barrett et al. (2021), the fact that a small group of people in these vignette studies do make continuity judgments for dead agents is a result of explicit afterlife beliefs that are culturally acquired. Barlev and Shtulman (2021) similarly argue that mind-body dualism observed in afterlife scenarios and the widespread belief in disembodied beings (ghosts, spirits, God etc.) result from learned rather than intuitive dualism. The prevalence of these beliefs is due to the social transmission advantage that stems from being (minimally) counterintuitive (cf. Banerjee et al., 2013; Boyer, 1994).

Indeed, various empirical patterns suggest that laypeople's dualistic reasoning involves considerable cultural scaffolding. First, several studies found that the tendency to distinguish between body and mind *increased* rather than decreased with age (Astuti & Harris, 2008; Bering, Hernández-Blasi, et al., 2005; Bering & Bjorklund, 2004; P. L. Harris & Giménez, 2005; Watson-Jones et al., 2017). Second, many of the intuitive dualism arguments are based on observations among young children, but even

5-year-olds have been exposed to considerable cultural discourse. As mentioned by P. L. Harris (2011b, p.37), children are more likely to be told that their dead grandmother or pet is still thinking about them and loves them, than that the emotional connection ceases at death. This might be similar to the cultural idea of the heart as being the place of love, which was demonstrated by children's belief that kindness would be transferred upon heart transplantation (C. N. Johnson, 1990). Still, based on this it would be mistaken to argue that the notion of the heart as the place of love is innate, in the sense of developing without substantial cultural guidance. Third, it has been shown that the framing (religious vs. secular) of a narrative influences the likelihood of stating that mental processes continue after death (Astuti & Harris, 2008; Bek & Lock, 2011; P. L. Harris & Giménez, 2005; Watson-Jones et al., 2017), again emphasizing the sensitivity of implicit afterlife beliefs and dualistic reasoning to cultural cues. Fourth, while mind-body dualism has been observed across various cultures (e.g., in the US, Madagascar, Brazil, Ecuador, Ukraine, Vanuatu, China; Astuti & Harris, 2008; Chudek et al., 2013; Cohen et al., 2011; Huang et al., 2013), substantial cultural differences in the categorization of different states have been documented (Huang et al., 2013; Weisman et al., 2021). A recent study, for instance, reported cross-cultural universality in reasoning about biological and cognitive states, but cultural variation in socio-emotional 'heart-like' states (Weisman et al., 2021). Additionally, following Bering's (2002) original 'dead person experiment', Huang et al. (2013) replicated the continuity of mental states and the cessation of psychobiological states in a Chinese sample (both immediately after and 2 days after the passing). However, in this study participants judged that auditory and visual perceptual states would continue, while they did not in the original study, indicating cross-cultural variation in these beliefs.

There are also systematic individual differences in the extent to which people reason dualistically. Perhaps most obvious is the link between religious beliefs and mind-body dualism; most religions involve some form of an afterlife that typically emphasizes continuity of the soul/spirit/mind of the deceased. Indeed, the assumed naturalness of mind-body dualism has been used as an argument to explain why religious beliefs are widespread and intuitive (Bering, 2006; Bloom, 2007). Empirical evidence also supports the link between religiosity and folk dualism, such that religious individuals are more likely to explicitly hold dualistic beliefs and make more continuity judgments about deceased people (i.e., display implicit afterlife beliefs; Riekki et al., 2013). Notably, religious individuals have been found to be even more likely to attribute mental capacities to deceased individuals than to living individuals in a vegetative state (K. Gray et al., 2011). At the same time, some studies have found that continuity judgments are even prevalent among atheists. For instance, over 50% of extinctivists (i.e., individuals who do not believe in an afterlife) judged high-level mental processes such as emotional and epistemic states to continue after death (Bering, 2002). In addition, atheists have also been found to hold explicit dualistic beliefs, albeit to a lesser degree than religious believers (T. A. Nelson et al., 2020). Finally, experimental manipulations aimed at investigating the role of culture and setting in folk dualism also capitalized on the relevance of religion; Astuti and Harris (2008) and P. L. Harris and Giménez (2005), for instance, found that continuity judgments occurred more often in response to a narrative involving religious burial rites than a narrative focused on a corpse. In addition, Watson-Jones et al. (2017) found that while in the US, religious framing enhanced continuity for both biological and psychological processes, in Vanuatu, an island nation in the South Pacific, religious framing mostly enhanced continuity judgments for biological processes specifically.

In sum, in the literature there is some evidence for the cross-cultural universality of implicit afterlife beliefs, the relation with an individual's religious beliefs (or the lack thereof) and the role of the framing of the narrative. In the current preregistered study, we aimed to replicate previous findings that (1) mental states are more likely to be judged to continue than bodily states ($\mathcal{H}_1$), that (2) individual religiosity is associated with increased continuity judgments ($\mathcal{H}_2$), and that (3) a framing manipulation emphasizing religious practices increases continuity judgments ($\mathcal{H}_3$). While there is preliminary evidence for these main effects, replication seems crucial, especially since previous studies were non-preregistered and only based on small samples (ranging from 46 to 260 adults)[2] and a few cultures (Hoogeveen & van Elk, 2021; Lindsay, 2015; Schmidt, 2009). Moreover, it is unclear how exactly religion as an individual difference factor or as a contextual manipulation is related to mind-body dualism (i.e., a state-by-religiosity interaction effect; $\mathcal{H}_4$ and a state-by-framing interaction effect; $\mathcal{H}_5$). That is, using the vignette approach by asking participants to make continuity judgments for both mental and physical states, religiosity might be associated with more continuity judgments uniformly across both mental and bodily states (cf. P. L. Harris & Giménez, 2005), relatively more continuity of mental states (vs. bodily states; increased dualism; cf. H. C. Barrett et al., 2021) or relatively more continuity of bodily states (i.e., reduced or no dualism; cf. Watson-Jones et al., 2017).

In addition, five complementary preregistered hypotheses were tested. First, we expected explicit afterlife beliefs to be positively related to implicit afterlife beliefs (i.e., overall continuity ratings; $\mathcal{H}_6$) and to mind-body dualism (i.e, a state-by-afterlife beliefs interaction; $\mathcal{H}_7$). Second, based on the work by Forstmann et al. (2012), we assessed mind-body dualism with a pictorial self-rating item showing two circles representing the mind and the body that are separate or overlapping to various degrees. We expected participants' ratings on this item to be positively related to mind-body dualism measured as the difference between continuity of mental and bodily states in the vignette ($\mathcal{H}_8$). Finally, while we expect some universality in the presence of folk dualism, the size of the mind-body difference might very well differ substantially across countries. Specifically, mirroring the religiosity effect at the individual level, we expected that the level of cultural religiosity within a country would be positively related to overall continuity beliefs ($\mathcal{H}_9$) and to the size of the state effect (i.e., the mental states vs. bodily states difference) in that respective country ($\mathcal{H}_{10}$).

We presented participants with a vignette describing a woman who had recently died. In a between-subjects manipulation, the death was either framed in religious terms featuring a religious authority (e.g., a priest) and references to an afterlife ('now that she's with God...') or in secular terms featuring a medical doctor and no further references ('now that she's dead...'; P. L. Harris and Giménez, 2005). Then we asked participants to judge the continuity of six states, three of which we classified as *bodily states* –feeling hungry, having an active brain, hearing– and three of which we classified as *mental states* –wanting, knowing, loving. We note that the literature is somewhat ambiguous about the categorization and evaluated continuity of perceptual

[2] Previous studies assessing afterlife continuity among adults included 84 (Bering, 2002), 46 (Astuti & Harris, 2008), 79 (Watson-Jones et al., 2017), and 260 participants (H. C. Barrett et al., 2021).

states such as seeing and hearing. For instance, Bering (2002) found that perceptual states were judged to cease, while in later work perceptions were among the cognitive states that were judged to continue in contrast to psychobiological states (Bering & Bjorklund, 2004). Using a more bottom-up approach based on interviews about continuity in hypothetical disembodiment scenarios, Cohen et al. (2011) suggested a categorization of body-dependent and body-independent processes, where perception is considered body-independent. This also fits with the findings of Huang et al. (2013), who found that perceptual states were judged to continue in a Chinese sample. Finally, Weisman et al. (2017), Weisman et al. (2021) proposed three categories of lay concepts of the mind: 'body-like', 'heart-like', and 'mind-like', which correspond to bodily versus social and emotional versus perceptual and cognitive states. An exploratory factor analysis showed that hearing mostly clustered with mind-like states, although not universally (Weisman et al., 2021). In our main analysis, we followed our preregistration based on the original distinction where perceptual states are considered bodily states (Bering, 2002). In addition, given the ambiguity in the literature, we conducted an exploratory analysis investigating the clustering of the six different states and ran a robustness check with the hearing item categorized as a mental state.

## 8.2 Disclosures

### 8.2.1 Data, materials, and preregistration

The current study was preregistered on the Open Science Framework; readers can access the preregistration, as well as all materials for the study, the anonymized raw and processed data (including relevant documentation), and the R code to conduct all analyses (including all figures), on the OSF (https://osf.io/3p78n/). Any deviations from the preregistration are highlighted in this manuscript.

### 8.2.2 Reporting

We report how we determined our sample size, all data exclusions, and all manipulations in the study. As this study is part of a larger cross-cultural data collection project (see Hoogeveen, Haaf, et al., 2022; Hoogeveen, Sarafoglou, Aczel, et al., 2022), we only describe measures relevant to the mind-body dualism sub-project.

### 8.2.3 Ethical approval

The study was approved by the local ethics committee at the Psychology Department of the University of Amsterdam (Project #2018-SP-9713). Additional approval was obtained from local IRBs at the Adolfo Ibáñez University (Chile), the Babes-Bolyai University (Romania), James Cook University (Singapore), Royal Holloway, University of London (UK), the University of Connecticut (US), and the Max Planck Society, as well as the Senate Department for Education, Youth and Family from the Ministry of Education in Berlin (Germany). All participants were treated in accordance with the Declaration of Helsinki.

## 8.3 Methods

### 8.3.1 Participants

In total, 10,535 participants completed the online experiment. Of these, 340 participants (3.23%) were excluded because they failed the attention check, leaving an analytic sample of $N = 10,195$ participants from 24 countries. Participants were recruited from university student samples, from personal networks, and from 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 a financial remuneration, the possibility for a reward through a raffle, course credits, or received no compensation. There were no a priori exclusion criteria; everyone over 18 years old could participate. Participants were forced to answer all multiple choice questions, hence there was no missing data. The countries were convenience-sampled (i.e., through personal networks), but were selected to cover 6 continents and include different ethnic majorities and religious majorities (Christian, Muslim, Hindu, Jewish, Eastern religions, as well as highly secular societies). See Table 8.1 for the descriptive statistics, method of recruitment and compensation per country and Table 8.2 for a breakdown of religious affiliations per country.

### 8.3.2 Sampling Plan

We preregistered a target sample size of $n = 400$ per country and 20-25 target countries. The preregistered sample size and composition allowed us to look at overall effects, effects within countries, and between countries. As we applied a Bayesian statistical framework, we needed a minimum of 20 countries to have sufficient data for accurate estimation in cross-country comparisons (Hox et al., 2012). However, we were mainly interested in overall effects - rather than effects for individual countries. With approximately 8,800 participants, we would have sufficient data to reliably estimate overall effects, especially since the state effect (mind vs. body) is within-subjects. We planned to terminate data collection on November 30th, 2019, but retained data from ten participants who completed the survey after this termination date.

### 8.3.3 Materials

The relevant variables for the current study were individual religiosity, target state category (mental state vs. bodily state), the manipulated framing of the narrative (secular vs. religious) and the binary continuity judgments for each state. Participant religiosity was measured using standardized items taken from the World Values Survey (WVS; World Values Survey, 2010), covering religious behaviours (institutionalized such as church attendance and private such as prayer/meditation), beliefs, identification, values, and denomination. Besides having high face-validity, these measures have been applied cross-culturally in other studies (Lindeman et al., 2015; Lun and Bond, 2013; Stavrova, 2015; see also Hoogeveen, Haaf, et al., 2022). A Bayesian reliability analysis using the `Bayesrel` package (Pfadt & van den Bergh, 2020) indicated good internal consistency of the religiosity measure, McDonald omega = 0.930 [0.927, 0.931] (all item-rest correlations > 0.61). All individual religiosity items were transformed

**Table 8.1:** Descriptive Statistics per Country

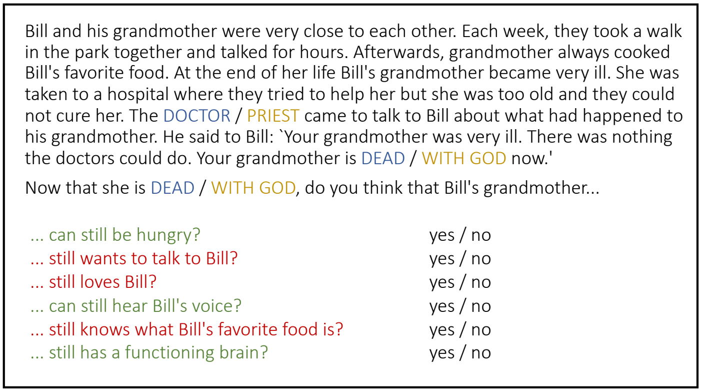
| Country | N | Age (SD) | Women | Religiosity | Sample | Compensation |
|---|---|---|---|---|---|---|
| Australia | 463 | 48.3 (16.0) | 48.4% | 0.52 | online panel | money |
| Belgium | 320 | 34.6 (13.1) | 55.6% | 0.24 | mixed | raffle |
| Brazil | 402 | 28.8 (10.4) | 73.1% | 0.51 | mixed | none; credits |
| Canada | 351 | 33.2 (10.5) | 52.4% | 0.28 | online panel | money |
| Chile | 308 | 30.8 (9.9) | 59.1% | 0.33 | mixed | raffle |
| China | 390 | 32.1 (8.4) | 55.9% | 0.32 | online panel | money |
| Croatia | 309 | 28.0 (6.9) | 78.3% | 0.41 | mixed | raffle |
| Denmark | 415 | 27.9 (10.3) | 71.3% | 0.26 | mixed | raffle |
| France | 405 | 40.6 (12.8) | 64.2% | 0.29 | online panel | money |
| Germany | 1,287 | 27.5 (9.0) | 62.2% | 0.32 | mixed | raffle |
| India | 394 | 30.4 (6.5) | 36.3% | 0.73 | online panel | money |
| Ireland | 434 | 42.6 (15.0) | 51.8% | 0.48 | online panel | money |
| Israel | 501 | 27.9 (10.1) | 73.5% | 0.37 | students | credits |
| Italy | 342 | 27.2 (8.2) | 50.9% | 0.26 | mixed | none; money |
| Japan | 424 | 40.6 (10.0) | 43.9% | 0.29 | online panel | money |
| Lithuania | 291 | 24.1 (7.0) | 83.2% | 0.35 | students | none |
| Morocco | 329 | 32.1 (11.8) | 16.1% | 0.70 | online panel | money |
| Netherlands | 482 | 57.6 (14.7) | 25.3% | 0.28 | online panel | money |
| Romania | 539 | 24.4 (7.4) | 85.2% | 0.55 | mixed | raffle |
| Singapore | 308 | 22.2 (3.4) | 62.0% | 0.45 | students | credits |
| Spain | 337 | 41.9 (13.9) | 31.2% | 0.21 | online panel | money |
| Turkey | 362 | 39.2 (11.1) | 24.6% | 0.33 | online panel | money |
| UK | 400 | 36.2 (12.7) | 65.8% | 0.23 | online panel | money |
| US | 402 | 35.8 (14.4) | 51.0% | 0.45 | mixed | none; money |
| Total | 10,195 | 33.8 (13.8) | 55.9% | 0.38 | - | - |

*Note.* Religiosity refers tot he self-reported level of individual religiosity based on 9 items, transformed on a 0-1 scale. Sample indicates the sample composition based on the method of recruitment per site.

**Table 8.2:** Religious Denomination per Country

| | Religious group | | | | | | |
|---|---|---|---|---|---|---|---|
| Country | Christian | Muslim | Hindu | Buddhist | Jewish | Other | None |
| Australia | 44.3% | 5.4% | 0.2% | 0.9% | 0.4% | 1.9% | 46.9% |
| Belgium | 28.4% | 2.5% | 0.0% | 0.6% | 0.3% | 0.6% | 67.5% |
| Brazil | 30.1% | 0.0% | 0.0% | 1.0% | 0.2% | 14.4% | 54.2% |
| Canada | 26.5% | 1.1% | 0.9% | 1.1% | 2.0% | 1.4% | 67.0% |
| Chile | 25.6% | 0.0% | 0.6% | 1.6% | 3.2% | 2.3% | 66.6% |
| China | 3.6% | 0.0% | 0.0% | 10.5% | 0.0% | 1.0% | 84.9% |
| Croatia | 54.4% | 0.3% | 0.0% | 0.6% | 0.3% | 0.6% | 43.7% |
| Denmark | 35.7% | 2.2% | 0.0% | 0.0% | 0.0% | 0.0% | 62.2% |
| France | 38.8% | 6.2% | 0.0% | 0.2% | 0.0% | 1.2% | 53.6% |
| Germany | 54.4% | 3.3% | 0.1% | 0.1% | 0.3% | 1.2% | 40.7% |
| India | 13.2% | 3.6% | 60.4% | 0.3% | 0.3% | 0.8% | 21.6% |
| Ireland | 54.4% | 1.6% | 0.2% | 0.0% | 0.2% | 0.9% | 42.6% |
| Israel | 2.2% | 3.2% | 0.0% | 0.0% | 11.6% | 2.0% | 81.0% |
| Italy | 17.5% | 0.0% | 0.0% | 0.9% | 0.0% | 0.0% | 81.6% |
| Japan | 0.9% | 0.2% | 0.0% | 15.3% | 0.0% | 1.2% | 82.3% |
| Lithuania | 39.2% | 0.0% | 0.3% | 0.0% | 0.0% | 0.7% | 59.8% |
| Morocco | 0.3% | 78.1% | 0.0% | 0.0% | 0.0% | 1.5% | 20.1% |
| Netherlands | 27.0% | 0.0% | 0.0% | 0.0% | 0.6% | 3.1% | 69.3% |
| Romania | 77.2% | 0.2% | 0.0% | 0.2% | 0.2% | 2.2% | 20.0% |
| Singapore | 20.5% | 4.9% | 3.9% | 20.5% | 0.0% | 5.2% | 45.1% |
| Spain | 39.8% | 0.0% | 0.0% | 0.0% | 0.0% | 1.2% | 59.1% |
| Turkey | 0.0% | 42.5% | 0.0% | 0.0% | 0.3% | 2.5% | 54.7% |
| UK | 22.2% | 0.5% | 0.8% | 0.5% | 0.8% | 1.0% | 74.2% |
| US | 44.0% | 1.2% | 0.5% | 0.2% | 3.2% | 3.0% | 47.8% |
| Total | 32.0% | 5.7% | 2.6% | 2.0% | 1.0% | 2.0% | 54.6% |

*Note.* Percentage of people indicating membership of the respective religious groups. Note that the response options were particularized per country. Here we show the 5 most prevalent groups.



Bill and his grandmother were very close to each other. Each week, they took a walk in the park together and talked for hours. Afterwards, grandmother always cooked Bill's favorite food. At the end of her life Bill's grandmother became very ill. She was taken to a hospital where they tried to help her but she was too old and they could not cure her. The DOCTOR / PRIEST came to talk to Bill about what had happened to his grandmother. He said to Bill: `Your grandmother was very ill. There was nothing the doctors could do. Your grandmother is DEAD / WITH GOD now.'

Now that she is DEAD / WITH GOD, do you think that Bill's grandmother...

| | |
|---|---|
| ... can still be hungry? | yes / no |
| ... still wants to talk to Bill? | yes / no |
| ... still loves Bill? | yes / no |
| ... can still hear Bill's voice? | yes / no |
| ... still knows what Bill's favorite food is? | yes / no |
| ... still has a functioning brain? | yes / no |

**Figure 8.1:** The mind-body dualism narrative as used in the present study. The framing (religious framing indicated in yellow, secular framing in blue) was varied between participants. The states (mental states indicated in red, physical states indicated in green) were presented in randomized order. The name of the target person and the specific religious authority were adjusted to the language and cultural context of each country.

on a 0-1 scale (to make each item contribute equally to the scale), tallied to create a religiosity score per participant, and grand-mean standardized for the analyses. The experimental stimuli consisted of a short narrative about a young person whose grandmother dies (see Figure 8.1). The framing was manipulated (between-subjects) by either introducing a priest (or comparable religious authority) or a doctor to mention the grandmother's death and stating that she is either *with God now* or *dead now*, respectively. Participants then indicated whether they thought that the grandmother was still capable of (1) *being hungry*, (2) *hearing voices*, still had (3) *a functioning brain*, still could (4) *know things*, (5) *love*, and (6) *want things*. The first three processes were classified as bodily states (psychobiological/perceptual) and the last three as mental states (emotional/cognitive). The narratives and process items were based on the materials used by P. L. Harris and Giménez (2005). The name of the target person and the specific religious authority were adjusted to the language and cultural context of each country (e.g., a priest, a rabbi, an imam).

For the complementary hypotheses we additionally used the item on afterlife beliefs from the religiosity scale ('To what extent do you believe in a life after death?'), a pictorial dualism self-rating item, and two items assessing cultural norms of religiosity in one's country. The pictorial dualism item was taken from Forstmann et al. (2012), which was adjusted from the self-other inclusion scale by Aron et al. (1992). The self-rating item had seven response options, showing two circles representing the mind and the body that are separate or overlapping to various degrees. The cultural norms items assessed participants' perception of the importance of religious beliefs and behaviors for the average citizen in their country. See the online Appendix for the full materials, including the pictorial dualism item.

### 8.3.4 Procedure

Participants received a link to the Qualtrics survey, either by email, social media or through an online platform. After reading the instructions and providing informed consent, they first completed items for a separate study about religiosity and trustworthiness and source credibility for spirituality and science (see Hoogeveen, Haaf, et al., 2022)[3]. Subsequently, they were presented with the short narrative in either the religious or secular context, provided continuity judgments for the six process items, and completed the manipulation check to validate that they recalled the type of authority (religious vs. medical). Finally, they provided demographics, a quality of life scale, the religiosity items, and were given the opportunity to provide comments. It took about 10 minutes to complete the entire survey (median completion time was 11.4 minutes).

### 8.3.5 Data Analysis

Analyses were carried out in R[4]. The models were built using the package `brms` (Bürkner, 2017), which relies on the Stan language (Carpenter et al., 2017). The `bridgesampling` package (Gronau et al., 2020) was used to estimate the log marginal likelihood of the models of interest and calculate Bayes factors. The multilevel Bayesian modelling approach allows us to systematically evaluate the evidence in the data under different models: (i) in every country the effect is truly null; (ii) all countries share a common nonzero effect; (iii) countries differ, but all effects are in the same (predicted) direction; and (iv) in some countries the effect is positive whereas in others the effect is negative (Haaf & Rouder, 2017; Rouder, Haaf, Davis-Stober, et al., 2019). The models differ in the extent to which they constrain their predictions, from the most constrained (i) to completely unconstrained (iv). We refer to these models as the null model, the common effect model, the positive effects model, and the unconstrained model, respectively. Note that while the predictions from model (iii) are less constrained than those from model (ii), it is more difficult to obtain evidence for small effects under model (iii) because it assumes that the effect is present in every

[3] We acknowledge that these preceding items may have affected participants in their response to the current vignettes. However, we consider it unlikely that religion was strongly primed by these items, as they solely involved subtle cues of religion (i.e., an image of a woman wearing a necklace with a religious symbol and a statement by a spiritual guru). Any questions probing participants' religiosity directly were presented after the mind-body dualism task

[4] For all analyses, we used R (Version 4.0.2; R Core Team, 2020) and the R-packages *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2018), *bayesplot* (Version 1.8.0; Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019), *brms* (Version 2.14.4; Bürkner, 2017, 2018), *cmdstanr* (Version 0.3.0.9000; Gabry & Češnovar, 2020), *coda* (Version 0.19.4; Plummer, Best, Cowles, & Vines, 2006), *corrplot2017* (Wei & Simko, 2017), *curl* (Version 4.3; Ooms, 2019), *digest* (Version 0.6.27; Antoine Lucas et al., 2020), *dplyr* (Version 1.0.5; Wickham, François, Henry, & Müller, 2021), *ggplot2* (Version 3.3.3; Wickham, 2016), *gridExtra* (Version 2.3; Auguie, 2017), *invgamma* (Version 1.1; Kahle & Stamey, 2017), *MASS* (Version 7.3.53; Venables & Ripley, 2002), *Matrix* (Version 1.3.2; Bates & Maechler, 2021), *MCMCpack* (Version 1.5.0; Martin, Quinn, & Park, 2011), *msm* (Version 1.6.8; Jackson, 2011), *mvtnorm* (Version 1.1.1; Genz & Bretz, 2009), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *posterior* (Version 0.1.3; Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2020), *Rcpp* (Version 1.0.6; Eddelbuettel & François, 2011; Eddelbuettel & Balamuta, 2018), *rethinking* (Version 2.13; McElreath, 2020), *rstan* (Version 2.21.3; Stan Development Team, 2020a), *scales* (Version 1.1.1; Wickham & Seidel, 2020), *StanHeaders* (Version 2.21.0.7; Stan Development Team, 2020b), *tidyr* (Version 1.1.3; Wickham, 2020), *tinylabels* (Version 0.1.0; Barth, 2020), and *wesanderson* (Version 0.3.6; Ram & Wickham, 2018).

country, rather than only in the aggregate sample. When applied to our hypothesis for the mental versus physical state effect, evidence for (i) would indicate that people from these 24 countries do not differentially evaluate continuity of physical and mental states after death, evidence for (ii) would indicate that on average people from these 24 countries consider mental states more likely to continue than physical states (or vice versa), evidence for (iii) would indicate that in all of the 24 countries, people consider mental states more likely to continue than physical states (or vice versa), but there is cultural variation in the size of this effect, and evidence for (iv) would indicate that in some countries people consider mental states more likely to continue than physical states, and in other countries people consider physical states more likely to continue than mental states, indicating cultural variation in the direction (and size) of the effect.

For the full model including all main effects and relevant interaction effects, we specified the following unconstrained model. Let $Y$ denote the continuity responses per participant aggregated over the three binary items per state, where 0 indicates discontinuity and 1 indicates continuity and $Y = 0, \ldots, 3$. Further, let $Y_{ijkl}$ be the continuity judgment for the $i$th participant, $i = 1, \ldots, N$, in the $j$th country, $j = 1, \ldots, 24$, for the $k$th state category, $k = 1, 2$ (physical or mental states, respectively), and the $l$th framing condition, $l = 1, 2$ (secular or religious framing, respectively). The responses $Y_{ijkl}$ are modeled using an aggregated binomial model with a logit link to transform probabilities into real numbers $\in (-\infty, \infty)$:

$$Y_{ijkl} \overset{ind}{\sim} \text{Binomial}(3, p_{ijkl}),$$

$$\text{logit}(p_{ijkl}) = \alpha_j + x_k \beta_j + u_i \delta_j + c_l \gamma_j + v_{ki} \theta_j + w_{kl} \zeta_j.$$

where $\text{logit}(p_{ijkl})$ is the combined effect of observations, countries, and state categories on the tendency to indicate ‘continues.’ Note that $\text{logit}(p_{ijkl}) = 0$ reflects a probability of 0.5 of indicating continuity. The term $\alpha_j$ serves as the baseline continuity intercept for the $j$th country. The indicator $x_k = -0.5, 0.5$ if $k = 1, 2$, respectively, where $k = 1$ indicates the physical state condition and $k = 2$ indicates the mental state condition. The term $\beta_j$ is the $j$th country’s main effect of state category on continuity judgments. The variable $u_i$ gives the $i$th participant’s standardized religiosity score and $\delta_j$ is the $j$th country’s main effect of religiosity. The indicator $c_l = -0.5, 0.5$ if $l = 1, 2$, respectively, where $l = 1$ indicates the secular framing condition and $l = 2$ indicates the religious framing condition. The term $\gamma_j$ is then the $j$th country’s main effect of framing. The indicator $v_{ki}$ gives the state-by-religiosity interaction term and $\theta_j$ is the corresponding interaction effect for the $j$th country. Finally, indicator $w_{kl}$ gives the state-by-framing interaction term and $\zeta_j$ is the corresponding interaction effect for the $j$th country.

## 8.4 Results

### 8.4.1 Descriptive results

On average, people made continuity judgments for 31.76% of the states, with 16.13% for physical states and 47.39% for mental states. In Figure 8.4A these observed rates are further unpacked per framing condition and level of religiosity. Additionally,

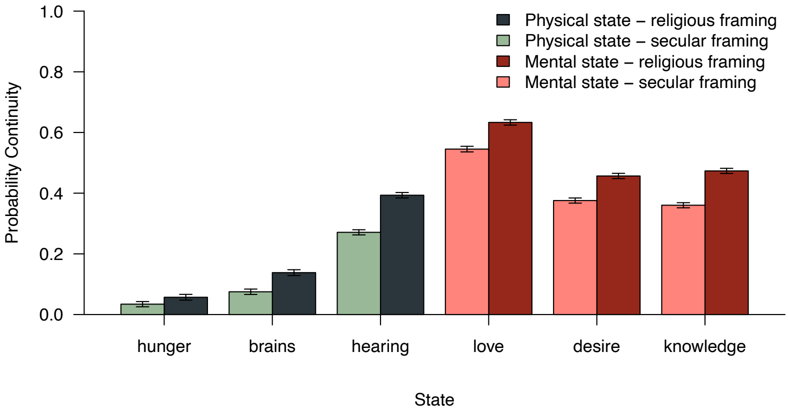


**Figure 8.2:** Descriptive pattern per state. Probability of continuity judgment per item, displayed per framing condition. The states were measured within-subjects and the framing was manipulated between-subjects. Error bars reflect the 95% confidence interval.

60.95% of participants judged at least one state to continue after death, while 2.01% reported all six states to continue.

At the same time, the modal response across most countries is complete cessation rather than continuity: in only 5 out of 24 countries, the modal sum score across the six items was either 3 or 4, in all other countries it was 0 (see Figure 8.5). Specifically, only in China, India, Japan, Romania, and Singapore were participants more likely to indicate continuity of some states than complete cessation of all states. Across the aggregated sample, the mode is also complete cessation.

In addition, the proportion of people that display implicit afterlife beliefs (i.e., rated continuity of states after death) in the absence of explicit afterlife beliefs (i.e., self-reported belief in life after death) is much smaller than the proportion of people endorsing explicit afterlife beliefs but implicitly rejecting continuity in an afterlife. That is, most people in the sample indicate that they at least somewhat believe in an afterlife (i.e., score $> 1$ on the 7-point Likert scale) and rate at least one state to continue in the narrative task (55.2%). Additionally, 19.6% of participants both explicitly and implicitly reject the possibility of an afterlife. Then there are 19.4% who explicitly state that they somewhat believe in an afterlife, but implicitly reject continuity of any states. Yet only 5.8% of participants explicitly reject an afterlife but implicitly allow for states to continue after death.

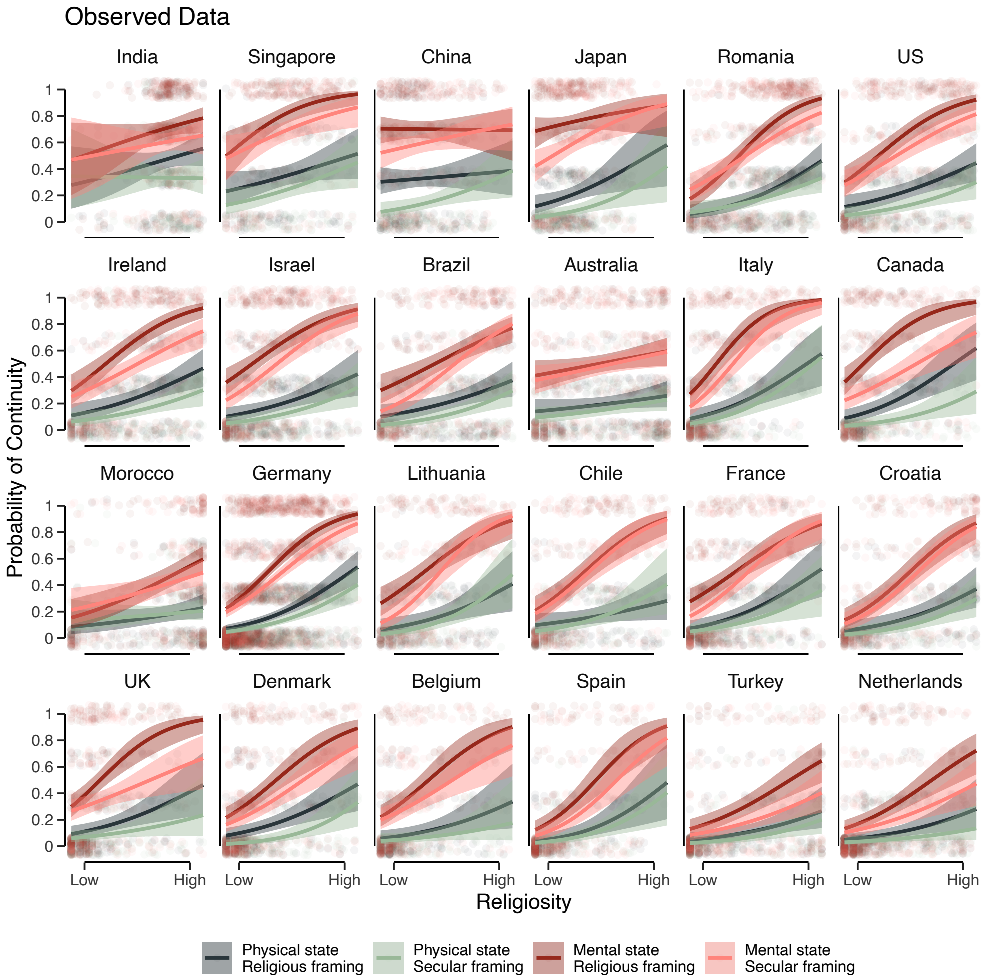


**Figure 8.3:** Descriptive pattern of results per country. Countries are ordered by the overall probability of making a continuity judgment (from left to right, top to bottom). Dark red lines denote probabilities for mental states in the religious framing condition, orange lines denote probabilities for mental states in the secular framing condition, dark blue lines denote probabilities for the physical states in the religious framing condition, and green lines denote probabilities for physical states in the secular framing condition. The shaded bands around the lines denote the 95% confidence interval. Data points are jittered to enhance visibility. Probabilities are averaged over the three items per state category.

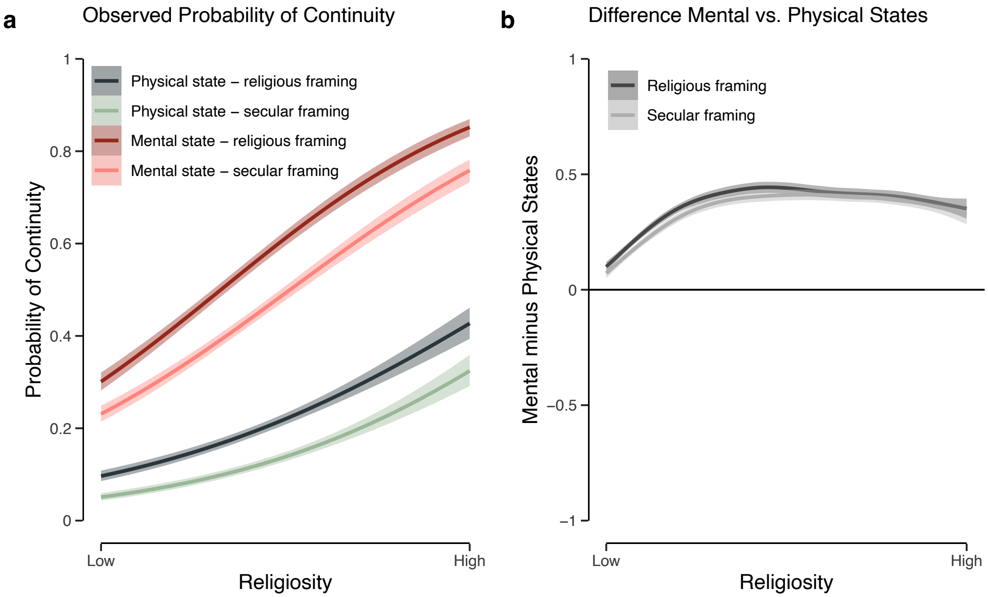


**Figure 8.4:** Descriptive pattern of results. Panel **a.** displays the probability of making a continuity judgment per state category (physical vs. mental), framing (secular vs. religious) and individual level of religiosity. Panel **b.** shows the observed difference in probability of mental vs. physical processes (state effect) for each level of religiosity and framing conditions. That is, a positive score on the y-axis indicates higher continuity attributed to mental states compared to physical states. At all levels of religiosity, continuity is more likely to be attributed to mental states than physical states, though the difference increases with higher religiosity. The shaded bands around the lines denote the 95% confidence interval.

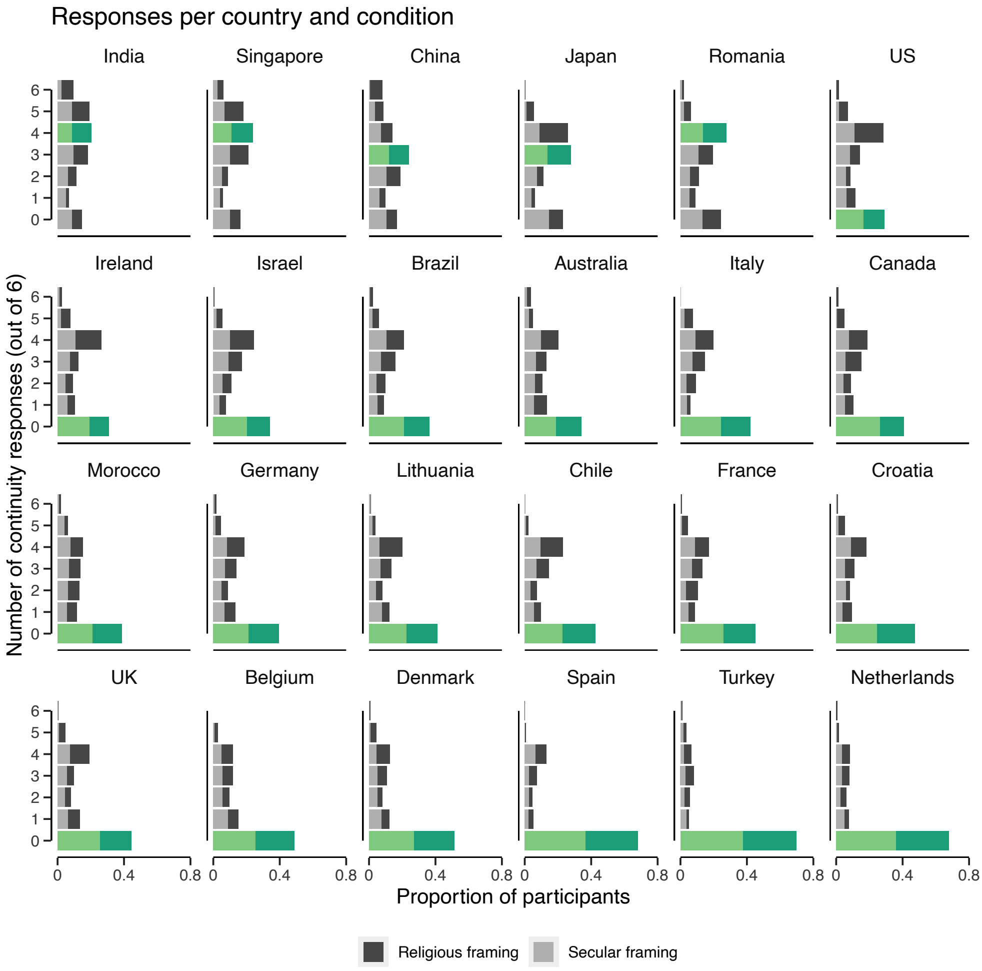


**Figure 8.5:** Proportion of participants and number of continuity responses per country. Countries are ordered by the overall probability of making a continuity judgment (from left to right, top to bottom). Dark grey bars reflect responses for the religious framing condition and light grey bars reflect responses for the secular framing condition. The modal number of continuity responses per country is indicated in green. Continuity responses were out of 6 states.

### 8.4.2 Confirmatory results

As can be seen in Table 8.3, we found substantial evidence in favor of our hypotheses for the state effect ($\mathcal{H}_1$), the religiosity effect ($\mathcal{H}_2$), the framing effect ($\mathcal{H}_3$), and the state-by-religiosity interaction effect ($\mathcal{H}_4$), yet strong evidence against the state-by-framing effect ($\mathcal{H}_5$).

First, regarding the state effect, mental processes are judged as more likely to continue after death than psychobiological processes, to a varying degree across countries: $\text{BF}_{+0} = \infty$; $\text{BF}_{+1} = 10^{26}$, $\mu_\beta = 1.71$ [1.55, 1.86], $\sigma_\beta = 0.35$ [0.25, 0.50]. This effect translates into an increase of 0.326 [0.129, 0.513] on the probability scale. Second, religiosity is positively associated with continuity judgments, to a varying degree across countries: $\text{BF}_{+0} = \infty$; $\text{BF}_{+1} = 10^{87}$, $\mu_\delta = 0.84$ [0.71, 0.96], $\sigma_\delta = 0.28$ [0.21, 0.39]. In other words, the most religious participants are 46.5% [13.9%, 72.0%] more likely to make continuity judgments than the least religious participants. Third, people are more likely to make continuity judgments when framed in a religious context than in a secular (medical) context, to a varying degree across countries: $\text{BF}_{+0} = 10^{146}$; $\text{BF}_{+1} = 10^{11}$, $\mu_\gamma = 0.52$ [0.41, 0.61], $\sigma_\gamma = 0.22$ [0.15, 0.32]. That is, people are 9.8% [0.9%, 21.4%] more likely to make continuity judgments in the religious framing condition than in the secular framing condition. Fourth, the difference in continuity judgments between mental and physical states becomes larger with increased religiosity, to a varying degree across countries: $\text{BF}_{10} = 10^{16}$; $\text{BF}_{+1} = 3143$, $\mu_\theta = 0.24$ [0.14, 0.33], $\sigma_\theta = 0.18$ [0.11, 0.28]. That is, overall, the most religious participants make an estimated 43.4% [23.2%, 57.8%] more continuity judgments about mental processes than about physical processes, while this difference is only 17.2% [3.9%, 41.8%] for the least religious participants. Note, however, that while the model comparison indicated substantial evidence for the interaction effect, the unconstrained model slightly outperforms the positive-effects model: $\text{BF}_{u+} = 1.19$. This is due to the fact that when looking at the countries separately, for 7 of them, the credible interval of the interaction effect includes zero (see Figure 8.6). Fifth, the difference in continuity judgments between mental and physical states is not larger in a religiously-framed than in a secularly-framed context: $\text{BF}_{01} = 40.34$, $\mu_\zeta = $ -0.09 [-0.19, 0.00], $\sigma_\zeta = 0.08$ [0.00, 0.22].[5]

[5] In Appendix 8.C, we additionally report exploratory analyses on the religiosity-by-framing interaction and the three-way state-by-religiosity-by-framing interaction effects. However, the data do not indicate substantial evidence for either of these interaction effects.

**Table 8.3:** Bayes factor model comparison and parameter estimates for the key effects

| | Bayes factors | | | | Parameter estimates | |
|---|---|---|---|---|---|---|
| Effect | $\mathcal{M}_0$ | $\mathcal{M}_1$ | $\mathcal{M}_+$ | $\mathcal{M}_u$ | $\mu$ | $\sigma$ |
| State Effect | 0.00 | 0.00 | **1.00** | 0.09 | 1.71 [1.55, 1.86] | 0.35 [0.25, 0.50] |
| Religiosity Effect | 0.00 | 0.00 | **1.00** | 0.08 | 0.84 [0.71, 0.96] | 0.28 [0.21, 0.39] |
| Framing Effect | 0.00 | 0.00 | **1.00** | 0.09 | 0.52 [0.41, 0.61] | 0.22 [0.15, 0.32] |
| State-by-Religiosity Effect | 0.00 | 0.00 | 0.84 | **1.00** | 0.24 [0.14, 0.33] | 0.18 [0.11, 0.28] |
| State-by-Framing Effect | **1.00** | 0.02 | 0.00 | 0.11 | -0.09 [-0.19, 0.00] | 0.08 [0.00, 0.22] |

*Note.* The preferred model for each effect is assigned value 1.00 and displayed in bold. The remaining values are the Bayes factors for the respective model relative to this preferred model. Subscripts reflect constraints on the critical parameter; $_0$ indicates no effect, $_1$ indicates a common (positive) effect, $_+$ indicates a varying positive effect, and $_u$ indicates an unconstrained effect. Parameter estimates (median and 95% credible interval) are taken from the unconstrained model for $\mathcal{H}_5$. $\sigma$ reflects the between-country variation in the respective effect.

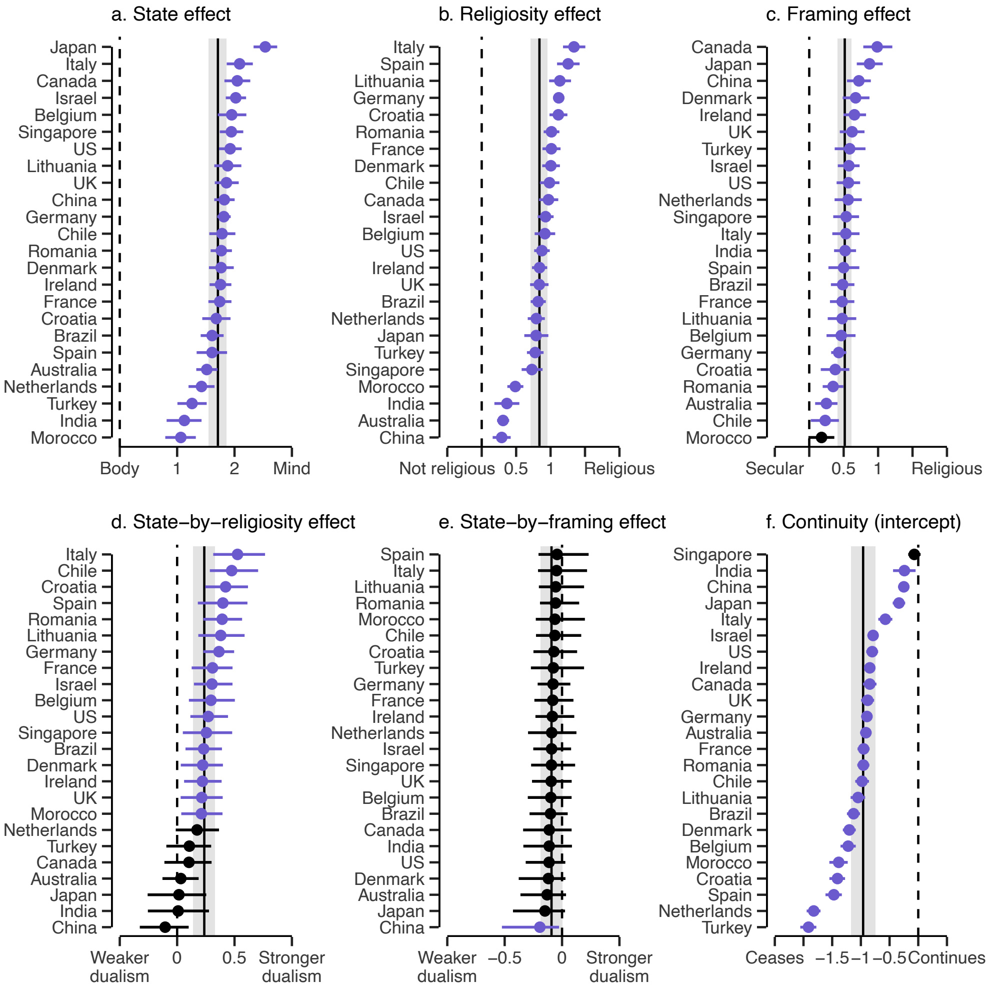


**Figure 8.6:** Estimated country-level effects (posterior medians) in increasing order. a. state contrast effects. b. religiosity effects. c. framing effects. d. state-by-religiosity interaction effects. e. state-by-framing interaction effects. f. intercepts. Each dot represents a country. Estimates with credible intervals colored in purple exclude zero and estimates with credible intervals colored in black include zero. The errorbars give the 95% credible interval for each country. The vertical lines denote the posterior median of the overall mean of the respective effect with the 95% credible interval in the shaded bands. The dashed lines indicates zero.

#### 8.4.2.1 Explicit afterlife beliefs

To test the hypothesis that explicit afterlife beliefs are related to both overall continuity judgments (i.e., implicit afterlife beliefs) and mind-body dualism, we constructed the models used to test $\mathcal{H}_2$ with the item on afterlife beliefs as the predictor. The Bayes factor analysis provided strong evidence for $\mathcal{H}_6$ that explicit afterlife beliefs are positively related to the overall probability of making continuity judgments, to a varying degree across countries ($\text{BF}_{+0} = \infty$; $\text{BF}_{+1} = 10^{96}$ , $\mu_\delta = 0.90$ [0.78, 1.01], $\sigma_\delta = 0.28$ [0.21, 0.39]). In addition, afterlife beliefs were also related to the tendency to differentiate between mental and physical states (i.e, $\mathcal{H}_7$), to a varying degree across countries ($\text{BF}_{+0} = 10^{12}$; $\text{BF}_{+1} = 3.01$, $\mu_\theta = 0.19$ [0.13, 0.26], $\sigma_\theta = 0.10$ [0.03, 0.18])

#### 8.4.2.2 Pictorial dualism item

As preregistered, we also assessed whether a pictorial dualism self-rating item predicted overall continuity ratings and mind-body dualism. The Bayes factor model comparison gave evidence against the pictorial item predicting mind-body dualism operationalized as the difference in continuity between mental and physical states: $\text{BF}_{10} = 0.07$; $\text{BF}_{01} = 13.76$, $\mu_\theta = 0.02$ [-0.03, 0.07], $\sigma_\theta = 0.06$ [0.00, 0.12]).

#### 8.4.2.3 Country-level cultural norms

Mirroring the religiosity effect at the individual level, we expected a positive relation between both the overall continuity judgments and cultural norms of religion and between mind-body dualism operationalized as the state effect and cultural norms of religion. To investigate this effect, we correlated cultural norms aggregated at the country-level with country-level estimates of the intercepts ($\alpha_j$) and state-effects ($\beta_j$) in the models. First, we find some weak evidence against a positive correlation between the country-level overall probability of continuity and cultural norms of religion: $\text{BF}_{+0} = 0.32$; $\text{BF}_{0+} = 3.09$. Second, we obtained moderate evidence against a positive correlation between country-level estimates of dualism (i.e., the state effect) and cultural norms of religiosity aggregated at the country-level: $\text{BF}_{+0} = 0.13$; $\text{BF}_{0+} = 7.66$ (see Figure 8.7). In fact, if anything, the correlation appears to be negative, rather than positive; the estimated size of the correlation coefficient is -0.48 [-0.71, -0.13].[6] This suggests that participants from countries where religion is more normative are *not* more likely to make continuity judgments or reason dualistically. Instead, in more religious countries, people may be less likely to distinguish between physical and mental states.

#### 8.4.2.4 Robustness checks

Here, we report the results of five alternative analysis choices that the results should be robust against. First, based on the ambiguity in the literature and the results of the exploratory factor analysis (see exploratory results below), we classified 'hearing' as a mental state rather than a bodily state. Second, we reran the analyses excluding

[6] If we release the directional constraint, we get strong evidence in favor of a correlation: $\text{BF}_{10}$= 15.69.

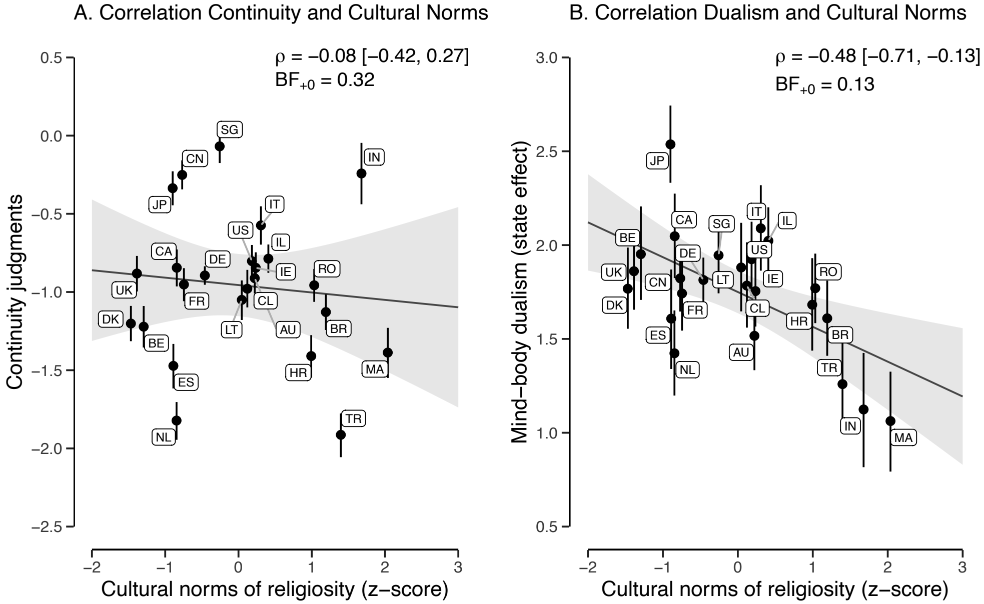


**Figure 8.7:** Correlation between country-level cultural norms of religion and continuity judgments (panel A.) and mind-body dualism (i.e., state effects; panel B.).

participants who failed to correctly identify a target figure mentioned in the narrative (i.e., a priest or a doctor). Third, we included level of education and age as covariates in the analyses, which were identified as potential confounding variables that warranted inclusion in the statistical models (see the online Appendix). Fourth, we preregistered a lower limit of 300 participants per country and hence reran the analyses while excluding data from Lithuania since $n = 291$. Fifth, we conducted an additional check with the (suboptimal) preregistered prior settings. That is, in the preregistration, we specified half-cauchy priors on the standard deviation. However, the prior predictive checks showed that the fat tails of the cauchy distribution resulted in implausible predictions on the probability scale (see Appendix 8.A for details). Following recommendations by McElreath (2020) and Betancourt et al. (2015), we used the half-normal(0,1) prior on the country-level standard deviation instead. This resulted in more reasonable prior predictions (see Appendix 8.A). As shown in Table 8.4, the results are qualitatively equal across the different robustness checks: we obtained strong support for a varying positive effect of state ($\mathcal{H}_1$), religiosity ($\mathcal{H}_2$), framing ($\mathcal{H}_3$), and a state-by-religiosity interaction ($\mathcal{H}_4$), but strong evidence against a state-by-framing interaction ($\mathcal{H}_5$).

**Table 8.4:** Bayes factor of different models for robustness checks

| Robustness set | $\mu$ [95% CI] | $\text{BF}_{10}$ | $\text{BF}_{+1}$ | Preferred |
|---|---|---|---|---|
| **State Effect** | | | | |
| Main analysis | 1.71 [1.55, 1.86] | $\infty$ | $10^{26}$ | $\mathcal{M}_+$ |
| Hearing as mental state | 1.16 [1.04, 1.27] | $\infty$ | $10^{11}$ | $\mathcal{M}_+$ |
| Excluding manipulation check failures | 1.75 [1.59, 1.88] | $\infty$ | $10^{20}$ | $\mathcal{M}_+$ |
| Education and age as covariates | 1.73 [1.57, 1.88] | $\infty$ | $10^{26}$ | $\mathcal{M}_+$ |
| Excluding Lithuania[a] | 1.70 [1.54, 1.85] | $\infty$ | $10^{26}$ | $\mathcal{M}_+$ |
| Cauchy$^+$(0,2) prior on $SD$[a] | 1.71 [1.54, 1.86] | $\infty$ | $10^{25}$ | $\mathcal{M}_+$ |
| **Religiosity Effect** | | | | |
| Main analysis | 0.84 [0.71, 0.96] | $\infty$ | $10^{87}$ | $\mathcal{M}_+$ |
| Hearing as mental state | 0.88 [0.75, 1.01] | $\infty$ | $10^{116}$ | $\mathcal{M}_+$ |
| Excluding manipulation check failures | 0.85 [0.72, 0.96] | $\infty$ | $10^{75}$ | $\mathcal{M}_+$ |
| Education and age as covariates | 0.86 [0.73, 0.97] | $\infty$ | $10^{80}$ | $\mathcal{M}_+$ |
| Excluding Lithuania[a] | 0.83 [0.70, 0.94] | $\infty$ | $10^{85}$ | $\mathcal{M}_+$ |
| Cauchy$^+$(0,2) prior on $SD$[a] | 0.84 [0.71, 0.96] | $\infty$ | $10^{87}$ | $\mathcal{M}_+$ |
| **Framing Effect** | | | | |
| Main analysis | 0.52 [0.41, 0.61] | $10^{135}$ | $10^{11}$ | $\mathcal{M}_+$ |
| Hearing as mental state | 0.56 [0.45, 0.66] | $10^{174}$ | $10^{16}$ | $\mathcal{M}_+$ |
| Excluding manipulation check failures | 0.52 [0.41, 0.63] | $10^{127}$ | $10^{14}$ | $\mathcal{M}_+$ |
| Education and age as covariates | 0.51 [0.41, 0.61] | $10^{133}$ | $10^{10}$ | $\mathcal{M}_+$ |
| Excluding Lithuania[a] | 0.52 [0.41, 0.62] | $10^{132}$ | $10^{11}$ | $\mathcal{M}_+$ |
| Cauchy$^+$(0,2) prior on $SD$[a] | 0.51 [0.41, 0.62] | $10^{135}$ | $10^{10}$ | $\mathcal{M}_+$ |
| **State-by-Religiosity Effect** | | | | |
| Main analysis | 0.24 [0.14, 0.33] | $10^{16}$ | 3127 | $\mathcal{M}_u$ |
| Hearing as mental state | 0.14 [0.05, 0.22] | $10^{5}$ | 7.66 | $\mathcal{M}_u$ |
| Excluding manipulation check failures | 0.24 [0.14, 0.33] | $10^{15}$ | 180 | $\mathcal{M}_+$ |
| Education and age as covariates | 0.24 [0.14, 0.34] | $10^{17}$ | 6360 | $\mathcal{M}_u$ |
| Excluding Lithuania[a] | 0.23 [0.13, 0.33] | $10^{15}$ | 2593 | $\mathcal{M}_u$ |
| Cauchy$^+$(0,2) prior on $SD$[a] | 0.24 [0.14, 0.33] | $10^{16}$ | 1160 | $\mathcal{M}_u$ |
| **State-by-Framing Effect** | | | | |
| Main analysis | -0.09 [-0.19, 0.00] | 0.02 | 0.04 | $\mathcal{M}_0$ |
| Hearing as mental state | -0.11 [-0.20, -0.02] | 0.02 | 0.03 | $\mathcal{M}_0$ |
| Excluding manipulation check failures | -0.07 [-0.17, 0.02] | 0.03 | 0.03 | $\mathcal{M}_0$ |
| Education and age as covariates | -0.09 [-0.18, 0.01] | 0.03 | 0.02 | $\mathcal{M}_0$ |
| Excluding Lithuania[a] | -0.10 [-0.19, -0.01] | 0.02 | 0.03 | $\mathcal{M}_0$ |
| Cauchy$^+$(0,2) prior on $SD$[a] | -0.09 [-0.18, 0.01] | 0.03 | 0.01 | $\mathcal{M}_0$ |

*Note:* Across all five sets of robustness checks, the results are qualitatively equal to those of the main analyses; the data indicate (a) strong state, religiosity, and framing effects that vary between countries but are consistently positive (mind > body; religious > non-religious; religious framing > secular framing), (b) a varying state-by-religiosity interaction effect (though sometimes the unconstrained model is preferred), and (c) no state-by-framing effect. Subscripts reflect parameter constraints; $_0$ indicates the null model, $_+$ indicates a varying positive effect, and $_1$ indicates a common effect.

[a] These options followed from strict adherence to the preregistration.

### 8.4.3 Exploratory Results

#### 8.4.3.1 Atheist extinctivists

In addition to the overall effects, we also investigated continuity judgments among self-reported atheists who explicitly state ‘not at all’ to believe in life after death ($n$ = 1513). As expected, for atheist extinctivists, the estimated intercept is considerably lower than for the overall sample: -2.92 (5.11%) vs. -0.96 (27.67%), respectively, as is the effect of mental versus physical state: 1.63 (i.e., an increase of 8.2% on the probability scale) vs. 1.71 (i.e., an increase of 32.9% on the probability scale). This is also displayed in the Figure 8.8. We note that the credible intervals for the estimates are quite wide for some countries where few people identify as atheists and deny an afterlife (e.g., India and Singapore). In general, the same pattern of results is observed for the atheist extinctivists as for the overall sample; the Bayes factor model comparison indicates most evidence for a varying positive effect of *state* ($\text{BF}_{+0} = 10^{67}$; $\text{BF}_{+1} = 16.93$) and of *framing* ($\text{BF}_{+0} = 10^{15}$; $\text{BF}_{+1} = 29.17$). Again, there is no evidence that the religious framing manipulation results in relatively stronger continuity judgments for mental states compared to physical states (i.e., state-by-framing interaction; $\text{BF}_{01} = 7.67$; $\text{BF}_{0+} = 37.90$).

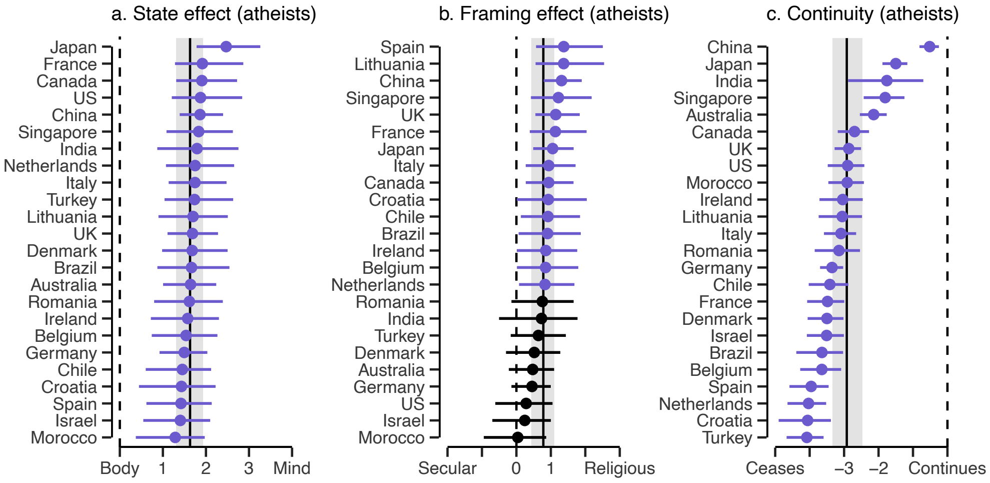


**Figure 8.8:** Estimated country-level effects (posterior medians) in increasing order for atheist extinctivists only. a. state contrast effects. b. framing effects. c. intercepts. Each dot represents a country. Estimates with credible intervals colored in purple exclude zero and estimates with credible intervals colored in black include zero. The errorbars give the 95% credible interval for each country. The vertical lines denote the posterior median of the overall mean of the respective effect with the 95% credible interval in the shaded bands. The dashed lines indicates zero.

**Table 8.5:** Bayes factor model comparison and parameter estimates for the key effects for atheists extinctivists only

| | Bayes factors | | | | Parameter estimates | |
|---|---|---|---|---|---|---|
| Effect | $\mathcal{M}_0$ | $\mathcal{M}_1$ | $\mathcal{M}_+$ | $\mathcal{M}_u$ | $\mu$ | $\sigma$ |
| State Effect | 0.00 | 0.06 | **1.00** | 0.09 | 1.63 [1.30, 1.93] | 0.44 [0.12, 0.85] |
| Framing Effect | 0.00 | 0.03 | **1.00** | 0.45 | 0.78 [0.44, 1.10] | 0.56 [0.23, 0.96] |
| State-by-Framing Effect | 0.69 | 0.09 | 0.02 | **1.00** | -0.39 [-0.85, 0.07] | 0.28 [0.01, 0.89] |

*Note.* The preferred model for each effect is assigned value 1.00 and displayed in bold. The remaining values are the Bayes factors for the respective model relative to this preferred model. Subscripts reflect constraints on the critical parameter; $_0$ indicates no effect, $_1$ indicates a common (positive) effect, $_+$ indicates a varying positive effect, and $_u$ indicates an unconstrained effect. Parameter estimates (median and 95% credible interval) are taken from the unconstrained model for $\mathcal{H}_5$. $\sigma$ reflects the between-country variation in the respective effect.

### 8.4.3.2 Exploratory factor analysis

Based on the ambiguity in the literature on the classification of perceptual states as either bodily states or mental states (Bering, 2002; Huang et al., 2013), we conducted a multilevel exploratory factor analysis to investigate the clustering of the 'hearing' item. We followed Weisman et al. (2021) in conducting an exploratory factor analysis across different samples. Specifically, for each country, we extracted the number of factors using parallel analysis with the `fa.parallel()` function from the `psych` package and then conducted EFA using ordinary least squares as implemented in the `fa()` function from the same package. As our data consisted of binary responses, we used tetrachoric correlations and oblique transformation. In order to maximize robustness, we ran the factor analysis procedure 100 times and report the median number of factors and factor loadings. The resulting factor loadings are visualized in Figure 8.9. As shown in the figure, across all countries except China[7] and India, the hearing item loaded most strongly on the mind-like or body-independent factor (though note that in Belgium, Italy, and the UK only one factor is extracted).

[7] Given the findings by Huang et al. (2013), it is noteworthy that China was one of the exceptions where the hearing item loaded most strongly on the body-dependent factor.

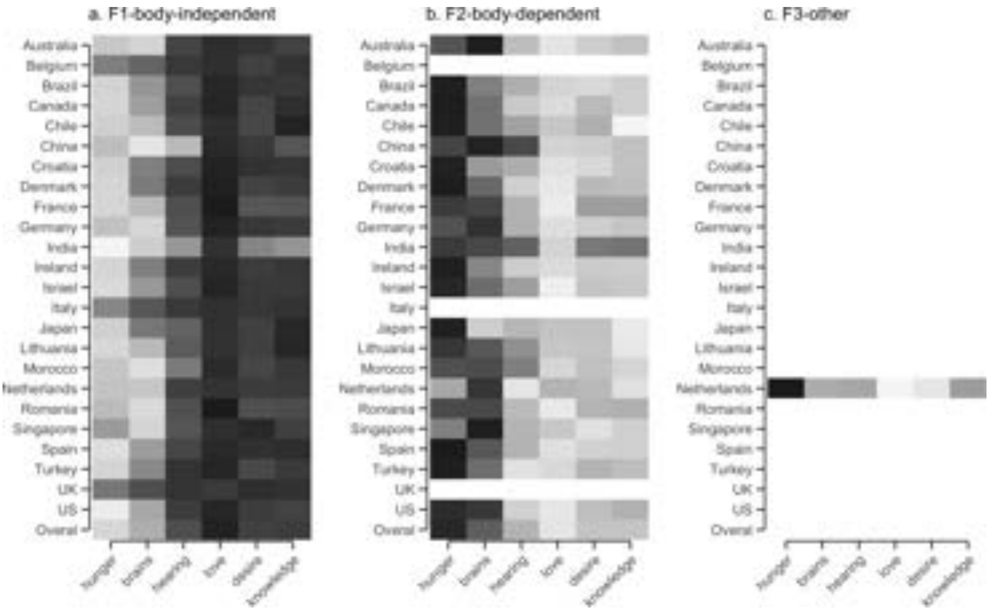


**Figure 8.9:** Factor loadings from EFAs per country. a. represents body-independent or mind-like factors, b. represents body-dependent or body-like factors and c. represents other factors. The shades of grey reflect the loading, with darker shades indicated stronger loading on the respective factor. Factors were extracted using tetrachoric correlations.

## 8.5 Discussion

In the current cross-cultural study, we replicated previous work showing that laypeople sometimes tend to reason dualistically about the continuity of states after death: across all 24 countries, the evaluated probability that mental states such as love and knowledge continue after death was higher than the evaluated probability for bodily states such as hunger or a working brain. In addition to this robust mind-body dualism effect, we also found that individual religiosity is consistently associated with increased implicit afterlife beliefs (i.e., overall continuity judgments for both mental and physical states). In all but one country –Morocco– a framing manipulation emphasizing religion also increased overall continuity judgments.

Even though these findings may appear straightforward at first sight, caution is warranted, because we should carefully distinguish between continuity judgments overall (which relate to implicit afterlife beliefs and may reflect that people believe both the mind and the body will continue to exist in some form) and the difference between continuity judgments for mental and physical states (which provides a proxy for people's dualistic thinking). One could argue that the mere continuation of any state might reflect dualistic reasoning as even the experience of post-mortem hunger implies a dissociation from pure bodily processes that are assumed to stop at death (e.g., inactivity of the stomach). In the current study, however, we followed previous work and characterized mind-body dualism as the difference in continuity judgments between physical/perceptual and mental/cognitive states.

Thus, in addition to these main effects, we investigated the relationship between religion and mind-body dualism: is religion associated with an increased tendency to distinguish between the continuity of mental and bodily states? In contrast to the main effect, we found that religiosity and a religious framing manipulation were not universally associated with increased mind-body dualism. Specifically, across 18 out of 24 countries, individual religiosity of the rater was related to more continuity judgments for mental states relative to physical states. In the remaining 6 countries, there was no such religiosity-by-state interaction effect. For the experimental framing manipulation, we did not find evidence that emphasis on religion increased mind-body dualism. On the contrary, in China, the effect went slightly into the opposite direction; the religious framing increased the relative continuity for physical compared to mental states. In all other countries, there was no state-by-framing interaction effect, nor was there a common effect in the aggregated sample.

While individual religiosity was consistently associated with a higher tendency for continuity judgments and –albeit somewhat less consistently– with more mind-body dualism, this association was not present at the country-level; the perceived normativity of religion within a country was not related to implicit afterlife beliefs, and, if anything, negatively related to mind-body dualism.[8] Overall continuity judgments were most prevalent in Asian countries (Singapore, China, India, and Japan). With the exception of India, these countries are not perceived as particularly religious based on the current data. However, cultural traditions related to immortality of the soul may exist outside of religious traditions. In China, for instance, less than 20% of the population is religiously affiliated (Grim, 2008), yet over 70% engages in ancestor worship, including venerating the spirits of deceased relatives (Hu, 2016). Indeed, many Chinese people indicate that the soul would persist after biological death, either in the afterlife or after reincarnation (Gut et al., 2021). So while implicit afterlife beliefs and religion are clearly linked, there are also other cultural traditions outside religion that may affect people's implicit afterlife beliefs.

In sum, our results suggest that the tendency to reason dualistically about people's capabilities after death is universal; across all 24 countries, we found robust evidence that mental states are judged as more likely to continue than bodily states. Even among atheist extinctivists who explicitly deny the existence of the afterlife, 16.9% of participants judged at least one state to continue after death and again in each of the 24 countries a state effect emerged, reflecting the tendency to attribute a higher likelihood of continuity for mental compared to physical states.

At the same time, these findings and their interpretation should be put into perspective. In all but 5 countries, as well as in the aggregated sample, the modal response was complete cessation rather than continuity, and over one third of the total sample (39%) judged none of the states to continue. Notably, if we take continuity beliefs in the narrative task as a measure of implicit afterlife beliefs, we find that more people endorse explicit afterlife beliefs in the absence of implicit afterlife beliefs (∼20%), than the other way around (∼5%). This pattern seems problematic

[8] In Appendix 8.C we report an additional analysis in which we correlated the country-level estimates for continuity judgments and mind-body dualism with census data on religion for included each country. Again, we found (weak) evidence for the absence of a positive correlation. In contrast to the cultural norms analysis, we did not find evidence in favor of a negative correlation either.

for intuitive dualism accounts, which suggest that especially implicit afterlife beliefs should be prevalent and widespread, more so than explicit and culturally transmitted afterlife beliefs.

Of course, the validity of the deceased grandmother vignette as a measure of implicit afterlife beliefs could be questioned, as has the general validity of vignette designs in experimental research (e.g., Argyris, 1975; Collett & Childs, 2011). On the one hand, there are clear benefits of using narratives to measure certain beliefs, attitudes, and intentions, including experimental control and the accessibility of ethically or practically difficult to manipulate scenarios (Aguinis & Bradley, 2014), such as someone dying. On the other hand, drawbacks include limited external validity and generalizability. First, responses might be influenced by social desirability (Gould, 1996). This issue seems most problematic in the context of vignettes targeting personal experiences ('imagine that you are...') as opposed to the type of third-person scenario used in the present study (Collett & Childs, 2011). Still, a form of demand characteristics could have played a role, resulting in an overestimation of implicit afterlife beliefs. Perhaps participants did not literally believe that the dead grandmother could still feel love, but simply responded within the context of the story, as if immersing themselves in a fairy-tale. This might explain why even atheists exhibited some implicit afterlife beliefs in the narrative task; rather than their continuity responses reflecting a divergence between explicit and implicit beliefs and hence a contraction in their beliefs, they may simply have 'played along' with the task. Second, responses to the vignette might reflect unintended peculiarities of the specific narrative (Gould, 1996). That is, the observed pattern could be idiosyncratic to the presented narrative and underestimate true implicit afterlife beliefs; perhaps people think that deceased individuals are in principle capable of feeling love and having knowledge, but that this does not hold for the grandmother in the narrative for some reason. However, we consider this explanation rather unlikely, as we see no obvious reason for participants to assume that this particular grandmother would not love her grandson anymore if she is indeed still capable of feeling love.

Overall, we believe that our results are more in line with the 'intuitive materialism' account (H. C. Barrett et al., 2021) than the 'intuitive dualism' account: at least in 19 out of 24 countries the default view seems that physical death ends all mental processes. Following the parallel systems account, which holds that humans can be construed as both intentional agents and as physical bodies, mental capacities are typically only attributed to living people and not to the deceased. In some cases, however, agency can be perceived in the absence of a physical body, allowing for afterlife beliefs about disembodied minds, spirits, and supernatural entities.

The question remains what mechanism underlies continuity beliefs and what determines which states are most likely to be judged to cease and to continue. Instead of a specialized cognitive mechanism for afterlife beliefs (cf. Bering, 2002; Bloom, 2005), a general mechanism for person continuity may provide a more parsimonious account. One option would be a form of psychological essentialism applied to individuals (Blok et al., 2001; Blok et al., 2005). On this account, we track individuals through perceived causal connectivity in time and space, and through radical transformations (Liittschwager, 1994; Rips et al., 2006). Continuation of certain states in the afterlife, thus, may be a reflection and natural consequence of the everyday strategies used for tracking individuals (Newman et al., 2006). The exact elements that are relevant and salient for what counts as marks of the same individual (e.g., their memories, their emotions, their body parts or special marks), both in everyday life and in the afterlife, may be culture-specific. In other words, what matters is a perception of a continued individual (their essence or perhaps 'soul'), while what sort of mental or bodily states are emphasized in the afterlife could vary across cultures. For instance, it could be the case that those cultures that emphasize individuals' psychological dispositions over their social relation (as many Western cultures do; see Henrich, 2020) would tend to conceptualize afterlife in term of psychological continuity (at least among religious individuals). Likewise, those cultures that emphasize individuals' social relations might be conceptualizing afterlife in terms of social embodiment (as Hodge, 2011a suggested). Furthermore, different iconography and depictions of afterlife across cultures might stress slightly different bodily aspects of the individuals in the afterlife. Finally, across various Asian cultures people use different methods to determine someone's reincarnation. Besides alleged memories of the past life, for instance, some Mongolian Buddhists mark their deceased with coal or chalk, and inspect their newborns for birth marks in the same bodily area (see also C. White, 2015).

Our findings show that the tendency to make continuity judgements depends on the framing; participants who read a narrative featuring religious or spiritual elements (e.g., a priest, God) were ∼10% more likely to make continuity judgments than participants who read a narrative without these elements. Notably, the religious versus secular framing manipulation also covaried with a focus on medical versus spiritual features; a doctor announcing death versus a priest announcing an afterlife. This difference in emphasis might also explain why the framing effect even emerged among atheist extinctivists, who explicitly do not believe in the existence of God or an afterlife; while the framing manipulation did not 'prime' their religious beliefs, it may have emphasized a more metaphysical vs. a medical conception of a dead body. In addition, the difference in wording between conditions might also have contributed to the observed framing effect: compared to 'grandmother is dead now', 'grandmother is with God now' may already imply continuity. Based on the anthropomorphic God concepts that people employ (J. L. Barrett & Keil, 1996), it could well be that participants visualize the grandmother literally and physically at a different location - which results in an overall increase in endorsement of continuity for both physical and mental states. Alternatively, the 'is with God now' wording may have amplified continuity responses as a result of the tendency to 'play-along' with the scenario.

As observed by P. L. Harris (2011a), many children and adults alike subscribe to both a biological and a religious conception of death. Depending on the framing, each of these conceptions may dominate attitudes and behaviors regarding afterlife processes. According to Van Leeuwen (2014), these framing-sensitive afterlife beliefs are an example of what he calls 'religious credence': they are not factual beliefs, but attitudes that are only relevant in certain contexts–such as a burial ritual. However, the context sensitivity of these beliefs may not be unique for religious beliefs (Levy, 2017). Instead, it is the intuitiveness of representations, whether religious or mundane, that guides people's responses, which is why even atheists who do not believe in an afterlife sometimes indicate that love continues after death and why people who deny the existence of a soul are unwilling to 'sell' their soul to an experimenter (Haidt et al., 2000).

Indeed, the exact nature of the continuity beliefs observed in our experiment remains unclear:: do people truly believe that the deceased person has an independent mind maintaining these capacities, or does it rather reflect some sort of persistence of positive associations or feelings that the person had before they died (e.g., a loving person can still feel love)? That is, do people believe that the deceased grandmother can literally hear her grandson, or is this 'hearing' a metaphorical idea of a sustained connection between grandmother and grandson? Even though one assumes that all (mental) functioning stops at death, one might still prefer to hold on to social relations and emotions that were present before the passing. In that sense, perhaps the notion of persistent love is intuitive because the alternative that she does not love him anymore feels uncomfortable. Future research could possibly address the nature of these continuity beliefs by manipulating the valence and relevance of the states. If people are more likely to indicate that love and kindness continue than anger and jealousy, this might suggest that some idealization of the deceased plays a role (e.g., Allison et al., 2009; Bering, McLeod, et al., 2005; Eylon & Allison, 2005). This 'death positivity bias' may serve the purpose of giving comfort when losing a loved one (Attig, 1996). Similarly, a difference in continuity judgments covarying with the mundaneness of the emotion (e.g., loving his wife versus loving to watch Netflix) might also signal a bias in how we remember the dead, which spills over to the capacities attributed to them.

The exploratory factor analysis indicated that the perceptual state ('hearing') is mostly perceived as a mental rather than a bodily state. In some Asian countries (India, China), however, this does not seem to be the case. The clustering of perceptual states might also depend on the object of the state (i.e., to which the specific state is directed). Here we asked about the grandmother hearing her grandson's voice; results might be different for 'hearing the equipment in the hospital room' or 'hearing the traffic outside'.

While our study has demonstrated a robust and universal pattern, it leaves many questions open for future research. First, how are explicit and implicit afterlife beliefs related? Our results suggests that they often converge, but that explicit afterlife beliefs might actually be more prevalent than implicit ones. Still, the causal relation between both implicit and explicit afterlife beliefs remains unclear. On the one hand, religious or spiritual beliefs could make people more receptive to the possibility of continuation of (mental) states (cf. H. C. Barrett et al., 2021). On the other hand, it could also be that the intuitiveness of mental continuation serves as an evolutionary explanation for the appeal of religion (e.g., Bering, 2002; Bloom, 2007). Second, while we observed cross-cultural variation in overall continuity beliefs, as well as mind-body dualism, our data do not permit a fine-grained analysis of cultural differences in participants' conception of these beliefs. The exploratory factor analysis suggests that across most countries, people distinguish between body-dependent (hunger, brains) and body-independent (hearing, love, desire, knowledge) processes, yet some difference between countries emerged. Third, in addition to the valence of the states, attitudes toward the target person might also affect continuity evaluations. Are people more likely to make continuity judgments for likable individuals and relatives compared to "bad guys" and strangers? Fourth, future research might investigate what individual differences besides religiosity predict implicit afterlife beliefs. For instance, what is the role of scientific training on these beliefs? Does a neuroscientific view of the brain and mind preclude continuity beliefs or can they coexist? Fifth, what, if any, are the behavioural consequences of these mind-body dualism beliefs? Forstmann et al. (2012), for instance, found that mind-body dualism was negatively associated with health behaviors, following the rationale that viewing the body as a mere container leads to taking less care of it.[9] It thus remains to be demonstrated how afterlife beliefs have down-stream effects on our behavior. Sixth, how do the different conceptions of 'intuitiveness' relate to one another? Levy (2020) distinguishes between intuitive dualism as a cultural universal - acquired without significant cultural scaffolding - and intuitive dualism as exerting influence over implicit cognition. Our findings clearly indicate that dualism is not intuitive in the second sense for all participants, but do not speak directly to the first.

In conclusion, our results suggest both universality and cross-cultural variation in reasoning about mental processes after biological death. Using a large sample from 24 different countries, we replicated previous findings that people tend to reason dualistically as they consider mental states more likely to continue after death than bodily states and that a framing manipulation emphasizing a religious conception of death increases overall continuity judgments, though not mind-body dualism. In addition, we showed that individual religiosity in general, and explicit afterlife beliefs in particular, are predictive of both overall implicit afterlife beliefs and mind-body dualism. At the same time, the pattern of the data does not imply universal intuitive dualism. Specifically, the modal response across the majority of countries and the aggregated sample was complete cessation of all states and explicit afterlife beliefs were more prevalent than implicit afterlife beliefs. Based on these data, an intuitive materialism account, assuming a default conception that all mental activity ends at physical death, yet allowing for culturally acquired explicit afterlife beliefs, appears more plausible than an intuitive dualism account.

[9] We note that these were social priming studies, the reliability of which in general has been called into question (e.g. Cesario, 2014; Doyen et al., 2012; Gilder & Heerey, 2018; Pashler et al., 2013; Shanks et al., 2013).

## Appendix 8.A Prior Predictive Checks

In order to systematically and thoroughly assess the adequacy of the priors, we should look at some settings for both the priors on the intercepts, the effects, the variability between countries and the correlation matrix. We can use previous studies to inform our options.

We will consider the following:

- intercept:

1. normal(0,1)
2. normal(0,5)
3. student-t(3,0,2.5): brms default

- effect:

1. normal(0,5)
2. normal(0,1)
3. normal(0,0.5)
4. normal(0,1000): approximation of the brms default of a flat prior

- standard deviation between countries:

1. exponential(1): as suggested by McElreath (2020).
2. inverse-gamma(3,0.5): assuming a standard deviation below 0.5
3. cauchy(0,2): as preregistered
4. normal(0,1)
5. student-t(3,0,2.5): brms default

- correlation matrix:

1. lkj(1): flat distribution for the correlation matrix
2. lkj(2): putting slightly less mass on extreme correlation values (i.e., -1 and 1)

What do we know? Based on data from previous studies that have been conducted across different cultures, we can get an idea of the expected intercepts and size of the effects. The mean state effect –the difference in the probability of continuity responses for mental vs. bodily states– across these 12 sites, taken from 4 previous studies is 0.16, so 16% with a standard deviation of 0.16 (15.70%). For the 10 framing effects in the previous studies, the mean difference between a theistic/spiritual prime and the neutral/control condition is 0.10 (10.20%) with a standard deviation of 0.17 (17.50%). Based on these data, we would expect experimental effects of about 10-20% and a standard deviation between studies/countries of about 15-20%.

What do we want? In the simulation, we draw samples from the prior distributions and look whether the distributions of the country-level intercept (i.e., the overall probability of saying that a given state will continue) and the predictions on the country-level experimental state effect (i.e., difference in probability of saying 'continues' between mental and physical states) make sense. If priors are too vague the distributions become bimodal, suggesting that all participants in a given country either judge all states to cease or continue. We aim to find prior distributions that are relatively uninformative while still allowing making sensible predictions.

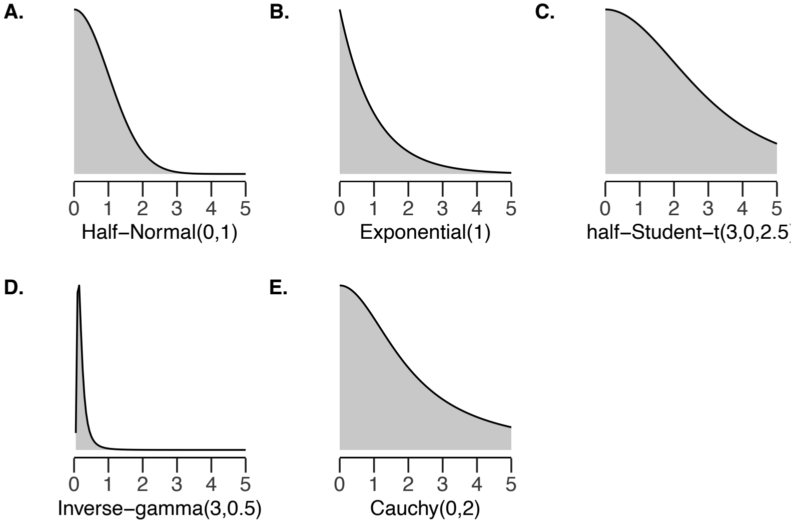


**Figure 8.10:** Different prior settings for the between-country variation in the effects of interest.

What do we conclude? We found that the LKJ settings do not have a strong influence on the chosen parameters. We therefore show only the LKJ(2) parameter case, as we think correlations between country-level effects of -1 or 1 are less likely than more modest correlation values a priori.

First, the normal(0,5) prior on the intercept translates into extreme predictions on the probability scale, resulting in a unrealistic bimodal distribution with most mass close to 0 and 1. The normal distribution with standard deviation 1, on the other hand, seems to make reasonable predictions about the overall probability of continuity, allowing for all values between 0 and 1 with most mass around 0.5. Second, based on visual inspection, is seems both the exponential(1) and the half-normal(0,1) prior for the between-country variation make sensible predictions. The inverse-gamma(3,0.5) seems a bit too strict and the preregistered cauchy(0,2) and the brms-default student-t(3,0,2.5) are too wide to translate into reasonable predictions on the probability scale. Finally, a normal distribution with a standard deviation of 1 seems to make the best predictions for the experimental state effect, putting most mass on smaller differences, but still allowing for effects up to 75% (as observed in one previous study). Based on these prior predictions, we decided to use the normal(0,1) prior for the intercept and the effect, the half-normal(0,1) for the variation between countries, and the LKJ(2) for the correlation matrix.

As becomes evident in Figure 8.14, predictions from both our preregistered prior settings and the brms default settings are completely unrealistic; both predict that all responses with be either complete cessation or continuity. The brms default priors are much too wide, resulting in predicting an unlikely difference of 100% between conditions. The preregistered priors, on the other hand, predict a modest effect, but due to the wide prior on the variation between countries, this results in a very strong prediction of observing no effect. Note, however, that in this case, because we have so

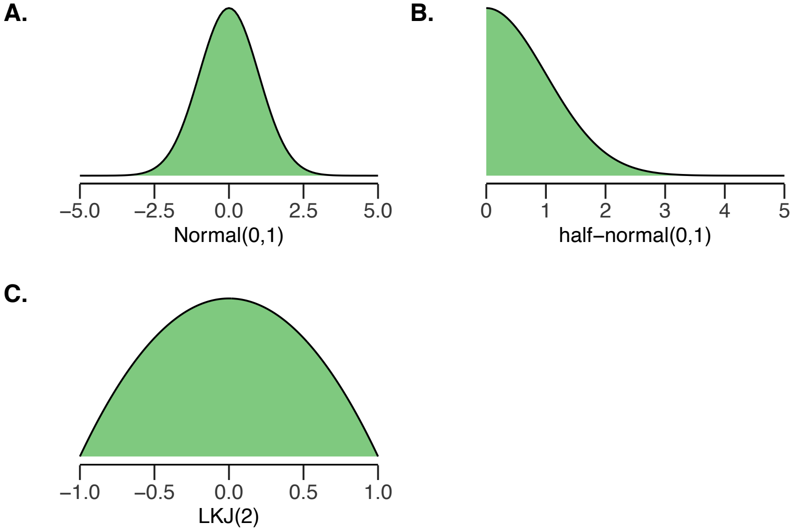
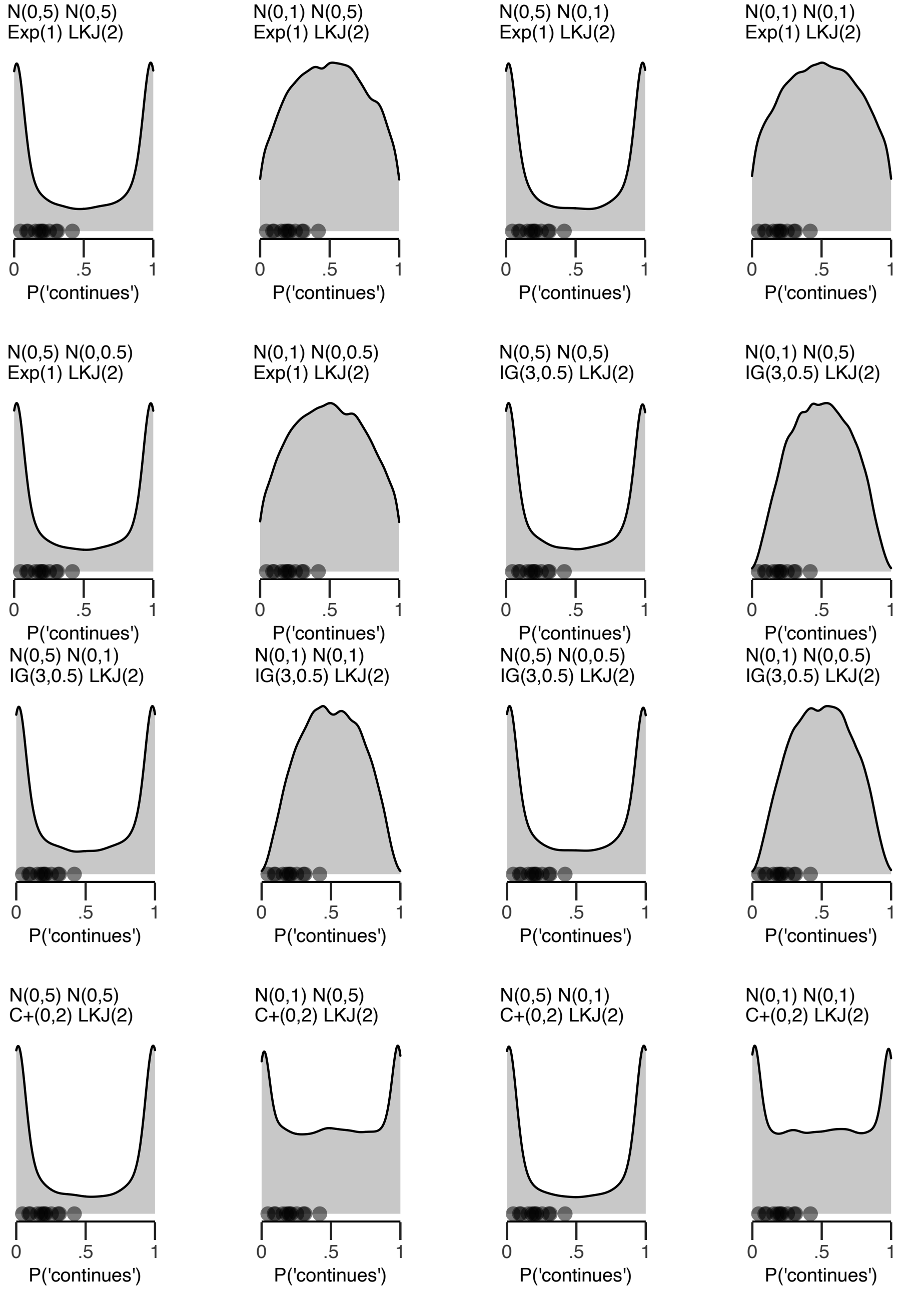


**Figure 8.11:** Chosen prior settings for the main analysis. A. shows the prior on the intercept and the effect, B. shows the prior on the variability between countries, and C. shows the prior on the correlation matrix.

much data, the data will always outweigh the priors, resulting in a reasonable posterior distribution, regardless of the exact prior specifications (see robustness checks).


N(0,5) N(0,5)
Exp(1) LKJ(2)
N(0,1) N(0,5)
Exp(1) LKJ(2)
N(0,5) N(0,1)
Exp(1) LKJ(2)
N(0,1) N(0,1)
Exp(1) LKJ(2)
N(0,5) N(0,0.5)
Exp(1) LKJ(2)
N(0,1) N(0,0.5)
Exp(1) LKJ(2)
N(0,5) N(0,5)
IG(3,0.5) LKJ(2)
N(0,1) N(0,5)
IG(3,0.5) LKJ(2)
N(0,5) N(0,1)
IG(3,0.5) LKJ(2)
N(0,1) N(0,1)
IG(3,0.5) LKJ(2)
N(0,5) N(0,0.5)
IG(3,0.5) LKJ(2)
N(0,1) N(0,0.5)
IG(3,0.5) LKJ(2)
N(0,5) N(0,5)
C+(0,2) LKJ(2)
N(0,1) N(0,5)
C+(0,2) LKJ(2)
N(0,5) N(0,1)
C+(0,2) LKJ(2)
N(0,1) N(0,1)
C+(0,2) LKJ(2)
0
.5
1
P('continues')

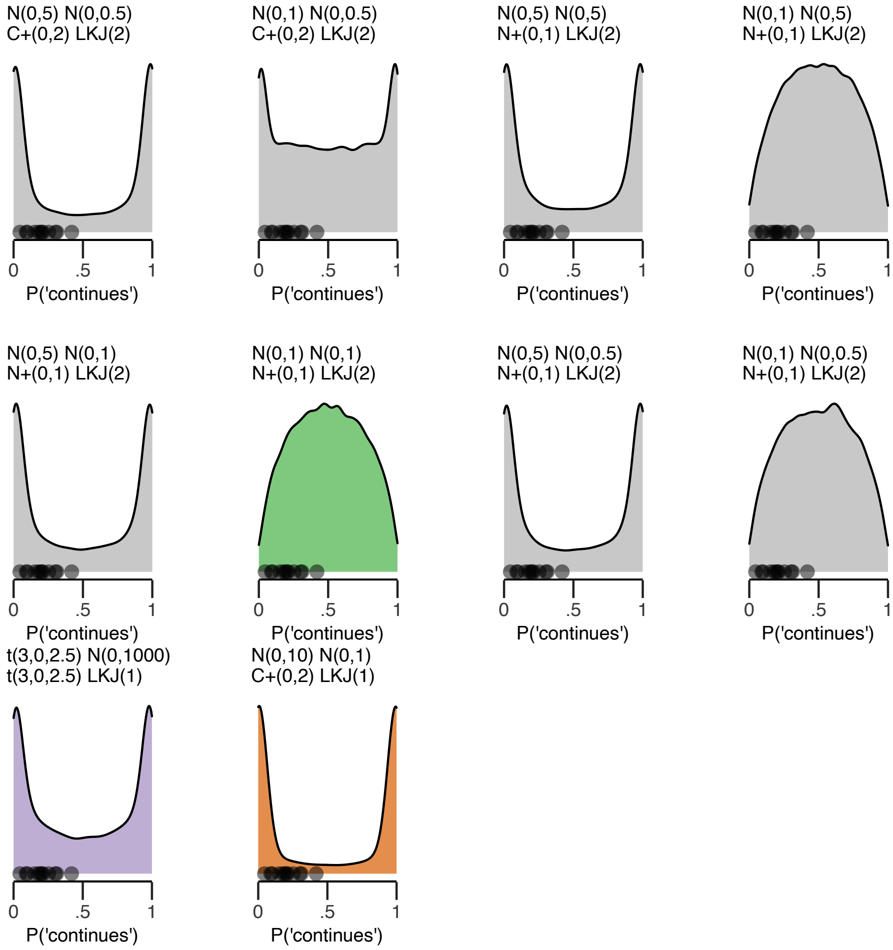
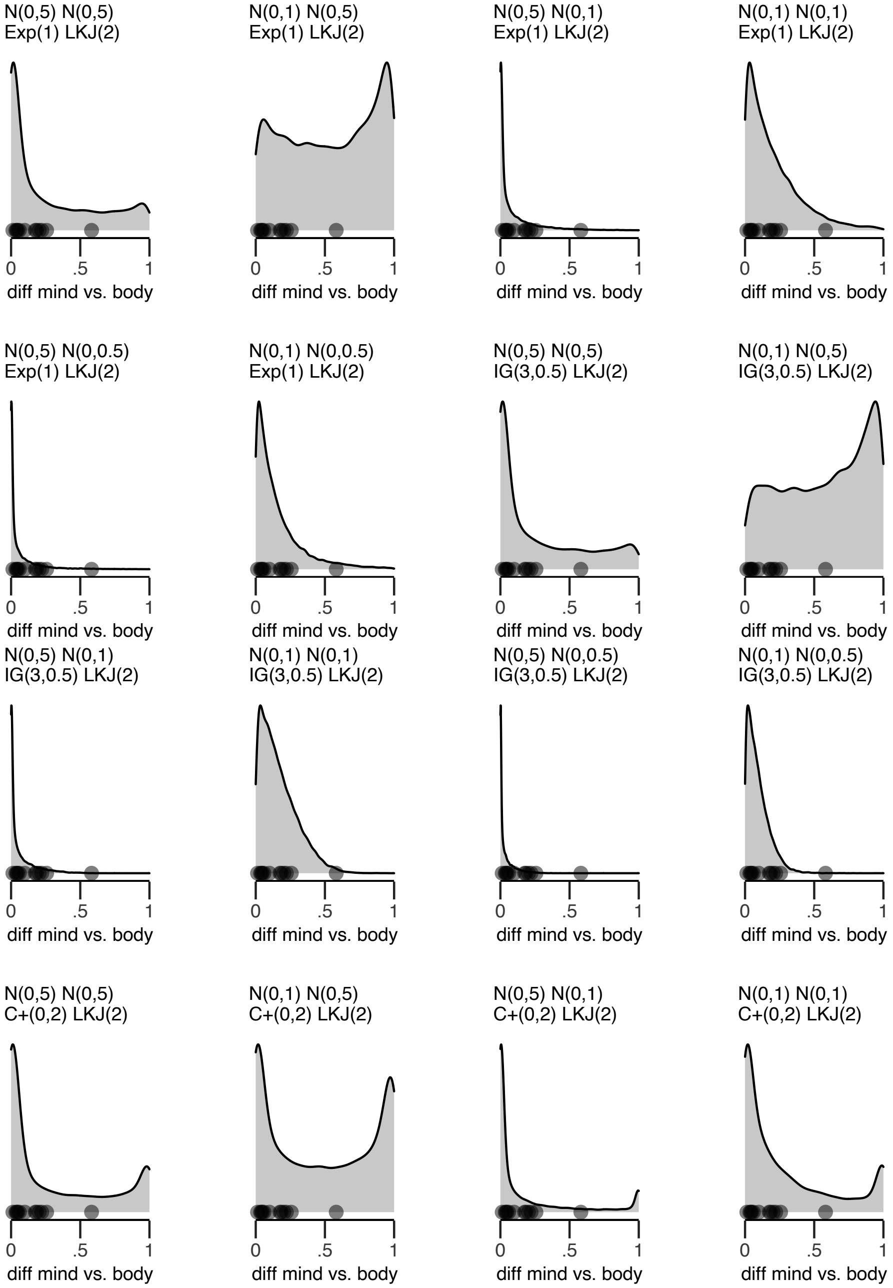


**Figure 8.12:** Prior predictive distributions for the overall probability of continuity (i.e., the intercept) under all considered prior settings on the intercept, effect, and between-country variation. The points on the x-axis reflect the observed continuity in previous studies. The distribution for the chosen prior settings is displayed in green, for the brms default settings in purple, and for the preregistered settings in orange. The distributions are denoted as follows: N = normal, N+ = the half-normal, Exp = exponential, IG = inverse gamma, C+ = half-cauchy, t = Student t, and LKJ = Lewandowski-Kurowicka-Joe.


N(0,5) N(0,5)
Exp(1) LKJ(2)
N(0,1) N(0,5)
Exp(1) LKJ(2)
N(0,5) N(0,1)
Exp(1) LKJ(2)
N(0,1) N(0,1)
Exp(1) LKJ(2)
N(0,5) N(0,0.5)
Exp(1) LKJ(2)
N(0,1) N(0,0.5)
Exp(1) LKJ(2)
N(0,5) N(0,5)
IG(3,0.5) LKJ(2)
N(0,1) N(0,5)
IG(3,0.5) LKJ(2)
N(0,5) N(0,1)
IG(3,0.5) LKJ(2)
N(0,1) N(0,1)
IG(3,0.5) LKJ(2)
N(0,5) N(0,0.5)
IG(3,0.5) LKJ(2)
N(0,1) N(0,0.5)
IG(3,0.5) LKJ(2)
N(0,5) N(0,5)
C+(0,2) LKJ(2)
N(0,1) N(0,5)
C+(0,2) LKJ(2)
N(0,5) N(0,1)
C+(0,2) LKJ(2)
N(0,1) N(0,1)
C+(0,2) LKJ(2)
0
.5
1
diff mind vs. body

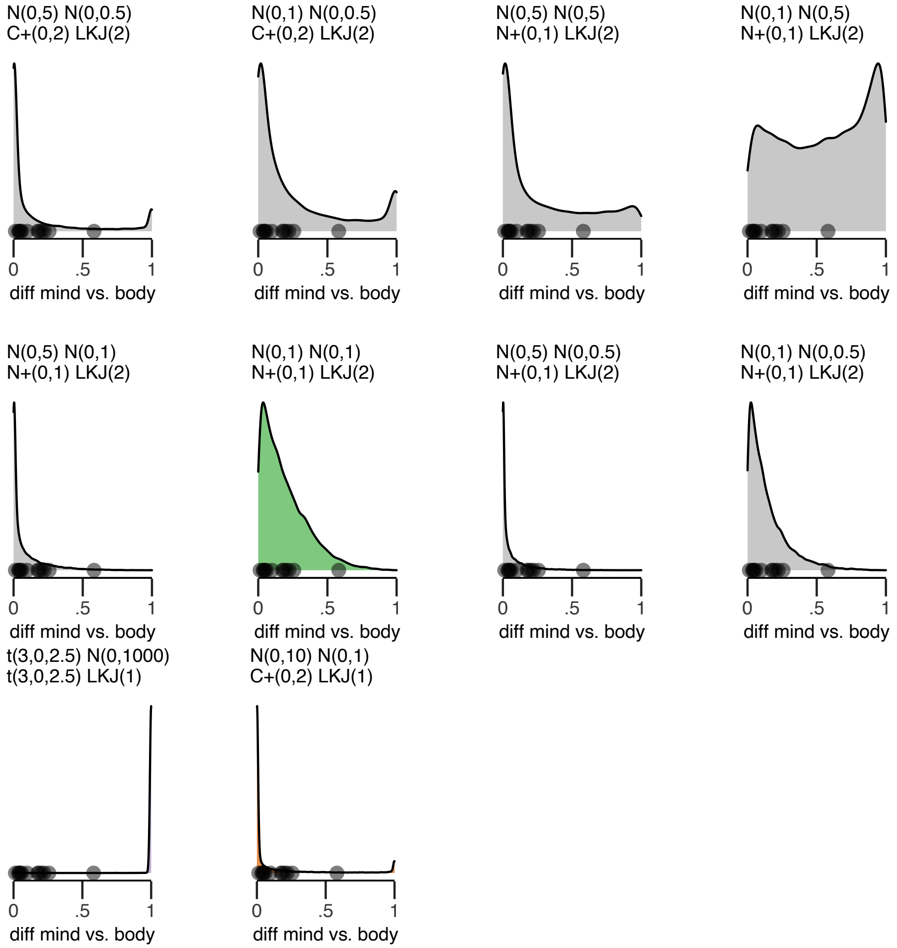


**Figure 8.13:** Prior predictive distributions for the difference between experimental conditions (i.e., mental vs. bodily states or religious vs. secular framing) under all considered prior settings on the intercept, effect, and between-country variation. The points on the x-axis reflect the observed effect in previous studies. The distribution for the chosen prior settings is displayed in green, for the brms default settings in purple, and for the preregistered settings in orange. The distributions are denoted as follows: N = normal, N+ = the half-normal, Exp = exponential, IG = inverse gamma, C+ = half-cauchy, t = Student t, and LKJ = Lewandowski-Kurowicka-Joe.

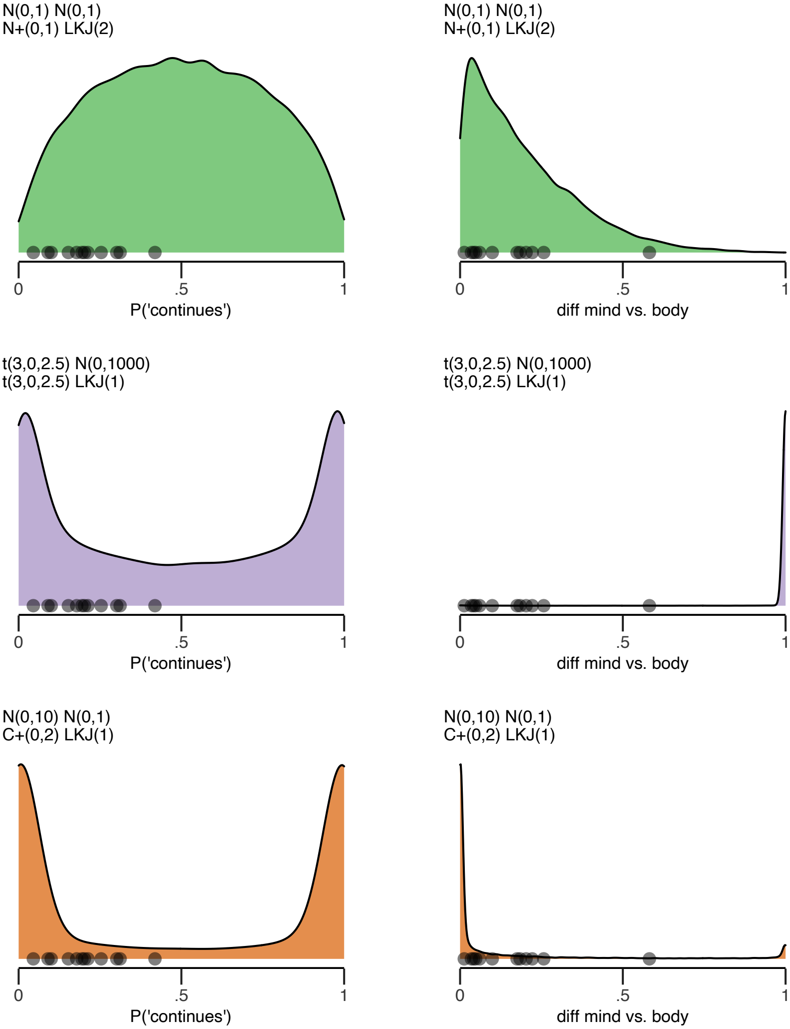


**Figure 8.14:** Prior predictions for the chosen prior settings (in green), the brms default settings (in purple), and the preregistered settings (in orange). The distributions are denoted as follows: N = normal, N+ = the half-normal, C+ = half-cauchy, t = Student t, and LKJ = Lewandowski-Kurowicka-Joe.

## Appendix 8.B Additional Model Statistics

### 8.B.1 MCMC Diagnostics

To investigate convergence of the MCMC chains, we extracted the $\hat{R}$ values for all model parameters. The smallest and largest $\hat{R}$ values were 1.00 for the correlation between the slope of the state effect and the state-by-framing effect and 1.00 for the individual level religiosity effect, respectively. The traceplots for these smallest and largest $\hat{R}$ values are shown in Figure 8.15a and b.

The ratio of effective samples versus total samples $\hat{N}_{\text{eff}}/N$ was calculated per parameter to assess to what extent autocorrelation in the chains reduces the certainty of the posterior estimates (Geyer, 2011). Ideally, $\hat{N}_{\text{eff}}$ is as large as possible (Vehtari et al., 2021). The $\hat{N}_{\text{eff}}/N$ for each of the 315 estimated parameters is displayed in Figure 8.15c. Note that $\hat{N}_{\text{eff}}$ can be larger than the the total number of iterations (in this case: $N = 20000$) when the samples are anti-correlated or antithetical (Carpenter, 2018). The smallest $\hat{N}_{\text{eff}} = 3016$ for the overall intercept. For many parameters, $\hat{N}_{\text{eff}}$ is at least half of the number of iterations, although for some parameters the ratio is rather low, indicating that there is some autocorrelation in the chains. Nevertheless, since brms uses the NUTS sampler (Hoffman & Gelman, 2014), even for complex models 'a few thousand' samples generally suffice for stable results (Bürkner, 2017). We therefore concluded that the effective sample size is sufficient for valid interpretation of the estimates and inference.

## Appendix 8.C Additional Analyses

We explored whether the data provide evidence for an interaction between religiosity and context, such that the religious framing effect on continuity judgments is enhanced for religious participants in particular. The Bayes factor model comparison provided most evidence for the unconstrained model: $\text{BF}_{u0} = 334$. As shown in Figure 8.16a, the unconstrained model is favored because in some countries the effect is positive, whereas in others it is negative. However, in only four countries (UK, Romania, Germany, and China) do the credible intervals exclude zero. Overall, there is no evidence in favor of a religiosity-by-framing interaction effect assuming that the framing effect on continuity is larger for religious participants: $\text{BF}_{10} = 0.87$; $\text{BF}_{01} = 1.15$ (this counts as basically no evidence either way).

Finally, we tested the evidence for a three-way interaction between state, religiosity and framing, such that mind-body dualism increases with religiosity, and particularly when framed in religious terms. The Bayes factor model comparison indicated some evidence in favor of a common three-way interaction: $\text{BF}_{10} = 17.96$. However, as shown in Figure 8.4b, it appears that compared to the secular framing, the religious framing slightly increases mind-body dualism for low religiosity in particular, but not for high religiosity. Based on the unclear pattern, the relatively small Bayes factor given the amount of data, and the fact that the three-way interaction only appears in 3 out of the 24 countries (see Figure 8.16b), we do not consider this effect of relevance.

In addition to the country-level cultural norms measured in the survey, we also used external census data on religiosity to investigate if national levels of religiosity based on representative sample might be related to continuity judgments and mind-body dualism. The reason for adding this analysis is that perhaps people's perception of

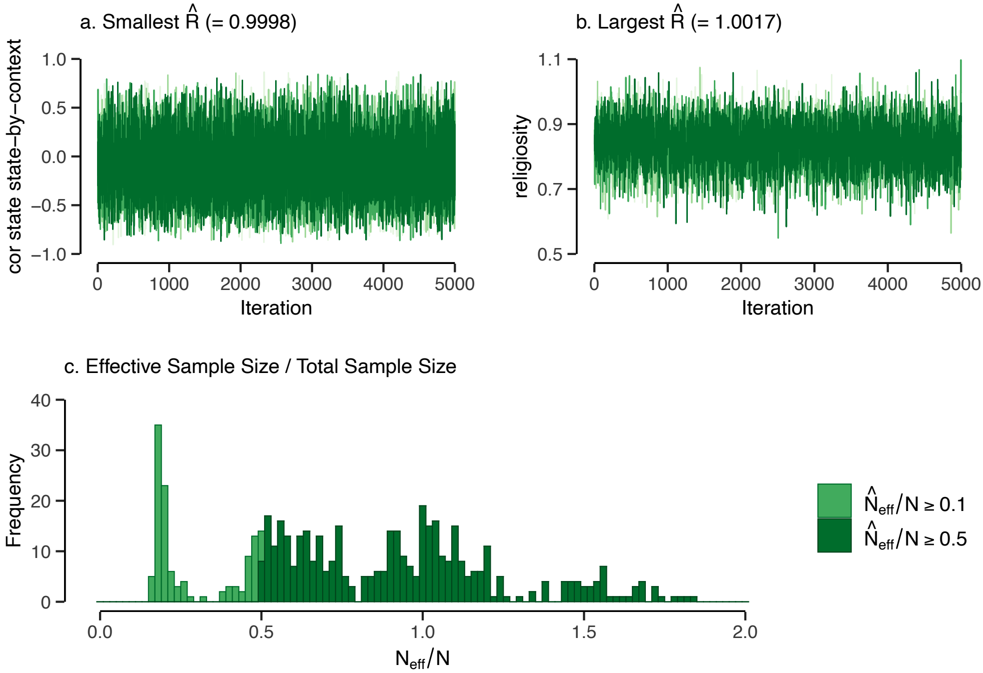


**Figure 8.15:** MCMC diagnostics. a. Chains for parameters with the smallest (correlation between the slope for the state effect and the state-by-framing interaction effect) and b. largest (individual level religiosity effect) rhat values. c. Ratio of the number of effective samples versus the total samples for each parameter in the full model.

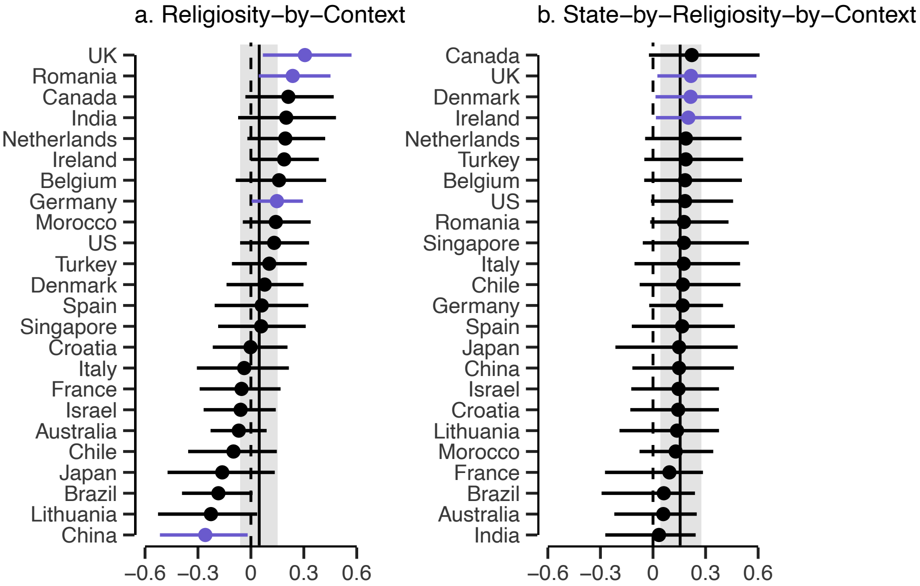


**Figure 8.16:** Estimated country-level effects (posterior medians) in increasing order. a. Religion-by-context interaction effects, where positive values indicate more continuity judgments for religious individuals in the religious framing condition. b. State-by-religiosity-by-context three-way interaction effects, where positive values indicate stronger mind-body dualism for religious individuals in the religious framing context. Each dot represents a country. Estimates with credible intervals colored in purple exclude zero and estimates with credible intervals colored in black include zero. The errorbars give the 95% credible interval for each country. The vertical lines denote the posterior median of the overall mean of the respective effect with the 95% credible interval in the shaded bands. The dashed lines indicates zero.

religiosity in their country does not correspond well to the actual levels of religiosity in their country. Therefore, we used data on global religious adherence from the Association of Religion Data Archives (ARDA; Brown & James, 2019) to complement the cultural norms correlational analysis.

We correlated country-level religious adherence (in percentage; from 2015) with country-level estimates of the intercepts ($\alpha_j$) and state-effects ($\beta_j$) in the models. First, similar to the cultural norms analysis, we found some anecdotal evidence against a positive correlation between the country-level overall probability of continuity and cultural religiosity: $\text{BF}_{+0} = 0.39$; $\text{BF}_{0+} = 2.55$. Second, we also obtained anecdotal to moderate evidence against a positive correlation between country-level estimates of dualism (i.e., the state effect) and cultural religiosity: $\text{BF}_{+0} = 0.25$; $\text{BF}_{0+} = 3.97$ (see Figure 8.17). Contrary to the cultural norms analysis, we do not find evidence in favor of a negative relation either; the data indicate no correlation in either direction.

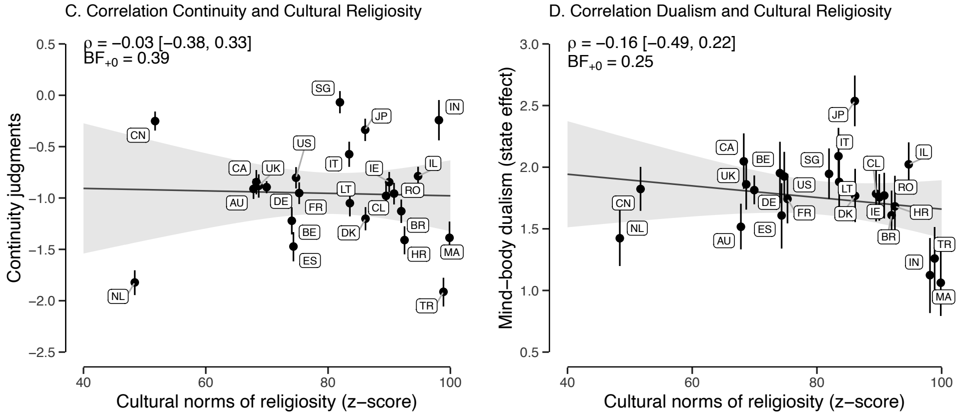


**Figure 8.17:** Census data on religion and continuity judgments (panel A.) and mind-body dualism (i.e., state effects; panel B.).

# 9

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

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.

## 9.1 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

This chapter has been adapted from: Hoogeveen, S., Sarafoglou, A., Aczel, B., Aditya, Y., Alayan, A., Allen, P., Altay, S., Alzahawi, S., Amir, Y., Anthony, F.-V., Appiah, O., Atkinson, Q. D., Baimel, A., Balkaya-Ince, M., Balsamo, M., Banker, S., Bartoš, F., Becerra, M., Beffara, B., … Wagenmakers, E.-J. (2022). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior*. https://doi.org/10.31234/osf.io/pbfye.
The full author list appears in the Contributions section (p. 331).

al., 2017; L. K. George et al., 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 et al., 2003), more happiness (Abdel-Khalek, 2006; Lewis & Cruise, 2006), higher life satisfaction (Lim & Putnam, 2010), and even lower mortality (Ebert et al., 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, et al., 2022) reported in Chapter 10.

## 9.2 Theoretical Background

The literature on the psychology of religion is replete with positive correlations between (self-rated) religiosity and mental health (Abdel-Khalek, 2006; L. K. George et al., 2002; Koenig and Larson, 2001; Plante and Sherman, 2001; Seybold and 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 and Sanders, 2003; Koenig and 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 et al., 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 et al., 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; S. R. 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 et al., 2011; Ebert et al., 2020; Gebauer et al., 2017; Stavrova, 2015; Stavrova et al., 2013). This makes sense from the literature on person-culture fit (Dressler et al., 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 et al., 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 a 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).

## 9.3 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 & Uhlmann, 2015; 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 et al., 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 and 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 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 et al., 2020; Rosmarin et al., 2010). Yet new studies keep emerging (e.g., 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 et al., 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 et al., 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 et al., 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, 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 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 et al., 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

[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 et al., 2020; Rohrer, 2018 for a perspective on causal inference in non-experimental studies).

osf.io/xg8y5), although the details of the processing and analysis of the many-analysts data were not preregistered.

## 9.4 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, Haaf, et al., 2022; Hoogeveen and van Elk, 2021). 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 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

[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).

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.

## 9.5 Disclosures

### 9.5.1 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 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 9.2 for an overview of all resources. Online Appendices can be accessed via https://osf.io/9kpfu/.

### 9.5.2 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.,

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

Sarafoglou et al., 2022; see Chapter 11). 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).

### 9.5.3 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.

## 9.6 Methods

### 9.6.1 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 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 9.1. In addition, Figure 9.1 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.

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

**Table 9.1:** 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.

### 9.6.3 Materials

#### 9.6.3.1 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 (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,

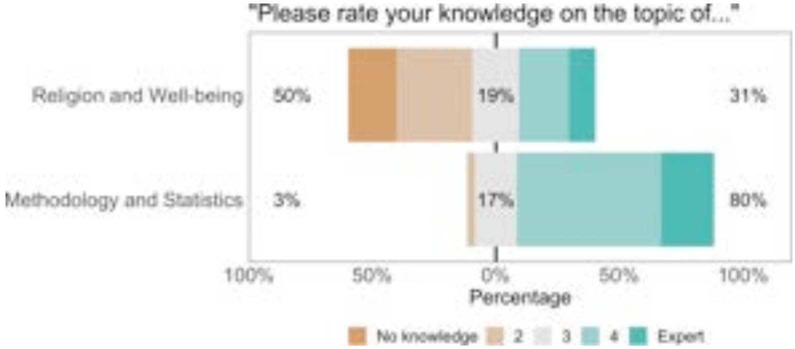


**Figure 9.1:** Responses to the survey questions on self-rated topical and methodological knowledge. The top bar represents the teams' answers about their knowledge regarding religion and well-being and the bottom bar represents the teams' answers about their knowledge regarding methodology and statistics. 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.

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).

### 9.6.4 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 9.2 for an overview of the procedure.

**Table 9.2:** 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.

## 9.7 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.

### 9.7.1 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 9.3) and (2) both the (scale of the) independent and dependent variables might vary across teams, we considered a beta coefficient the most suitable effect 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).

#### 9.7.1.1 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 9.2, the results are remarkably 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 9.3 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.

#### 9.7.1.2 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 9.4 the results for research question

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

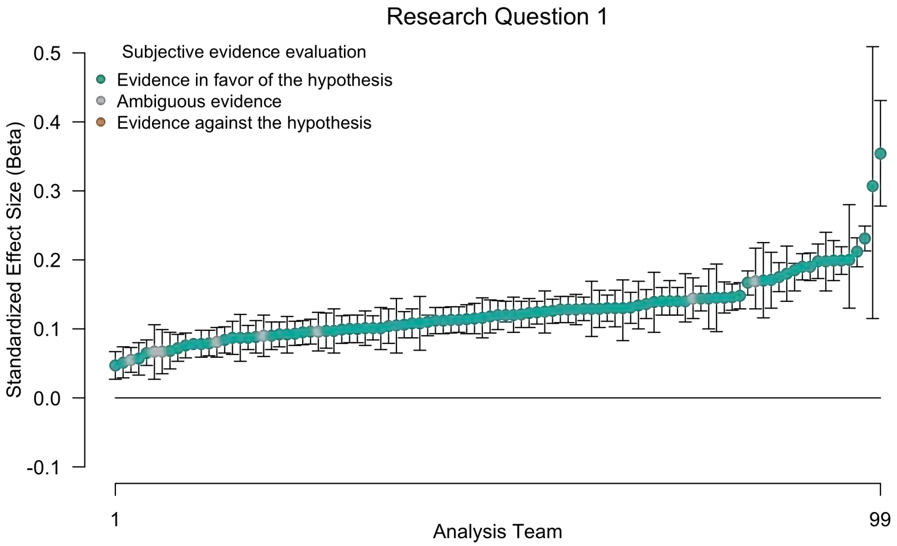


**Figure 9.2:** Beta coefficients for the effect of religiosity on self-reported well-being (research question 1) with 95% confidence or credible intervals. Green/blue points indicate effect sizes of teams that subjectively concluded that there is *good evidence for a positive relation* between individual 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.

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 9.5 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 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.

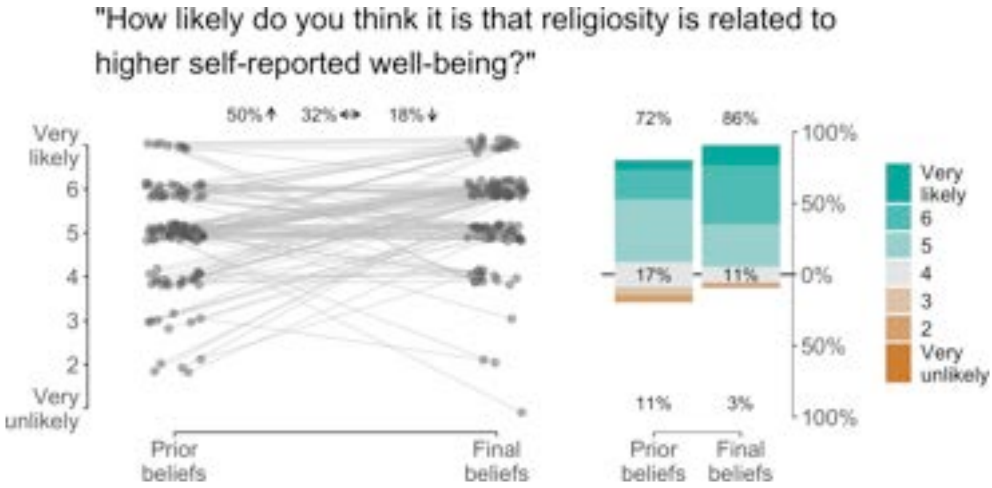


**Figure 9.3:** Responses to the survey questions about the likelihood of hypothesis 1. The left side of the figure shows the change in beliefs for each analysis team. Fifty percent of the teams considered the hypothesis somewhat more likely after having analyzed the data than prior to seeing the data, 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.

### 9.7.2 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.

#### 9.7.2.1 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 9.6, 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.

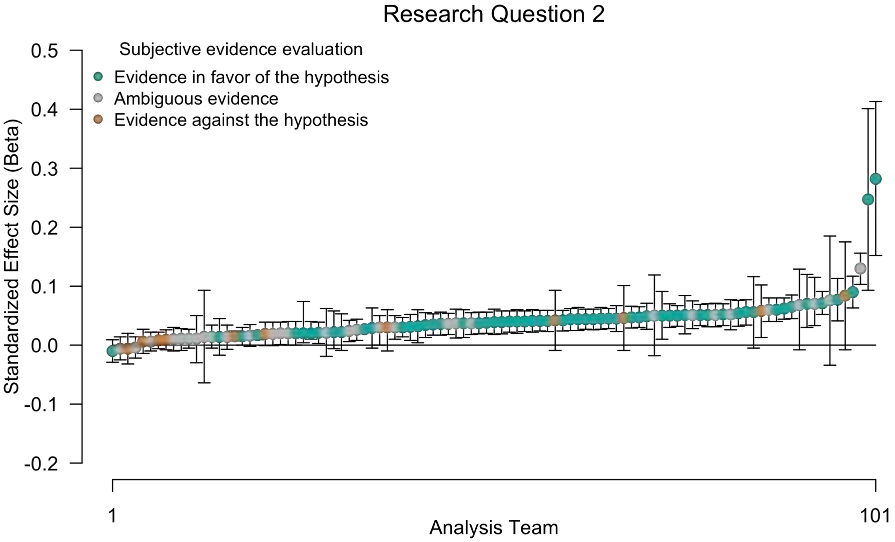


**Figure 9.4:** Beta coefficients for the effect of cultural norms of the relation between religiosity and self-reported well-being (research question 2) with 95% confidence or credible intervals. Green/blue points indicate effect sizes of teams that subjectively concluded that there is *good evidence 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.

#### 9.7.2.2 Analytic Approaches

Table 9.3 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 9.6, 89% of the teams indicated to be (very) confident, 4% was neutral, and 7% was not (at all) confident.[6]

#### 9.7.2.3 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

[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.

**Table 9.3:** 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.

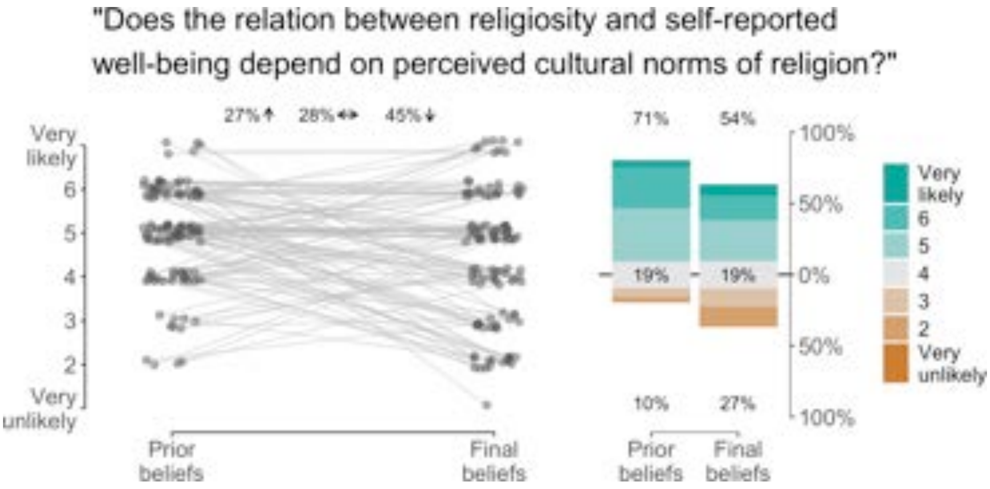


**Figure 9.5:** Responses to the survey questions about the likelihood of hypothesis 2. The left side of the figure shows the change in beliefs for each analysis team. Twenty-seven percent of the teams considered the hypothesis somewhat more likely after having analyzed the data than prior to seeing 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.

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 9.7 shows the included items as dependent variable aggregated over all teams for research question 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

[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.

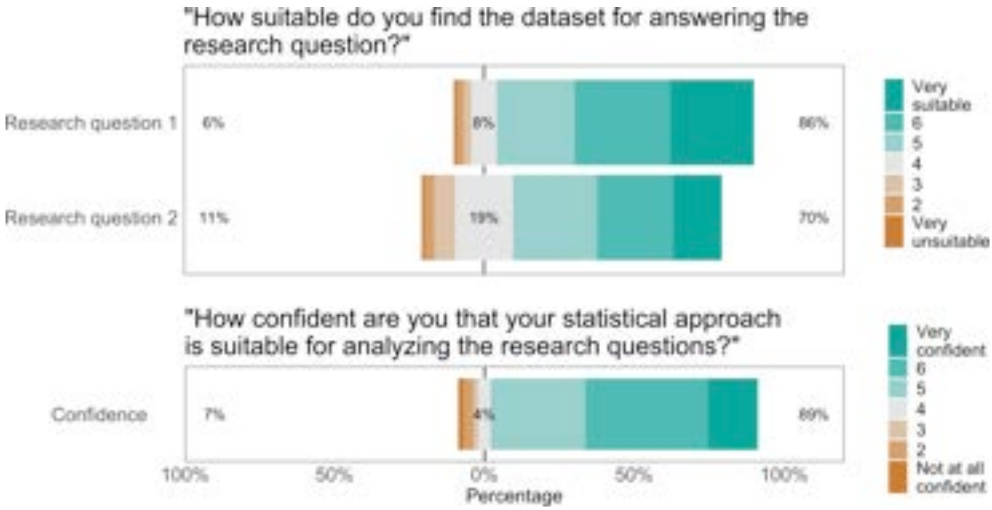


**Figure 9.6:** Responses to the survey questions about the suitability of the dataset for answering the research questions (top) and the teams' confidence in their analytic approach (bottom). For question 1, the top bar represents the teams' answers with respect to research question 1 and the bottom 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.

same dependent variable for research question 1 and 2.[8] In the online Appendix, we show the included items separately for each team (see https://osf.io/9kpfu/).

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 9.8 shows the included items as independent variable aggregated over all teams for research question 1 and research question 2. In the online Appendix, we show the included items separately for each team.

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*.

[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.

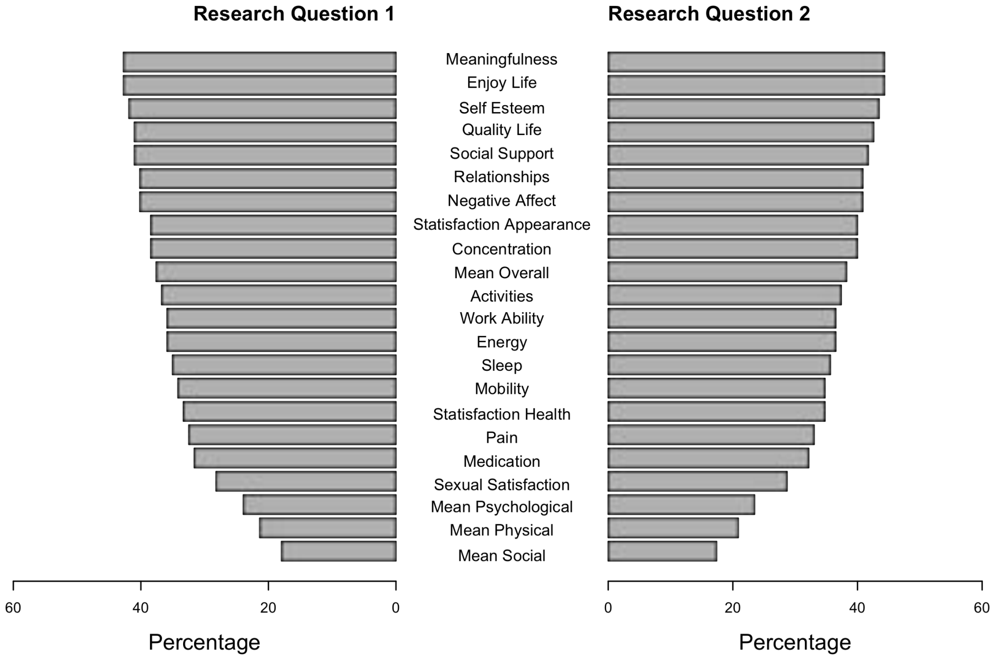


**Figure 9.7:** Items included as dependent variables for research question 1 (on the left) and research question 2 (on the right). Note that the averages for the well-being subscales ('Mean Psychological', 'Mean Social', 'Mean Physical'), as well as the overall average ('Mean Overall') were provided by the MARP team.

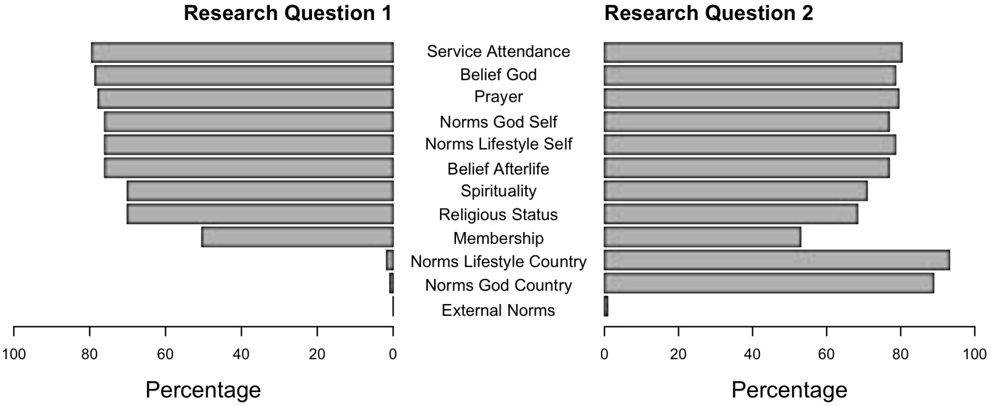


**Figure 9.8:** Items included as independent variables for research question 1 (on the left) and research question 2 (on the right).

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 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 9.9 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 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.

[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).

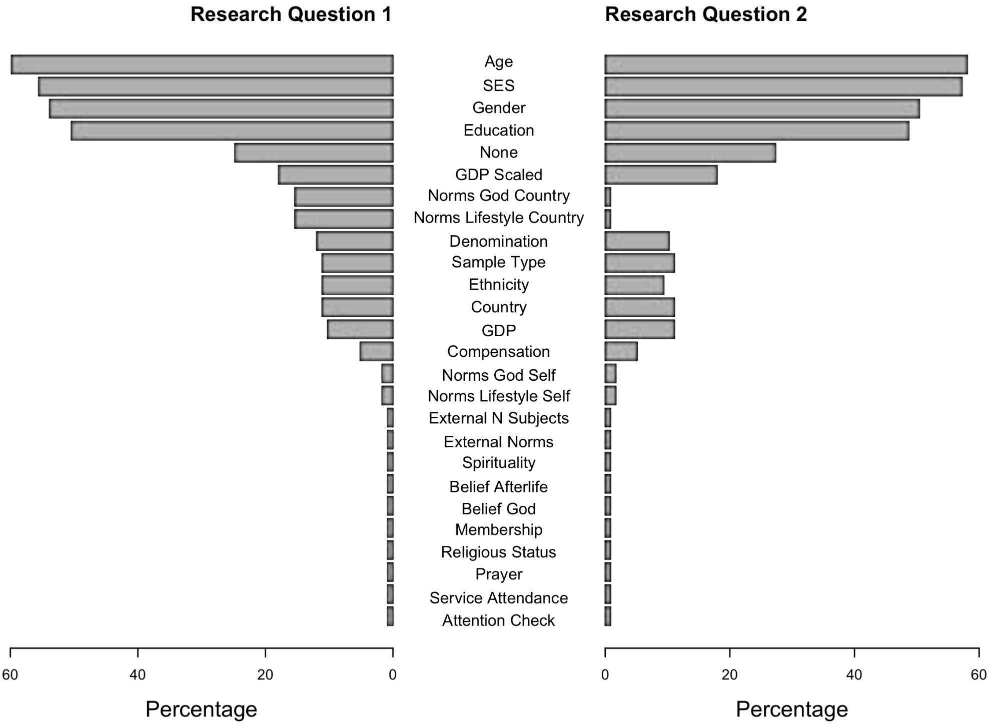


**Figure 9.9:** Items included as covariates for research question 1 (on the left) and research question 2 (on the right). Variables indicated as 'external' refer to covariates that are based on data not provided by the MARP team.

#### 9.7.2.4 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 9.10). 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.

#### 9.7.2.5 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, et al., 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

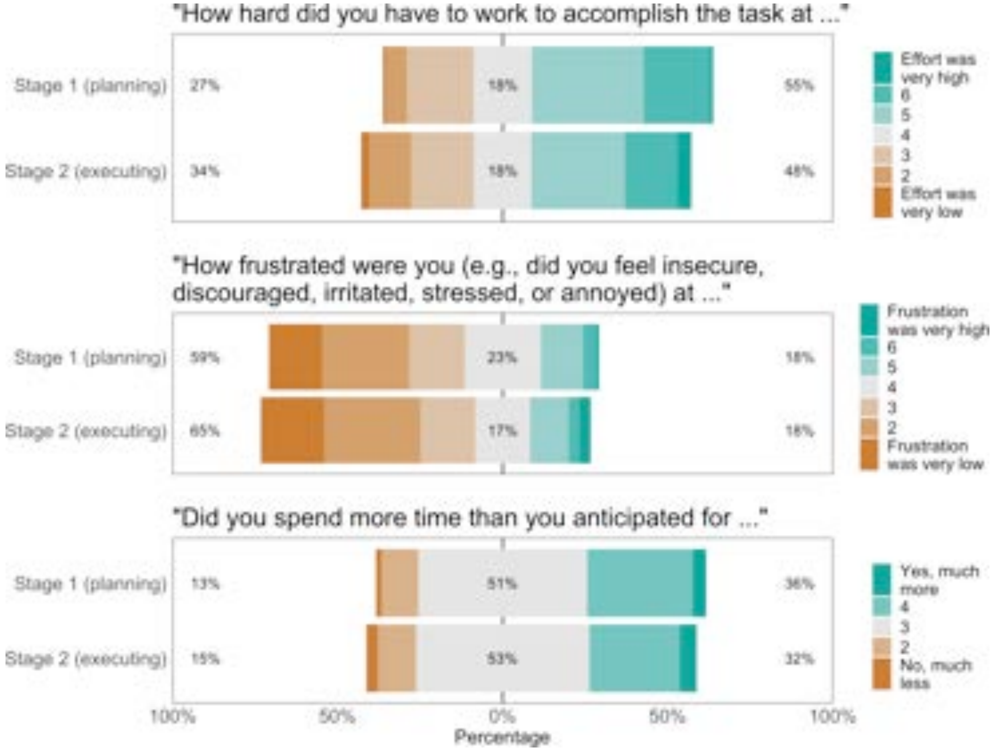


**Figure 9.10:** Responses to the survey questions about effort (top), frustration (middle), and workload (bottom). For each question, the top bar represents the teams' answers about stage 1 (planning) and the bottom bar represents the teams' answers about stage 2 (executing). For each item, the number 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.

question 2 based on the results of their analyses. Note, however, that based on the scatterplot in Figure 9.11D, 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 ($\text{BF}_{10} = 0.13$; $\text{BF}_{01} =$

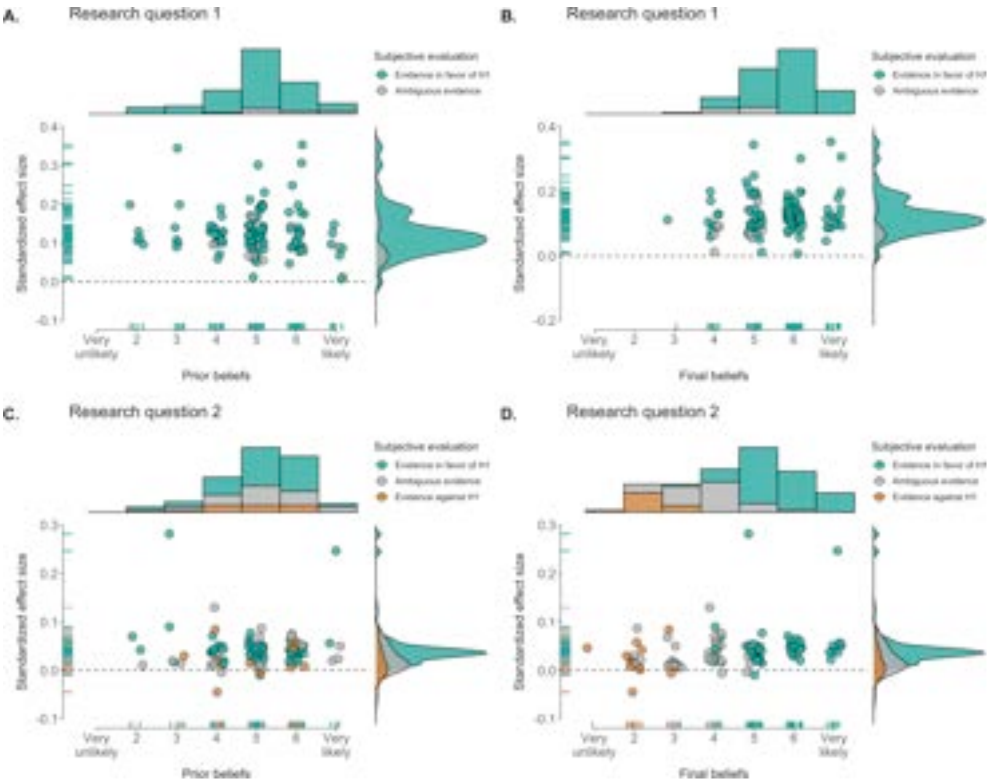


**Figure 9.11:** Reported effect sizes (beta coefficients) and subjective beliefs about the likelihood of the hypothesis. **A.** shows the relation between effect size and prior beliefs for research question 1, **B.** shows the relation between effect size and final beliefs for research question 1, **C.** 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.

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 (BF$_{10}$ = 0.48; 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 (BF$_{10}$ = 0.12; 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 (BF$_{10}$ = 0.16; BF$_{01}$ = 6.41, $\rho_s$ = -0.08, 95% credible interval [-0.24, 0.09]). See Figure 9.12 for scatterplots of the data.

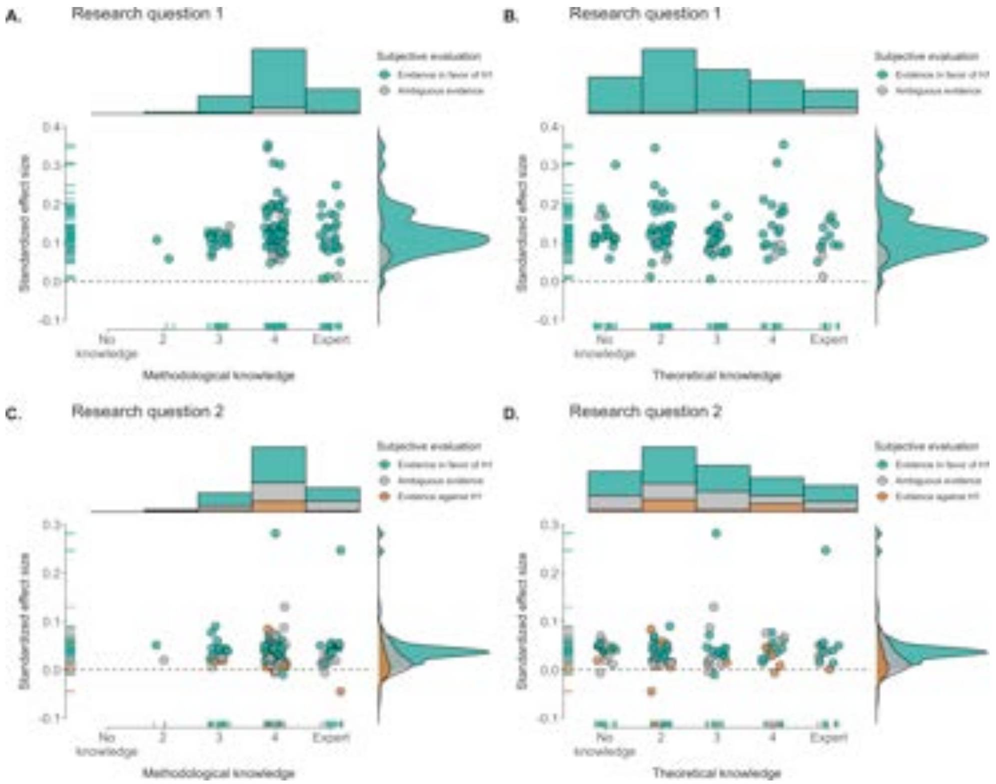


**Figure 9.12:** Reported effect sizes (beta coefficients) and self-reported team expertise. **A.** shows the relation between effect size for research question 1 and methodological knowledge, **B.** shows the relation between effect size for research question 1 and theoretical knowledge, **C.** shows 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.

## 9.8 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.

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, Sarafoglou, van Elk, et al., 2022) reported in Chapter 10. There, we also address the commentaries written by some of the analysis teams.

# 10

# Many-Analysts Religion Project: Reflection and Conclusion

In the main article on the Many-Analysts Religion Project (MARP) reported in Chapter 9 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 (Hoogeveen, Sarafoglou, Aczel, et al., 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.

This chapter has been adapted from: Hoogeveen, S., Sarafoglou, A., van Elk, M., & Wagenmakers, E.-J. (2022). Many-Analysts Religion Project: Reflection and conclusion. *Religion, Brain & Behavior*.
The first two authors contributed equally.

## 10.1 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 et al.; 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 et al.; 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).

## 10.2 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 et al. 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

[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/.

the topic of religion and well-being: since there are many large-scale surveys covering both constructs, this seems a feasible endeavor.

### 10.2.1 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, 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 et al.[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.

[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.

## 10.3 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 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 et al., 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 et al. 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 in Chapter 9 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).

## 10.4 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 on 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.

# 11

# Comparing Analysis Blinding With Preregistration in the Many-Analysts Religion Project

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.

## 11.1 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 con-

This chapter has been adapted from: Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E.-J. (2022). Comparing analysis blinding with preregistration in the many-analysts religion project. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.31234/osf.io/6dn8f.

ducting 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 et al., 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 et al. (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., 2017).

An alternative to preregistration is analysis blinding (Dutilh, Sarafoglou, et al., 2019; MacCoun & Perlmutter, 2015, 2018; MacCoun, 2020). 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, et al., 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.

### 11.1.1 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; Hoogeveen, Sarafoglou, Aczel, et al., 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 their change of beliefs on the two MARP research questions.

### 11.1.2 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.

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.

## 11.2 Disclosures

### 11.2.1 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

**Table 11.1:** 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/ |

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 manuscript.

### 11.2.2 Data and Materials

Table 11.1 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/.

### 11.2.3 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 Hoogeveen, Sarafoglou, Aczel, et al. (2022).

### 11.2.4 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.

## 11.3 Methods

### 11.3.1 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.

### 11.3.2 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.

### 11.3.3 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).

### 11.3.4 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.

### 11.3.5 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.

#### 11.3.5.1 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).

#### 11.3.5.2 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.

#### 11.3.5.3 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, Haaf, et al., 2022; Hoogeveen & van Elk, 2021). 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.

#### 11.3.5.4 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 https://osf.io/qdzwn/). 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.*

#### 11.3.5.5 Dependent Variables: Experienced Workload, Effort, 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]

#### 11.3.5.6 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

[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.

the measure "Anticipated Workload" by summing up the indicated values for stage 1 and stage 2 for each team.

#### 11.3.5.7 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.

### 11.3.6 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 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.

### 11.3.7 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, 2019).

To test hypothesis 4, we fitted two zero-inflated Poisson regression models as defined by Lambert (1992) and implemented in McElreath (2020). This model assumes that with probability $\theta$ 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($\lambda$) distribution. The first model included "analysis method" as predictor, the second model did not. McElreath (2020) expressed the logit-transformed parameter $\theta'$ 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 (2020) expressed the log-transformed parameter $\lambda'$ 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., 2020; Gronau, Sarafoglou, et al., 2017). 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 et al., 1992; Hoijtink, 2011; Klugkist et al., 2005; Klugkist, 2008; Sedransk et al., 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 Development Core Team, 2004).

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-

**Table 11.2:** 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.

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.

## 11.4 Results

### 11.4.1 Sample Characteristics

The career stages and research backgrounds featured in each team are shown in Table 11.2. As apparent from Figure 11.1, 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).

### 11.4.2 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.

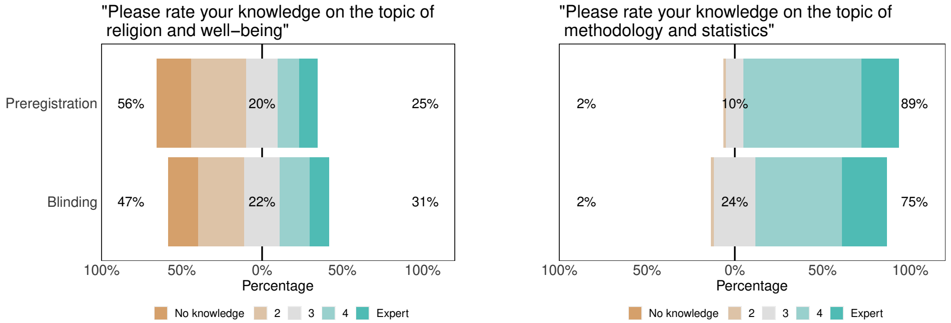


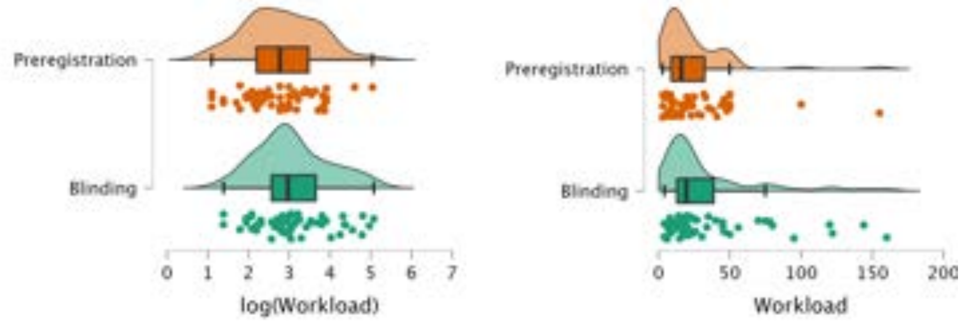
**Figure 11.1:** Responses to the survey questions on the teams' reported knowledge regarding religion and well-being (left panel) and knowledge regarding methodology and statistics (right panel). In each panel, the top bar represents responses from teams who preregistered and the bottom bar represents 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.

### 11.4.3 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 11.2 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 11.3 illustrates the responses of the reported workload separately for stage 1 and stage 2. The difference in total workload spend was the largest in 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.



**(a)** Raincloud plot for log workload **(b)** Raincloud plot for workload

**Figure 11.2:** Reported total workload of stage 1 and stage 2 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. 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.

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 11.4 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 11.4 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 11.3.

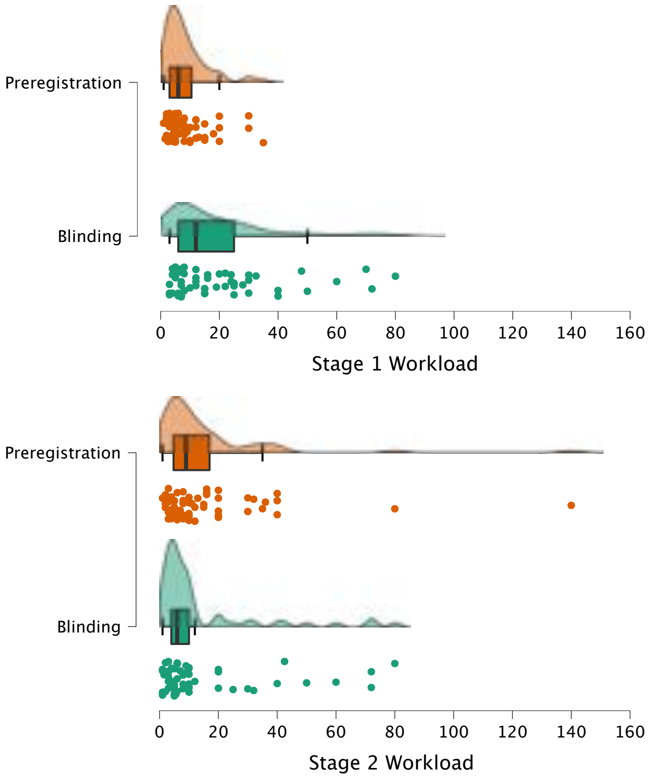


**Figure 11.3:** Reported total workload of stage 1 (top) and stage 2 (bottom) for each analysis team. The upper panel shows (in orange) responses of teams in the preregistration condition. The lower panel shows (in green) responses of teams in the analysis blinding condition. In stage 1, teams required 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.

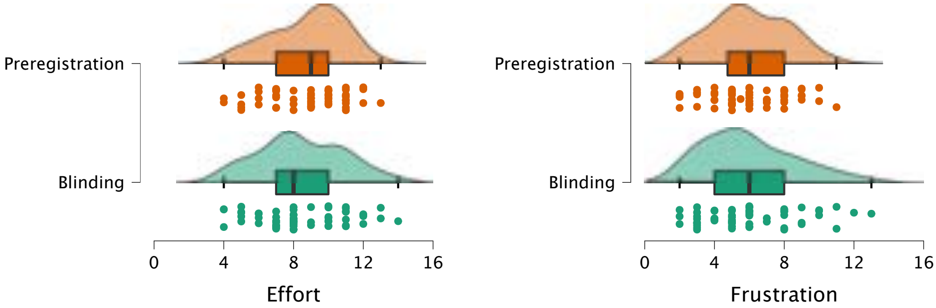


**Figure 11.4:** Responses to the survey questions about the perceived effort (left panel) and frustration (right panel) of planning and executing the analysis. The top panel shows responses of teams in the preregistration condition. The bottom panel shows responses of teams in the analysis blinding condition. The data suggests no or moderate evidence on whether analysis blinding was perceived as less effortful and frustrating, respectively. Points are jittered to enhance visibility.

We collected strong evidence in favor for our hypothesis, that is, $\mathrm{BF}_{r0} = 11.40$. The estimated probability that a team would deviate from their analysis plan was almost twice as high 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).

### 11.4.4 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, $\mathrm{BF}_{+0} = 1.511$.

## 11.5 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.

**Table 11.3:** 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.

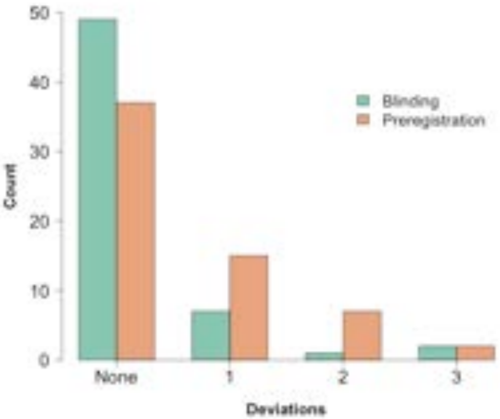


**Figure 11.5:** Reported deviations from planned analysis per condition. The green bars represent teams in the analysis blinding condition, the orange bars represent teams in the preregistration condition. More teams in the analysis blinding condition reported no deviations from their planned analysis and if they had deviated, they did so on less aspects than teams in the preregistration condition.

## 11.6 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 could 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" (Wicherts et al., 2016, p. 2). The quality of the submitted preregistrations could be rated with the coding protocol used by 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 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.

*Religion is a culture of faith; science is a culture of doubt.*

Richard P. Feynman

# 12 General Discussion

Religion is ubiquitous; the majority of the world's population practices a religion, wars have been fought over religious disputes, and hospitals and charities have been established through religious institutions. While the perceived influence of religion seems declining in the West (Poushter & Fetterolf, 2019), globally, religion is actually on the rise (Sherwood, 2018); this is mostly due to higher birthrates among religious individuals compared to their secular counterparts (Pew Research Center, 2015). To many people, religion remains an important factor in their daily lives (Pew Research Center, 2018).

Religion is sometimes considered an evolutionary puzzle (e.g., Bloch, 2008; Irons, 2001), since its costs appear larger than its immediate benefits: religious behaviors and rituals demand time, effort, and resources to be sacrificed and do not seem to have a direct fitness advantage. Over the last decades, scholars in the psychology and cognitive science of religion have been trying to solve this puzzle. Theories explaining the origin and function of religion can roughly be categorized into two classes: religion as a cognitive byproduct and religion as an adaptation. The cognitive byproduct accounts hold that religion originated as a spandrel of general psychological mechanisms, such as mentalizing, agency detection, intuitive dualism (J. L. Barrett, 2012; Bering & Bjorklund, 2004; Bloom, 2007; Boyer, 1994, 2001): On this account, humans are "born believers". Others argue that religion is an adaptation that proved evolutionary beneficial at some point in our species' history or presently: belief in moralizing Gods may promote prosociality and cooperation (Norenzayan, 2013; Norenzayan et al., 2016; Purzycki et al., 2016), (religious) rituals may be conducive to prosociality and sustaining social complexity (Sterelny, 2018; Xygalatas, Mitkidis, et al., 2013), and the costs of religious participation may serve to signal reciprocal altruism and commitment (Bulbulia, 2008), which boost credibility and cultural learning (Gervais, Willard, et al., 2011; Henrich, 2009). In addition to these evolutionary theories with 'ultimate explanations' (Mayr, 1961) for religion, psychologists have also offered 'motivational' theories with 'proximate explanations' of how religion fulfills basic psychological needs: belief in a heavenly afterlife may ease death anxiety (Greenberg et al., 1995; Jonas & Fischer, 2006; Jong et al., 2012), religious beliefs and behaviors may provide meaning in life (Park, 2005), or belief in a God intervening with earthly life may fulfill the need for structure and control (Kay et al., 2008).

In the current dissertation, we aimed to contribute to the scientific inquiry of the

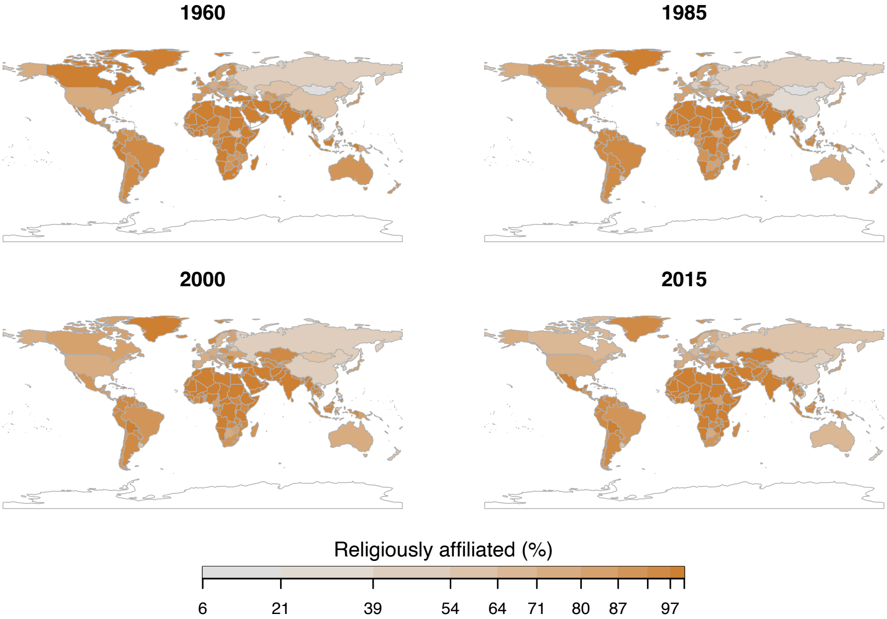


**Figure 12.1:** Global religious identification in 1960, 1985, 2000, and 2015. Data retrieved from the Association of Religion Data Archives (ARDA; Brown and James, 2019).

fascinating phenomenon that religion is. We did so not by proposing new theories or adding hypotheses, but by rigorously testing influential existing ones. Specifically, we tried to add to previous research by reexamining empirical effects that had been reported to support either of these theories, while (1) including large and diverse samples, (2) applying open science practices and Bayesian modeling techniques (3) conducting replications of key effects plus potential alternative explanations or effects (e.g., correlational instead of experimental effects), (4) critically assessing and visualizing patterns in the raw data, and (5) applying new tools to assess robustness (e.g., a many-analysts approach, analysis blinding).

## 12.1 Replicability in the Psychology of Religion

So what do we conclude about replicability in the psychology of religion? Clearly, the work described in this dissertation is far too limited to answer this question definitively. The unsurprising conclusion we can draw is: some effects are replicable and some are not. We managed to successfully (partially) replicate some effects: we found that religiosity was positively related to self-reported well-being (Chapter 9, 10); religiosity was predictive of the tendency to make post-mortem continuity judgments of psychological states, in particular for mental states (e.g., love) compared to bodily states (e.g., hunger; Chapter 8); religiosity was related to credibility ratings for gobbledegook statements, and a reduced relative difference for those from a scientist compared to a spiritual guru (Chapter 7). At the same time, we obtained convincing evidence for the absence of other effects: an experimental attenuation of personal control does not seem to activate a compensatory mechanism of belief in a controlling God (Chapter 4); neural markers of cognitive conflict processing do not seem to be associated with religiosity (Chapter 5); focusing on one's death does not seem to strengthen one's cultural identity (Chapter 6); visual displays of religion do not generally seem to increase perceived trustworthiness (see Appendix A for an analysis on these unpublished data from the CCRRP).

Thus, across all empirical studies reported in this dissertation, we successfully replicated $^3/_6$ – or $^3/_7$ if we include the ML4 reanalysis. While this replication rate might sound somewhat depressing, it is also unsurprising in the light of large scale replication projects that demonstrated successful replication rates ranging from 36% (Open Science Collaboration, 2015) to 90% (Soto, 2019), with an overall success rate of 64% across 307 replication studies included in systematic or multi-site (e.g., Many-Labs) replication projects (Nosek et al., 2022). Conceptually, we should not expect replicability rates of 100%; these rates would only be possible under an extremely conservative research agenda that only studies 'open doors' and would not generate progress in science. Ideally, the scientific practice should find a balance between conservatism and risky predictions that optimizes the diagnosticity of the data.

In hindsight, it is often tempting to consider research findings either too obvious ("my grandmother could have told you that") or ridiculous on their face ("how on earth could looking at a statue change religious beliefs?"). In Chapter 3 we argued that laypeople's accurate predictions show that the research community has been ignoring common sense a bit too much. For years, psychological scientists have paid a lot of attention to 'sexy' effects that turned out not to replicate and that were in fact also not considered plausible by scientists and non-scientists alike. At the same time, common sense is of course not enough; in science we need evidence to substantiate our claims, the mere fact that something sounds plausible is obviously not sufficient for accepting a claim in science. In the end, even obvious claims still require evidence. We noted that in the context of the bleak reality of the current replication rate in psychological science, laypeople are optimists. At the same time, I am optimistic that, as a result of the rapid changes that have been adopted in psychological science, perhaps laypeople's prediction of around 75% successful replications is in fact more realistic right now.

Importantly, since we used Bayesian inference, we could actually substantiate the absence of some of the effects of interest. That is, rather than not being able to reject the null-hypothesis, we actually obtained evidence supporting the null-hypothesis for some of the focal effects. In other words, even in the cases of null-results, we did learn something from the data.

## 12.2 Putting the Findings in Broader Perspective

So how do these findings fit into the broader context of theories on the correlates and consequences of religion? In other words, how do our findings speak to what characterizes religious believers and what function religion serves for them? In the following sections, I will relate our findings to two broad lines of research within the field, namely the cognitive angle on religion and error monitoring and the (social) psychological angle on religion fulfilling basic psychological needs. These theories mostly concern proximate explanations of religion, although the work on mind-body dualism has also been viewed through the evolutionary adaptationist or by-product lens (Bering, 2006; Boyer, 2006; Pyysiäinen, 2006).

### 12.2.1 Mechanisms Underlying Religion

You might recall the 'silly' experiment from Chapter 1, which showed that participants who viewed the sculpture of The Thinker reported lower religious beliefs than participants who viewed the sculpture of Diskobolos (a discus thrower) because analytic thinking was primed by the former. This experiment illustrates a large line of research focused on how religious beliefs are associated with reduced analytic thinking, skepticism, and error monitoring. For instance, dual-process models assume that religious individuals have a stronger tendency to rely on intuition rather than analytic reasoning (Pennycook et al., 2012; Pennycook et al., 2020; Risen, 2016; Shenhav et al., 2012). New work refines this theory and suggests that rather than a preference for intuitive thinking, a reluctance to question one's prior beliefs through actively open-minded thinking might be a better predictor of religious beliefs (Bronstein et al., 2019; Newton et al., 2021). This notion fits well with the predictive processing model by van Elk and Aleman (2017) and the cognitive resource depletion model by Schjoedt et al. (2013), which similarly assume that religious believers are less likely to override prior beliefs in case of conflicting sensory information; instead, the conflict between prior beliefs and sensory input is resolved by assigning more weight to (religious) priors and suppressing the influence of error signals.

The work presented in this dissertation provides mixed evidence with regard to the religion – skepticism link. On the one hand, we could not replicate the correlation between religiosity and conflict sensitivity at the behavioral nor at the neural level: religious believers did not display impaired performance on a cognitive control task, nor did they show reduced activity in the anterior cingulate cortex (ACC) in response to errors or conflicting information (Chapter 5). On the other hand, in the cross-cultural study we found that religiosity was positively related to credibility judgments for nonsense statements: for both gobbledegook from a scientist and from a spiritual guru, religious believers were less skeptical than non-believers (Chapter 7). As suggested in Chapter 5, this inconsistency might emphasize the context-sensitivity of the relation: while religiosity is characterized by a reduced tendency for analytic thinking and questioning incoming information in some contexts, religious believers are not *generally* gullible or insensitive to errors. Rather, in instances that implicate worldview beliefs or epistemic attitudes such as trust in authorities or loyalty to one's beliefs in the face of contradicting evidence, religiosity could be predictive. In other words, we might find that religious believers show reduced ACC activity when listening to a religious authority (Schjoedt et al., 2011) or when performing a cognitive control task with religiously prohibited stimuli (e.g., alcohol; Good et al., 2015), but not when listening to their neighbour talk about his stamp collection or selecting the correct gender of faces in a Stroop task. Especially in neuroscience studies of religion, ecological validity and staying close to the topic of interest, namely religious beliefs and experiences, appears crucial (cf. Schjoedt & van Elk, 2019).

### 12.2.2 Function of Religion

With regard to the function of religion, our work tentatively suggests that religion may provide meaning and comfort to believers, and perhaps a sense of control for some. First, religiosity was most strongly related to the item of the well-being scale assessing to what extent life is experienced as meaningful (Chapter 10). Second, religiosity is –unsurprisingly– associated with implicit afterlife beliefs: religious believers are more likely than non-believers to indicate that a deceased person may still be able to feel love and desire, have knowledge, and hear the voices of their loved ones (Chapter 8). These afterlife beliefs may provide comfort by implying that important social relations do not cease to exist upon biological death or that (mental) life itself simply does not end (Van Tongeren et al., 2017). Note, however, that we failed to find evidence for the key experiment of Terror Management Theory, which holds that fear of death should strengthen one's cultural identity, since cultural identities –including religious identities– can provide symbolic immortality (Chapter 6; Greenberg et al., 1994). It has been argued that religion is an especially attractive defense strategy against death anxiety, as it offers both symbolic immortality in the sense of cultural membership and literal immortality in the sense of an afterlife (Dechesne et al., 2003; Jong et al., 2012). Of course, we did not investigate to what extent mortality salience enhances religious beliefs in particular. However, given the non-replicability of various experimental manipulations aimed at shifting deep-grained beliefs, I'm not too optimistic that it would – the existing evidence is also mixed at best (Jong, 2021; Jong et al., 2012; Osarchuk & Tatz, 1973; Vail et al., 2012). Third, and relatedly, we did not find that an experimental manipulation of personal control affects religious beliefs as a source of external control (Chapter 4). In the US, though, the experience of control in one's life was negatively related to belief in a controlling God, suggesting that religion may indeed serve a purpose of providing a sense of order and control in the world. In the Netherlands, on the other hand, this relation was absent.

In short, our findings suggest that religious believers may indeed be less likely to skeptically assess new information, though probably not at the level of the neural response to cognitive errors or low-level perceptual conflict. In addition, religion may fulfill basic psychological needs for adherents such as providing meaning and a sense of order and control in life, and perhaps a way to come to terms with death. However, these conclusions are all based on cross-sectional, correlational findings, which do not allow for direct causal claims.

## 12.3 The Replication Script Revisited

In the following paragraphs, I will further discuss the main findings and contributions of this dissertation, using a slightly adjusted version of the replication script presented in Table 2.2 on page 25, that is repeated here for convenience (Table 12.1). While this script was designed for replication studies in particular, it may also be applied to any empirical study in general.

Selecting. (1) In the selection phase, we recommended selecting studies with medium chances of replication success. Based on the proportion of empirical studies in this dissertation that were replicated successfully (50%; 3/6) it seems that we

**Table 12.1:** Replication Script Revisited

| Stage | Step | Recommendation |
|---|---|---|
| Selecting | 1. | Opt for studies with medium chances of replication success. |
| | 2. | Consult experts in the field for their suggestions and intuitions. |
| | 3. | Investigate possibilities for replication+ projects that replicate and extend previous work in interesting ways (e.g., boundary conditions or cross-cultural universality). |
| Planning | 4. | Possibly: seek collaboration with colleagues in the field, for instance with authors of the original study. |
| | 5. | In cross-cultural projects: ask for feedback on cultural appropriateness of experimental materials. |
| | 6. | Preregister the research questions, hypotheses, methods, and analysis plan. |
| | 7. | Consider a Registered Report format. |
| Executing | 8. | Collect data. |
| | 9. | Possibly: use analysis blinding to retain flexibility yet avoid biases. |
| | 10. | Conduct analyses according to preregistered plan, and explore data for interesting patterns. |
| Reporting | 11. | Visualize the (raw) data. |
| | 12. | Write up results and invite discussions from scholars in the field. |
| | 13. | Share annotated data and code. |

*Note.* This script was inspired by the summary of guidelines reported in van Doorn, van den Bergh, et al. (2020).

succeeded in including both replication failures and successes – although we cannot conclude that the individual studies had a 50% prior chance of replication, of course. In Chapter 3 we suggested that laypeople's estimates of the plausibility of research effects may be used to quantify these prior odds, both in the context of new research hypotheses and replications. While we did not directly measure prior odds for the empirical studies reported in this dissertation using this method, in the many-analysts religion project (MARP) we did ask the analysts how likely they considered the hypotheses of interest *before* having seen the data, on a 7-point Likert scale ranging from "very unlikely" to "very likely". We found that 72% thought it plausible that (1) religiosity is positively related to self-reported well-being and 71% thought it plausible that (2) this relation is moderated by the perceived desirability of religion in a given country. The optimistic predictions of the analysts turned out to be correct: the first hypothesis received almost unanimous support and the second was corroborated by a 2/3 majority. Future studies may incorporate the forecasting of results in a more systematic way and potentially use the obtained prior odds to inform the analysis. For instance, effects that are deemed highly implausible might require more empirical evidence to convince the skeptic audience than claims that obey common sense.

(2) We encouraged consultation of experts in the field. For the CCRRP, we convened with the 'advisory board' to select the package of studies worth replicating.

This resulted in a bundle with one correlational study and three experiments, including both within- and between-subjects manipulations. Three of those applied existing frameworks: the much-debated well-being–religiosity association (Chapter 9), the mind-body dualism vignette design (Chapter 8), and the religious badges photo design. The last study employed a new design assessing source credibility effects in the context of science and spirituality (Chapter 7).

(3) Across all studies, we attempted to replicate the key effect of interest and extend the design in a way that would allow for secondary inferences. For example, in the direct replication of the compensatory control effect in Chapter 4, we included a trait measure of personal control, in addition to the experimental manipulation aimed at shifting state personal control. This allowed us to adopt an individual differences approach and correlate feelings of personal control to belief in a controlling God. In the US, we indeed found evidence that personal control experienced in one's life is negatively related to belief in a controlling God, suggesting that for some people, belief in a controlling supernatural entity might compensate for a lack of personal control. In the Netherlands, however, this negative correlation was absent. In addition, in the cross-cultural study on religiosity and well-being, we included a subjective measure of cultural norms of religion at the individual level, allowing for assessment of how an individual's perception of the role of religion in their culture affects the extent to which they reap the benefits of being religious.

Planning. (4) Following the recommendation to collaborate, we invited experts to join the project, both theoretical experts on religion as well as methodological experts. For the study on compensatory control, we collaborated with one of the original proponents of the CCT (i.e., Aaron Kay). Beyond benefiting from his theoretical expertise and experience with setting up the study, the collaboration also improved my subjective experience of conducting a replication study: instead of engaging in an adversarial process of trying to "prove them wrong", it felt like working together to assess the robustness and boundary conditions of the effect of interest. Indeed, in this case, the boundary turned out to be the experimental manipulation itself, as the procedure supposed to shift feelings personal control was ineffective.

(5) We recommended to be particularly sensitive to cultural appropriateness of stimuli in cross-cultural studies. In the CCRRP, we did make some effort to optimize our items in this regard. First, in consultation with anthropologists, we tried to ask concrete questions (e.g., "Imagine you lent this person $50, how likely do you think it is that she will give it back to you?") instead of abstract ones (e.g., "How trustworthy do you find this person?"), as concrete items are typically less context-sensitive and vulnerable to reference-group effects (Heine et al., 2002). Second, we tried to use validated measures that have been applied cross-culturally, ideally with existing translations. Specifically, we used the well-being survey from the World Health Organization (WHOQOL Group, 1998) and the religiosity items from the World Values Survey (World Values Survey, 2010). However, some cultural specificities might still be lost in this approach. For instance, one of the analysts in the MARP from Israel was highly surprised to see that only 11.6% of the Israeli sample identified as Jewish (see Table 8.2 on page 161). In our survey, participants first indicated if they belonged to a religious group and only when they answered affirmatively, they could specify their religious group, including being Jewish. According to the Israeli analyst, many

people do not consider themselves religious but they do identify as Jewish, which the setup of our study failed to capture.

(6) We preregistered the hypotheses, materials, and analysis plans for all empirical studies (see https://osf.io/xvqg2/, https://osf.io/xtasg, https://osf.io/8hwdv, https://osf.io/usgr9, https://osf.io/2cdht). This is not to say that we could seamlessly follow our anticipated plans. In fact, all but two studies (reported in Chapter 4 and 5) involved at least one deviation from the preregistration. Although the ubiquity of deviations (Claesen et al., 2021; Heirene et al., 2021) could be considered a drawback of preregistration (Sarafoglou et al., 2021), in practice, I never experienced any problems or criticism about not being able to exactly adhere to the preregistered plan.[1] In my experience, transparency about and a reasonable justification of any deviations is generally acceptable and does not jeopardize the confirmatory status of the study (unless you decide to change your hypotheses, I guess). A practical solution to the almost inevitable analytic deviations may be to report the superior but adjusted analysis in the main text and add the preregistered inferior one in the supplemental materials or as a robustness check, as we did in Chapter 7 and 8 when we changed the analytic strategy and prior settings.

(7) We recommended considering a Registered Report format, in which the preregistration is embedded in the peer review procedure. That is, the introduction, methods section, and analysis plan are written and submitted to a journal prior to data collection. The reviewers then evaluate the theoretical rationale for the study and the planned design and analysis. Upon approval, the manuscript receives "in principle acceptance", the data can be collected, and the manuscript is published no matter the results. Beyond eliminating publication bias, this format also has the added benefit that reviewers can comment on the design of the study at a stage when changes can still be implemented. For instance, in the study on compensatory control and belief in God (Chapter 4), we initially planned to recruit the American participants via Amazon Mechanical Turk and the Dutch participants via a representative panel. A reviewer, however, questioned whether the samples would be comparable (especially regarding socio-economic status and income), hence we decided to use another panel agency with access to both Dutch and American representative samples (i.e., Kieskompas).

Executing. (8) In the execution stage, we recommended collecting data. This was a particularly insightful advice that we followed across all empirical studies. Because, in the words of statistician W. Edwards Deming, "Without data you're just another person with an opinion." Nothing to add here.

(9) We suggested using analysis blinding, which we implemented in Chapter 11. As discussed above, preregistration is an important tool to safeguard the confirmatory status of an empirical study. However, in some cases, designing an exhaustive analysis plan may be cumbersome or unrealistic, for instance, when the analysis includes many contingencies and conditional pathways. Here, the method of analysis blinding –temporarily distorting crucial elements of the data in order to remove the effect of interest– could be a potent alternative. In Chapter 11 we reported an empirical comparison of preregistration and analysis blinding in the many-analysts' take on

[1] On the contrary, one reviewer specifically praised the transparent listing of deviations in the article reported in Chapter 7.

the MARP data. We found that subjective experiences (effort and frustration) and workload were comparable between methods, but that blinding may lead to fewer deviations from the planned analysis, since more aspects of the data can already be accounted for. We believe especially the combination of preregistration and analysis blinding could be a promising way forward. This combined approach forces the researcher to translate their ideas into concrete and specific hypotheses and procedures, yet allow for flexibility in the data analysis without introducing bias. We took this combined approach in our analysis of the preregistration versus blinding comparison reported in Chapter 11. Although this adds another layer to the planning of the study – hence more time and effort – once the researcher gets used to the procedure, the added investment may be trivial. Moreover, in the spirit of "slow science" (Duyvendak, 2019; Frith, 2020), acknowledging the merits of good research practices such as preregistration and analysis blinding may hopefully result in a research culture that truly values quality over quantity in terms of academic output (Benedictus et al., 2016; Smaldino & McElreath, 2016).

(10) We recommended executing the planned analysis and explore the data for additional interesting patterns. For all analyses reported in this dissertation we used Bayesian statistics. While only hearing the term 'Bayesian inference' may have terrified some readers, I believe Bayes factors are actually rather intuitive (arguably more so than *p*-values). For instance, in Chapter 3, we presented half of our forecasters with study descriptions only and half with study descriptions plus the Bayes factor for the original study. We found that laypeople were considerably better at predicting replicability when they had access to the evidence strength by means of Bayes factors than without this information (67% vs. 59% accuracy). In addition, accessible software such as JASP (JASP Team, 2019) or R packages such as `BayesFactor` (Morey & Rouder, 2018) and `brms` (Bürkner, 2017) allow researchers to easily apply Bayesian statistics themselves.[2]

Furthermore, I believe the flexibility of Bayesian modeling and the intuitiveness of model comparisons with Bayes factors present considerable advantages. For example, in Chapter 6, we quantified the evidence against the mortality salience effect of Terror Management Theory under many different constellations of the data. We found that across 29 out of 33 plausible analysis paths, the data provided evidence against the hypothesis that thinking about one's own death would boost one's cultural identity, compared to watching TV. The strength of the evidence ranged from 1.42-to-1 to 44.69-to-1 against the effect. For the remaining 4 paths, the data provided only anecdotal evidence in favor of the effect, with Bayes factors ranging between 1.11-to-1 and 1.61-to-1. Based on these results, we concluded that there is no evidence for a mortality salience effect. A further extension of this approach could be to fine-tune the multiverse analysis by assigning different weights to the different paths, based on a priori theoretical plausibility. For instance, one might argue, as Chatard et al. (2020) did, that designs in which the original authors were involved and datasets that included only participants of the cultural majority should be assigned more weight in the analysis as they constitute a fairer test of the theory.

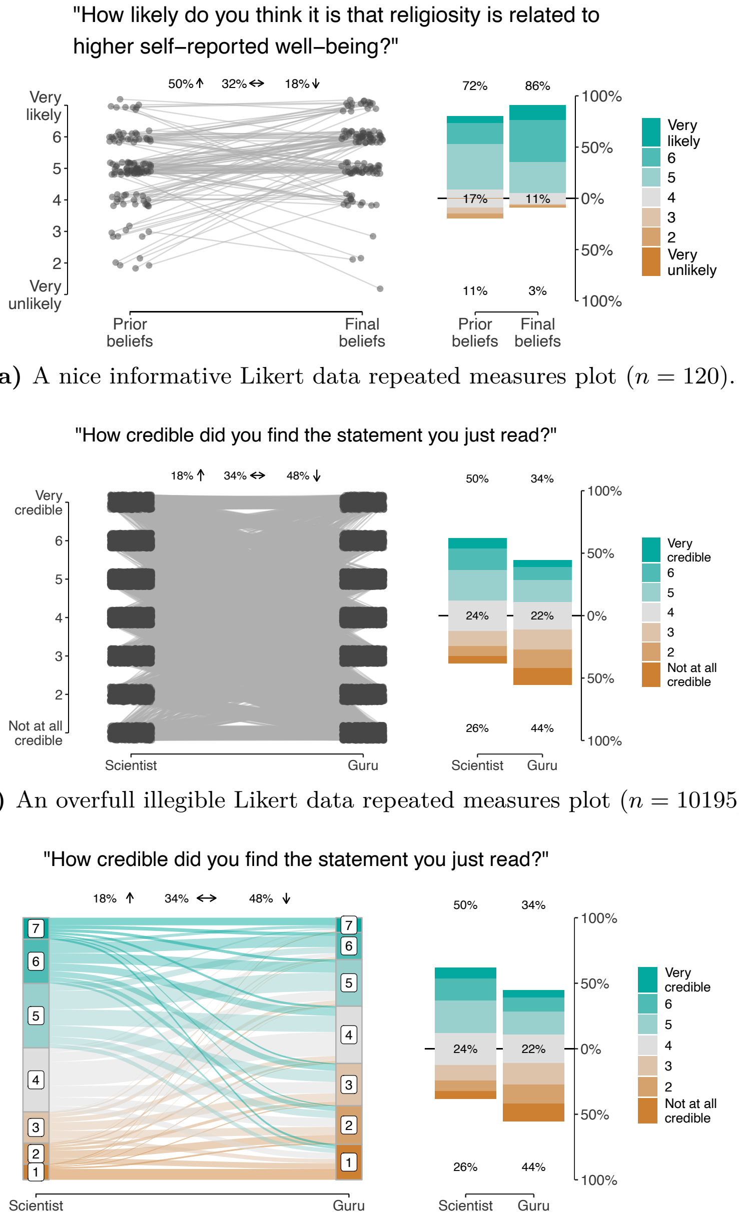
Additionally, in Chapters 7 and 8, we constructed various models reflecting differ-

[2] I can personally attest to that statement; before starting my PhD I'd never even heard of Bayesian statistics (and somehow bluffed my way through the application, sorry EJ) and now I'm generally comfortable with Bayesian analyses and a 100% in 'camp Bayes'.

ent predictions about the structure of the data, beyond variable inclusion. Typically, we compared to what extent the data favor a null-model (“no country does X”), a common-effect model (“all countries do X to the same degree”), a positive effects model (“all countries do X to different degrees”), and an unconstrained model (“some countries do X, other countries do Y”). For the effect of scientific versus spiritual authority on information credibility, for instance, we found that across all 24 tested countries, nonsense from a scientist is considered more credible than the same nonsense from a spiritual guru, but to different degrees: in Turkey the difference was about 1 unit on a 7-point Likert scale, in Japan about 0.4. The interaction between source and rater religiosity, on the other hand, was similar in size across countries (about -0.2; for every standard deviation increase in religiosity the difference in credibility between the scientist and guru became 0.2 Likert unit smaller). In Appendix A, I showed how the Bayes factor model comparison approach could be extended by including a spike-and-slab model that reflects the prediction that “some countries do not X whereas other countries do X to different degrees” (Haaf & Rouder, 2019).

Reporting. (11) With respect to reporting a study, I believe visualization of the raw data is a crucial but often omitted element of presenting one’s results. For instance, in Chapter 7, we found very strong evidence that participants spent more time processing the statement from the scientist than from the spiritual guru ($\text{BF}_{10} = 8050$). Figure 7.4a, however, shows that the effect is tiny (28.3 vs. 27.0 seconds) and in most countries, the 95% credible interval overlaps with zero. Furthermore, while we obtained extremely strong evidence for mind-body dualism in Chapter 8, the statistical evidence alone does not tell the whole story. That is, the pattern of responses visualized in Figure 8.5 op page 168 shows that although mental states (e.g., love) are more likely to be judged to continue after physical death than bodily states (e.g., hunger), across most countries the modal continuity response is zero (out of six). This observation clearly affects the conclusion on whether or not afterlife beliefs and/or mind-body dualism are natural and a cognitive default (cf. Bering, 2002; Bloom, 2005). Specifically, we suggested that *if* people make continuity judgement they seem more likely to believe that certain high-level states may persist after death than body-related states, yet people do not universally exhibit implicit afterlife beliefs in the first place: many people believe that all states cease upon biological death.

Across all empirical studies, we tried to visualize the raw data. A challenge in data visualization is to balance informativeness and interpretability without compromising on aesthetics. I believe “raincloud plots” (Allen et al., 2021) or scatterplots that feature densities (for continuous data) or histograms (for ordinal data) are promising visualization methods. For instance, the plot in Figure 12.2a was inspired by the raincloud plots designed by Allen et al. (2021). However, for Likert scale data, we should not use densities and box plots, since we have spikes of mass at discrete points, rather than continuous data. Therefore, we adjusted the existing raincloud plots, by replacing the densities and boxplots with “stacked bargraphs” that show the distribution of responses across the Likert scale in an intuitive manner. Future developments in data visualisation may focus on tools to clearly display large amounts of data. That is, if we were to use the design of Figure 12.2a to show the data from the CCRRP, for instance on source credibility, the large number of data points would make the plot useless (see Figure 12.2b). Instead, something along the lines of Figure



**(a)** A nice informative Likert data repeated measures plot ($n = 120$).

**(b)** An overfull illegible Likert data repeated measures plot ($n = 10195$).

**(c)** An alternative Likert data repeated measures plot ($n = 10195$).

**Figure 12.2:** Illustration of different plots for Likert scale repeated measures data. Version (a) works well for relatively small datasets, but becomes illegible with larger datasets (b). Version (c) might be an alternative to to visualize Likert data with many data points.

12.2c might be used to display repeated measures Likert scale data for large datasets. I believe standard plotting methods that allow researchers to easily visualize their data without much customization would be highly beneficial for the community.

(12) We recommended to write up the results and invite discussions from scholars in the field. I believe preprint services such as PsyArXiv are a great invention in this regard. Rather than waiting for the entire peer review and publication process (which can take months), researchers can now immediately upload and publicly share a manuscript once finalized. This way, findings can be disseminated and evaluated quicker (Bourne et al., 2017). For instance, the Bayesian reanalysis of the Many Labs 4 (ML4) findings described in Chapter 6 followed from the publication of the ML4 preprint by R. A. Klein et al. (2019) and the reaction by Chatard et al. (2020) that appeared a few weeks later. The combination of open data and code by the ML4 core team and the preprints uploaded on PsyArXiv allowed Chatard et al. (2020) and us to quickly engage in the (public) discussion on the findings. In addition to the time benefit, preprints can provide a record of priority without increasing the (unjustified) chance of being scooped and can boost citation rates once the manuscript is formally published (Bourne et al., 2017; Fu & Hughey, 2019).

A more rigorous and direct method to involve other scholars in the discussion of a study's results is by inviting them to contribute to the data analysis.[3] In Chapter 9, we took such a many-analysts approach to assess the analytic robustness of the hypothesized relation between religiosity and well-being. We found consistent evidence for a positive association that seems robust against a plethora of different analytic decisions and strategies. The almost complete consistency in outcomes among the 120 analyst teams was surprising, especially compared to previous crowd-sourced analysis projects (e.g., Bastiaansen et al., 2020; Botvinik-Nezer et al., 2020; Silberzahn et al., 2018). 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, but are quickly gaining popularity. We believe the two main merits of a many-analysts approach are that it provides (1) an indication of the robustness of the effect on interest, and (2) a concrete demonstration of the variety of theoretical angles and statistical strategies that may be added to researchers' toolboxes. In addition, our impression was that the analysts generally seemed to enjoy participating in the project, an aspect that should not be ignored in doing research.

(13) Our final recommendation was to share annotated data and code. For all of our projects, the data and analysis code are publicly accessible, typically on the OSF. In addition, we have also benefited from others sharing their data and code. As discussed above, the fact that the ML4 team published their data when they released the preprint, allowed the research community to evaluate their findings and conduct their own reanalyses.[4] Sharing and clearly annotating data and code is not only

[3] Another great way to initiate scholarly discussion is to publish a target article and invite commentaries, as we were allowed by the journal *Religion, Brain, & Behavior* for the MARP. However, this is often not possible of course.

[4] We noted that the ML4 data and code was not the easiest to read and reuse, but fortunately the

beneficial for the research community at large, but also for your future self. The code for the Bayes factor model comparison of the hierarchical models used in Chapters 7 and 8, for instance, was largely borrowed from work by Julia Haaf and Jeffrey Rouder, who share all their code for each paper in a clear manner on Github. Having customized this code for the CCRRP data subsequently allowed me to fairly quickly run the analyses and create figures for the unpublished data on religious displays reported in Appendix A.

## 12.4 What's Next (and What Not)?

In terms of future research, I see various promising avenues. In the following, I will discuss some suggestions for theoretical and methodological developments.

### 12.4.1 Theoretical Suggestions

From a theoretical point of view, I believe cross-cultural investigations of influential accounts such as anthropomorphic God concepts (J. L. Barrett & Keil, 1996; J. L. Barrett & Richert, 2003) and teleological reasoning (Kelemen, 1999, 2004) might be worthwhile. Using an approach similar to the CCRRP we could examine the extent to which an anthropomorphic God-image (e.g., God being constrained in space and time) is universally present or only in Abrahamic religions/cultures, or whether people cross-culturally endorse purpose-based explanations of natural phenomena ("giraffes have long necks *so that* they can eat from tall trees") and whether religiosity predicts this type of teleological reasoning. Furthermore, as suggested in Chapter 8, follow-up research might dive into various attributional biases related to afterlife processes, such as the death-positivity bias and the God-serving bias. For example, in the afterlife continuity vignette, we might expect more continuity judgments for positive emotions, particularly self-transcendent and social emotions such as love and gratitude than for negative emotions such as hate and vengeance (i.e., a death positivity bias); more continuity judgments for people who have led good lives than those who committed evil (reflecting belief in a just world), and more religious attributions for successes than for failures (i.e., a God-serving bias).

### 12.4.2 Methodological Suggestions

In terms of methodology, I believe there is considerable room for improvement in (1) thinking clearly about study designs and causal inference (e.g., Bulbulia, 2022; Bulbulia et al., 2021), (2) validating measures within a study, especially in cross-cultural research, and (3) critically assessing data quality related to response sincerity.

Designs and Inference. First, most successful replications reported in the current dissertation where those that assessed associations between religiosity and individual differences in beliefs and perceptions. That is, we found that religiosity is related to higher self-reported well-being, higher perceived credibility of ambiguous information, more implicit afterlife beliefs, and less experienced personal control (only in the US). The question remains, however, what these associations mean; as the old adage

lead author was very responsive to our queries.

goes, correlation does not imply causation. Does religion really make people happy? Does religious service attendance reduce anxiety or does anxiety prevent religious attendance? Are people who feel little control in their lives drawn to religion or do religious beliefs lead to outsourcing control? Do explicit religious beliefs translate into implicit afterlife beliefs (i.e., continuity judgments of psychological states after death) or do intuitive implicit afterlife beliefs make people receptive to religion? Or, are there third variables such as socio-economic status and upbringing that account for all these associations? An obvious reason why (social) psychologists love experiments is that they allow us to draw causal conclusions. However, as suggested by Marcus and McCullough (2021), it might be time to abandon lab-experiments attempting to shift religious beliefs or any correlate of interest (e.g., self-control) as these are simply not replicable. Instead, it may be more fruitful to focus on longitudinal studies, quasi-experimental designs (e.g., comparing students in religious vs. public schools), or other methods for causal inference in the absence of experimental data (Bulbulia, 2022; Pearl, 2019; Rohrer, 2018).

Validating Measures. This is obviously not to say that experimental manipulations related to religion will always fail; we obtained strong evidence that a scientific versus spiritual guru source manipulation affected credibility ratings of gobbledegook statements in Chapter 7. In addition, we found that the framing of a narrative emphasizing religion increased continuity judgments, suggesting that certain contextual manipulations might be effective in changing beliefs or attributions. However, as discussed in Chapter 8, it is not unequivocal that the framing manipulation truly influenced personal beliefs. Alternatively, participants may have been responding as if immersing themselves in a fairy-tale and "playing along" with the task. More generally, it may be worthwhile to include more qualitative measures in future research, in order to better gauge how participants interpret the questions and tasks at hand and to what extent this corresponds to the intended meaning by the researcher. For instance, for the personal control manipulation in Chapter 4 participants had to describe a positive situation in which they either felt in control or lacked control. Descriptions such as "Insurance Visa card payed up for an item not received or ordered" indeed involved little personal control, yet it can hardly be interpreted as a threat to personal control that might –on some unconscious level– instigate belief in divine intervention.[5]

Especially in cross-cultural research, which is becoming more popular in recent years (for good reason), it is important to validate that our measures are measuring what they are supposed to measure. Some concerns have been raised regarding the "measurement schmesurement" attitude in social sciences (Flake et al., 2017; Flake & Fried, 2020): researchers often make little effort to validate their measures, except for the occasional reporting of Cronbach's alpha. In cross-cultural research, the issue of measurement invariance across cultures is arguably crucial for valid inferences yet often ignored (Boer et al., 2018; Fischer & Karl, 2019; Hussey & Hughes, 2020; Ross et al., 2022; Schreiner et al., 2022). That is, in order to make cross-cultural comparisons, one should first check whether a given measure has the same properties and structure across different samples and contexts. With the CCRRP data, we have

[5] Though I do not deny that some people might think of Visa as the devil.

also been guilty of ignoring this important precondition.[6] I believe it is important for future cross-cultural research to validate that, say, a religiosity measure in Japan is interpreted in a similar fashion as in the US. Practical tools to assess multi-group invariance (e.g., Fischer & Karl, 2019) may be instrumental in this development.

Data Quality (Catching Trolls). Finally, research on supernatural beliefs may benefit from paying close attention to data quality. In addition to inattentiveness that can relatively easy be filtered out using attention check items, data quality may also be compromised by so-called “survey trolling”: participants responding insincerely for the sake of being provocative or funny (Lopez & Hillygus, 2018). For instance, Lopez and Hillygus (2018) found that while only 12%-15% of participants indicated a strong conviction in Hilary Clinton’s connection to a child-sex ring, half of those were flagged for being insincere in other parts of the survey (e.g., indicating multiple low-incident items such as having stabbed someone and having smoked before age 8, or admitting dishonest responding). While survey trolling may be less of a concern in measuring religious beliefs than conspiracy beliefs, it should probably be taken into account when assessing broader supernatural or paranormal beliefs, such as belief in ghosts, gnomes, or telepathy[7]. Future surveys on these phenomena might include “catch items” such as the low-incident questions mentioned above, self-reported dishonesty, or made-up conspiracy theories to flag disingenuous response patterns.

## 12.5 Closing Remarks

A personal take-away of conducting the studies reported in this dissertation is that methodology can already give an indication of the replication probability: subtle manipulations aimed at shifting deep-grained religious or cultural beliefs (Gervais & Norenzayan, 2012; Greenberg et al., 1995; Kay et al., 2008)? Probably not so successful (Haaf et al., 2020; Hoogeveen, Wagenmakers, et al., 2018; R. A. Klein et al., 2018). Individual differences in brain anatomy or functional activity patterns related to religious beliefs (Inzlicht et al., 2009)? Wouldn’t bet a lot of money on it (Hoogeveen, Snoek, et al., 2020; van Elk & Snoek, 2020). In general, finding robust individual differences in cognitive tasks and brain data might be futile, as the signal-to-noise ratio is low and hence requires hundreds of trials or thousands of participants (Marek et al., 2022; Rouder, Kumar, et al., 2019).

But despite the replication failures, I personally have never ‘lost faith’ in the scientific practice in general and in the psychology of religion in particular during this journey. The speed at which changes fostering more robust and replicable science are developed and adopted is truly inspiring. Trends towards open science and transparency, team science and international collaborations, crowd-sourced data collection and analysis, and recognition of replication research do not only improve the quality

[6] In fact, we preregistered to investigate measurement invariance for the well-being construct used in Chapter 9, but did not include this in our own analysis for the MARP data. Fortunately, the issue of measurement invariance was discussed by two commentaries on the MARP (Ross et al., 2022; Schreiner et al., 2022).

[7] A Dutch survey from 1985 indicated that 32% of the population expressed belief in telepathy, 12% in the existence of ghosts, and 3% in gnomes and elves, from which 2% based on personal experience. While I am hesitant to make unfounded statements about the Dutch society in the decade before I was born, I’d not be surprised if those rates included some survey trolling.

of science, but also the subjective experience of conducting research, at least for me. I have been amazed and humbled by the knowledge, expertise, and enthusiasm of everyone who contributed in some way to the studies reported in this dissertation. For me, the excitement of conducting projects in collaboration with researchers from all over the world, with various backgrounds and viewpoints strongly outweighs the occasional disappointment over obtaining another null-result. Hopefully, the increased focus on the quality of the research process, the data, and analytic strategy may remove the negative perception of null-results or ambiguous outcomes, which, like it or not, are probably inevitable in a healthy scientific research practice.

It seems that the psychology of religion, like psychological science in general, is not exempt from replication failures. Yet like general psychology, the field is quickly changing, and scholars and journals of religion are embracing preregistration, diversifying samples, collaborative science, and replication studies. In the end, I hope that this dissertation does not leave the reader depressed and disillusioned but rather optimistic and perhaps even inspired to adopt (new) practices such as preregistration, analysis blinding, crowd-sourced analyses, Bayesian hierarchical modeling, and cross-cultural collaborations. Importantly, I hope that we contributed to showing that replication research can be interesting and exciting. And obviously, I happily invite anyone to attempt to replicate the findings in this dissertation.

# Bibliography

Abdel-Khalek, A. M. (2006). Measuring happiness with a single-item scale. *Social Behavior and Personality: An International Journal*, *34*(2), 139–150. https://doi.org/10.2224/sbp.2006.34.2.139

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*, 14. https://doi.org/10.3389/fninf.2014.00014

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*(2), 130–139.

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., … Wagenmakers, E.-J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies (P. Rodgers, Ed.). *eLife*, *10*, e72185. https://doi.org/10.7554/eLife.72185

Agley, J. (2020). Assessing changes in US public trust in science amid the COVID-19 pandemic. *Public Health*, *183*, 122–125. https://doi.org/10.1016/j.puhe.2020.05.004

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, *17*(4), 351–371. https://doi.org/10.1177/1094428114547952

Albrecht, S. L., & Heaton, T. B. (1984). Secularization, higher education, and religiosity. *Review of Religious Research*, *26*, 43–58. https://doi.org/10.2307/3511041

Alcorta, C. S., & Sosis, R. (2005). Ritual, emotion, and sacred symbols. *Human Nature*, *16*(4), 323–359. https://doi.org/10.1007/s12110-005-1014-3

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., van Langen, J., & Kievit, R. A. (2021). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, *4*. https://doi.org/10.12688/wellcomeopenres.15191.2

Allison, S. T., Eylon, D., Beggan, J. K., & Bachelder, J. (2009). The demise of leadership: Positivity and negativity biases in evaluations of dead leaders. *The Leadership Quarterly*, *20*(2), 115–129. https://doi.org/10.1016/j.leaqua.2009.01.003

Alper, S., & Sümer, N. (2017). Control deprivation decreases, not increases, belief in a controlling God for people with independent self-construal. *Current Psychology*, 1–5. https://doi.org/10.1007/s12144-017-9710-9

Anczyk, A., Grzymala-Moszczyńska, H., Krzysztof-Swiderska, A., & Prusak, J. (2019). The replication crisis and qualitative research in the psychology of religion. *The International Journal for the Psychology of Religion*, *29*, 278–291. https://doi.org/10.1080/10508619.2019.1687197

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. https://doi.org/10.1177/0956797611434537

Anderson, M. S., Martinson, B. C., & De Vries, R. (2007). Normative dissonance in science: Results from a national survey of U.S. scientists. *Journal of Empirical Research on Human Research Ethics*, *2*(4), 3–14. https://doi.org/10.1525/jer.2007.2.4.3

Anvari, F., & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology*, *3*(3), 266–286. https://doi.org/10.1080/23743603.2019.1684822

Argue, A., Johnson, D. R., & White, L. K. (1999). Age and religiosity: Evidence from a three-wave panel analysis. *Journal for the Scientific Study of Religion*, 423–435. https://doi.org/10.2307/1387762

Argyris, C. (1975). Dangers in applying results from experimental social psychology. *American Psychologist*, *30*(4), 469–485. https://doi.org/10.1037/h0076834

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*(4), 596–612. https://doi.org/10.1037/0022-3514.63.4.596

Astuti, R., & Harris, P. L. (2008). Understanding mortality and the life of the ancestors in rural madagascar. *Cognitive Science: A Multidisciplinary Journal*, *32*(4), 713–740. https://doi.org/10.1080/03640210802066907

Atkinson, Q. D., & Bourrat, P. (2011). Beliefs about God, the afterlife and morality support the role of supernatural policing in human cooperation. *Evolution and Human Behavior*, *32*(1), 41–49. https://doi.org/10.1016/j.evolhumbehav.2010.07.008

Atkinson, Q. D., Claessens, S., Fischer, K., Forsyth, G. L., Kyritsis, T., Wiebels, K., & Moreau, D. (2022). Being specific about generalisability. *Commentary in MARP special issue.*

Attig, T. (1996). *How we grieve: Relearning the world.* New York: Oxford University Press.

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*, 26–41. https://doi.org/10.1016/j.media.2007.06.004

Bailey, R. C., & Garrou, D. G. (1983). Dating availability and religious involvement as influences on interpersonal attraction. *The Journal of Psychology*, *113*(1), 95–100. https://doi.org/10.1080/00223980.1983.9923562

Bains, S. (2011). Questioning the integrity of the John Templeton Foundation. *Evolutionary Psychology*, *9*(1), 92–115. https://doi.org/10.1177/147470491100900111

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*, 452–454.

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077. https://doi.org/10.1177/0956797616647519

Balkaya-Ince, M., & Schnitker, S. (2022). Advantages of using multilevel modeling approaches for the Many Analysts Religion Project. *Commentary in MARP special issue.*

&

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, *128*, 32–55. https://doi.org/10.1037/0096-3445.128.1.32

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry: The power of response time distributional analyses. *Current Directions in Psychological Science*, *20*(3), 160–166. https://doi.org/10.1177/0963721411408885

Banerjee, K., & Bloom, P. (2013). Would Tarzan believe in God? Conditions for the emergence of religious belief. *Trends in Cognitive Sciences*, *17*(1), 7–8. https://doi.org/10.1016/j.tics.2012.11.005

Banerjee, K., Haque, O. S., & Spelke, E. S. (2013). Melting lizards and crying mailboxes: Children's preferential recall of minimally counterintuitive concepts. *Cognitive Science*, *37*(7), 1251–1289. https://doi.org/10.1111/cogs.12037

Barber, N. (2011). A cross-national test of the uncertainty hypothesis of religious belief. *Cross-Cultural Research*, *45*(3), 318–333.

Barch, D. M., & Yarkoni, T. (2013). Introduction to the special issue on reliability and replication in cognitive and affective neuroscience research. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(4), 687–689. https://doi.org/10.3758/s13415-013-0201-7

Barlev, M., & Shtulman, A. (2021). Minds, bodies, spirits, and gods: Does widespread belief in disembodied beings imply that we are inherent dualists? *Psychological Review*, *128*(6), 1007–1021. https://doi.org/10.1037/rev0000298

Barrett, H. C., & Behne, T. (2005). Children's understanding of death as the cessation of agency: A test using sleep versus death. *Cognition*, *96*(2), 93–108. https://doi.org/10.1016/j.cognition.2004.05.004

Barrett, H. C., Bolyanatz, A., Broesch, T., Cohen, E., Froerer, P., Kanovsky, M., Schug, M. G., & Laurence, S. (2021). Intuitive dualism and afterlife beliefs: A cross-cultural study. *Cognitive Science*, *45*(6), e12992. https://doi.org/10.1111/cogs.12992

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Sciences*, *4*(1), 29–34. https://doi.org/10.1016/S1364-6613(99)01419-9

Barrett, J. L. (2018). Some common misunderstandings about cognitive approaches to the study of religion: A reply to Sterelny. *Religion, Brain & Behavior*, *8*(4), 425–428. https://doi.org/10.1080/2153599X.2017.1323787

Barrett, J. L., & Keil, F. C. (1996). Conceptualizing a nonnatural entity: Anthropomorphism in God concepts. *Cognitive Psychology*, *31*(3), 219–247. https://doi.org/10.1006/cogp.1996.0017

Barrett, J. L., Newman, R. M., & Richert, R. (2003). When seeing is not believing: Children's understanding of humans' and non-humans' use of background knowledge in interpreting visual displays. *Journal of Cognition and Culture*, *3*(1), 91–108. https://doi.org/10.1163/156853703321598590

Barrett, J. L., Richert, R. A., & Driesenga, A. (2001). God's beliefs versus mother's: The development of nonhuman agent concepts. *Child Development*, *72*(1), 50–65. https://doi.org/10.1111/1467-8624.00265

Barrett, J. L. (1998). Cognitive constraints on Hindu concepts of the divine. *Journal for the Scientific Study of Religion*, 608–619. https://doi.org/10.2307/1388144

Barrett, J. L. (2012). *Born believers: The science of children's religious belief*. Simon and Schuster.

Barrett, J. L., & Richert, R. A. (2003). Anthropomorphism or preparedness? Exploring children's god concepts. *Review of religious research*, 300–312.

Barrett, M. (2021). Ggdag: Analyze and create elegant directed acyclic graphs.

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., Chow, S.-M., de Jonge, P., Emerencia, A. C., Epskamp, S., Fisher, A. J., Hamaker, E. L., Kuppens, P., Lutz, W., Meyer, M. J., Moulder, R., Oravecz, Z., Riese, H., Rubel, J., … 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. https://doi.org/10.1016/j.jpsychores.2020.110211

Batson, C. D. (1975). Rational processing or rationalization? The effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology*, *32*(1), 176–184. https://doi.org/10.1037/h0076771

Battiston, P., Kashyap, R., & Rotondi, V. (2020). Trust in science and experts during the COVID-19 outbreak in Italy.

Bek, J., & Lock, S. (2011). Afterlife beliefs: Category specificity and sensitivity to biological priming. *Religion, Brain and Behavior*, *1*(1), 5–17. https://doi.org/10.1080/2153599X.2010.550724

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. https://doi.org/10.1037/a0021524

Benedictus, R., Miedema, F., & Ferguson, M. W. J. (2016). Fewer numbers, better science. *Nature*, *538*(7626), 453–455. https://doi.org/10.1038/538453a

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10.

Berent, I., & Platt, M. (2021). The true "me"—Mind or body? *Journal of Experimental Social Psychology*, *93*, 104100. https://doi.org/10.1016/j.jesp.2020.104100

Berger, J. O. (2006). Bayes Factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences, vol. 1 (2nd ed.)* (pp. 378–386). Wiley. https://doi.org/10.1002/9781118445112.stat00224

Berger, J. O., & Wolpert, R. L. (1988). *The Likelihood Principle (2nd ed.)* Institute of Mathematical Statistics.

Bergin, A. E. (1983). Religiosity and mental health: A critical reevaluation and meta-analysis. *Professional Psychology: Research and Practice*, *14*(2), 170–184. https://doi.org/10.1037/0735-7028.14.2.170

Bering, J. M., Hernández-Blasi, C., & Bjorklund, D. F. (2005). The development of afterlife beliefs in secularly and religiously schooled children. *British Journal of Developmental Psychology*, *23*, 587–607.

Bering, J. M. (2002). Intuitive conceptions of dead agents' minds: The natural foundations of afterlife beliefs as phenomenological boundary. *Journal of Cognition and Culture*, *2*(4), 263–308. https://doi.org/10.1163/15685370260441008

Bering, J. M. (2006). The folk psychology of souls. *Behavioral and Brain Sciences*, *29*(5), 453–462. https://doi.org/10.1017/S0140525X06009101

Bering, J. M., & Bjorklund, D. F. (2004). The natural emergence of reasoning about the afterlife as a developmental regularity. *Developmental psychology*, *40*(2), 217. https://doi.org/10.1037/0012-1649.40.2.217

Bering, J. M., McLeod, K., & Shackelford, T. K. (2005). Reasoning about dead agents reveals possible adaptive trends. *Human Nature*, *16*(4), 360–381. https://doi.org/10.1007/s12110-005-1015-2

Bernts, T., & Berghuijs, J. (2016). *God in Nederland 1966-2015*. Have, Ten.

Betancourt, M., Vehtari, A., & Gelman, A. (2015). Prior choice recommendations.

Birch, S. A. J., Akmal, N., & Frampton, K. L. (2010). Two-year-olds are vigilant of others' non-verbal cues to credibility. *Developmental Science*, *13*(2), 363–369. https://doi.org/10.1111/j.1467-7687.2009.00906.x

Bishop, L. (2009). Ethical sharing and reuse of qualitative data. *Australian Journal of Social Issues*, *44*(3), 255–272. https://doi.org/10.1002/j.1839-4655.2009.tb00145.x

Bloch, M. (2008). Why religion is nothing special but is central. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1499), 2055–2061. https://doi.org/10.1098/rstb.2008.0007

Blok, S., Newman, G., Behr, J., & Rips, L. J. (2001). Inferences about personal identity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 80–85.

Blok, S., Newman, G., & Rips, L. J. (2005). Individuals and their concepts. In W. Ahn, R. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the lab* (pp. 127–149). American Psychological Association.

Bloom, P., & Weisberg, D. S. (2007). Childhood origins of adult resistance to science. *Science*, *316*(5827), 996–997. https://doi.org/10.1126/science.1133398

Bloom, P. (2005). *Descartes' baby: How the science of child development explains what makes us human*. Random House.

Bloom, P. (2007). Religion is natural. *Developmental Science*, *10*(1), 147–151. https://doi.org/10.1111/j.1467-7687.2007.00577.x

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., Krypotos, A.-M., Lerche, V., Logan, G. D., Palmeri, T. J., van Ravenzwaaij, D., Servant, M., Singmann, H., Starns, J. J., Voss, A., Wiecki, T. V., Matzke, D., & Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the Diffusion Decision Model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *87*, 46–75. https://doi.org/10.1016/j.jmp.2018.09.004

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, A. J., Brown, S. D., & Forstmann, B. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, *66*, 115–133. https://doi.org/10.1016/j.cortex.2014.11.019

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713–734. https://doi.org/10.1177/0022022117749042

Bonett, D. G. (2012). Replication-extension studies. *Current Directions in Psychological Science*, *21*(6), 409–412. https://doi.org/10.1177/0963721412459512

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Evaluating the demand for control: Anterior cingulate cortex and conflict monitoring. *Psychological Review*, *108*(3), 624–652.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Botvinik-Nezer, R., Iwanir, R., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Dreber, A., Camerer, C. F., Poldrack, R. A., & Schonberg, T. (2019). fMRI data of mixed gambles from the Neuroimaging Analysis Replication and Prediction Study. *Scientific Data*, *6*(1), 106. https://doi.org/10.1038/s41597-019-0113-7

Bourne, P. E., Polka, J. K., Vale, R. D., & Kiley, R. (2017). Ten simple rules to consider regarding preprint submission. *PLOS Computational Biology*, *13*(5), e1005473. https://doi.org/10.1371/journal.pcbi.1005473

Boyer, P. (1994). *The naturalness of religious ideas: A cognitive theory of religion*. Univ of California Press.

Boyer, P. (2001). *Religion explained: The evolutionary origins of religious thought*. Basic Books.

Boyer, P. (2006). Prosocial aspects of afterlife beliefs: Maybe another by-product. *Behavioral and Brain Sciences*, *29*(5), 466–466.

Boyer, P. (2008). Religion: Bound to believe? *Nature*, *455*(7216), 1038–1039. https://doi.org/10.1038/4551038a

Bradshaw, M., & Ellison, C. G. (2009). The nature-nurture debate is over, and both sides lost! Implications for understanding gender differences in religiosity. *Journal for the scientific study of religion*, *48*, 241–251. https://doi.org/10.1111/j.1468-5906.2009.01443.x

Brandt, M. J., & Crawford, J. T. (2020). Worldview conflict and prejudice. In B. Gawronski (Ed.), *Advances in Experimental Social Psychology* (pp. 1–66). Academic Press. https://doi.org/10.1016/bs.aesp.2019.09.002

Brefczynski-Lewis, J. A., Lutz, A., Schaefer, H. S., Levinson, D. B., & Davidson, R. J. (2007). Neural correlates of attentional expertise in long-term meditation practitioners. *Proceedings of the National Academy of Sciences*, *104*, 11483–11488. https://doi.org/10.1073/pnas.0606552104

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Brinol, P., & Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology*, *20*(1), 49–96. https://doi.org/10.1080/10463280802643640

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, *8*(1), 108–117. https://doi.org/10.1016/j.jarmac.2018.09.005

Brown, D., & James, P. (2019). Religious characteristics of states dataset project - demographics v. 2.0 (RCS-dem 2.0), countries only. https://doi.org/10.17605/OSF.IO/7SR4M

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, *90*(1), 125–144. https://doi.org/10.1348/000712699161189

Bulbulia, J. A. (2008). Religious costs as adaptations that signal altruistic intention. *10*(1), 19–38.

Bulbulia, J. A. (2013). The arts transform the cognitive science of religion. *Journal for the Cognitive Science of Religion*, *1*, 141. https://doi.org/10.1558/jcsr.v1i2.141

Bulbulia, J. A. (2022). Causal models are needed to infer how religion affects mental health. *Commentary in MARP special issue.*

Bulbulia, J. A., Schjoedt, U., Shaver, J. H., Sosis, R., & Wildman, W. J. (2021). Causal inference in regression: Advice to authors. *Religion, Brain & Behavior*, *11*(4), 353–360. https://doi.org/10.1080/2153599X.2021.2001259

Bulbulia, J. A., & Slingerland, E. (2012). Religious studies as a life science. *Numen-international Review for The History of Religions*, *59*, 564–613. https://doi.org/10.1163/15685276-12341240

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, *4*(6), 215–222. https://doi.org/10.1016/S1364-6613(00)01483-2

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*(5), 365–376. https://doi.org/10.1038/nrn3475

Cacciatore, M. A., Browning, N., Scheufele, D. A., Brossard, D., Xenos, M. A., & Corley, E. A. (2018). Opposing ends of the spectrum: Exploring trust in scientific and religious authorities. *Public Understanding of Science*, *27*(1), 11–28. https://doi.org/10.1177/0963662516661090

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., … Wu, H. (2018). Evaluating replicability of social science experiments in *Nature* and *Science*. *Nature Human Behaviour*, *2*, 637–644.

Camerer, C. F., Dreber, A., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., & Razen, M. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

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*(11), 1938–1951. https://doi.org/10.1002/jclp.22681

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*. https://doi.org/10.18637/jss.v076.i01

Carpenter, B. (2018). We were measuring the speed of Stan incorrectly—it's faster than we thought in some cases due to antithetical sampling.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, *280*(5364), 747–749. https://doi.org/10.1126/science.280.5364.747

Center for Open Science. (2021). Open science framework.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, *9*, 40–48. https://doi.org/10.1177/1745691613513470

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology*, *66*(3), 460–473. https://doi.org/10.1037/0022-3514.66.3.460

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*, 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Chambers, C. D. (2017). *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*. Princeton University Press. https://doi.org/10.1515/9780691192031

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, *1*, 4–17. https://doi.org/10.3934/neuroscience.2014.1.4

Chang, M.-C., Chen, P.-F., Lee, T.-H., Lin, C.-C., Chiang, K.-T., Tsai, M.-F., Kuo, H.-F., & Lung, F.-W. (2021). The effect of religion on psychological resilience in healthcare workers during the coronavirus disease 2019 pandemic. *Frontiers in Psychology*, *12*, 628894. https://doi.org/10.3389/fpsyg.2021.628894

Charles, S. J., Bartlett, J. E., Messick, K. J., III, T. J. C., & Uzdavines, A. (2019). Researcher degrees of freedom in the psychology of religion. *The International Journal for the Psychology of Religion*, *29*(4), 230–245. https://doi.org/10.1080/10508619.2019.1660573

Chartier, C., Kline, M., McCarthy, R., Nuijten, M., Dunleavy, D. J., & Ledgerwood, A. (2018). The cooperative revolution is making psychological science better. *APS Observer*, *31*(10).

Chatard, A., Hirschberger, G., & Pyszczynski, T. (2020). A word of caution about Many Labs 4: If you fail to follow your preregistered plan, you may fail to find a real effect. https://doi.org/10.31234/osf.io/ejubn

Chen, G., Cox, R. W., Glen, D. R., Rajendra, J. K., Reynolds, R. C., & Taylor, P. A. (2018). A tail of two sides: Artificially doubled false positive rates in neuroimaging due to the sidedness choice with t-tests. *Human Brain Mapping*, 1037–1043. https://doi.org/10.1101/328567

Chen, S.-C., Szabelska, A., Chartier, C. R., Kekecs, Z., Lynott, D., Bernabeu, P., Jones, B. C., DeBruine, L., Levitan, C., Werner, K. M., Wang, K., Milyavskaya, M., Musser, E. D., Papadatou-Pastou, M., Coles, N. A., Janssen, S., Ozdogru, A., Storage, D., Manley, H., … Schmidt, K. (2018). Investigating object orientation effects across 14 languages. https://doi.org/10.31234/osf.io/t2pjv

Chia, E. K. F., & Jih, C.-S. (1994). The effects of stereotyping on impression formation: Cross-cultural perspectives on viewing religious persons. *The Journal of Psychology*, *128*(5), 559–565. https://doi.org/10.1080/00223980.1994.9914913

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*(4), 482–493. https://doi.org/10.1177/1745691610375562

Chudek, M., Heller, S., Birch, S., & Henrich, J. (2012). Prestige-biased cultural learning: Bystander's differential attention to potential models influences children's learning. *Evolution and Human Behavior*, *33*(1), 46–56. https://doi.org/10.1016/j.evolhumbehav.2011.05.005

Chudek, M., McNamara, R., Birch, S., Bloom, P., & Henrich, J. (2013). Developmental and cross-cultural evidence for intuitive dualism. *Psychological Science*, *20*.

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.

Clark, C. J., Winegard, B. M., Beardslee, J., Baumeister, R. F., & Shariff, A. F. (2020). RETRACTED: Declines in religiosity predict increases in violent crime—but not among countries with relatively high average IQ. *Psychological Science*, *31*(2), 170–183. https://doi.org/10.1177/0956797619897915

Clément, F., Koenig, M., & Harris, P. (2004). The ontogenesis of trust. *Mind & Language*, *19*(4), 360–379. https://doi.org/10.1111/j.0268-1064.2004.00263.x

Cohen, E., & Barrett, J. (2008). When minds migrate: Conceptualizing spirit possession. *Journal of Cognition and Culture*, *8*(1-2), 23–48. https://doi.org/10.1163/156770908X289198

Cohen, E., Burdett, E., Knight, N., & Barrett, J. (2011). Cross-cultural similarities and differences in person-body reasoning: Experimental evidence from the United Kingdom and Brazilian Amazon. *Cognitive Science*, *35*(7), 1282–1304. https://doi.org/10.1111/j.1551-6709.2011.01172.x

Collett, J. L., & Childs, E. (2011). Minding the gap: Meaning, affect, and the potential shortcomings of vignettes. *Social Science Research*, *40*(2), 513–522. https://doi.org/10.1016/j.ssresearch.2010.08.008

Collett, J. L., & Lizardo, O. (2009). A power-control theory of gender and religiosity. *Journal for the Scientific Study of Religion*, *48*(2), 213–231. https://doi.org/10.1111/j.1468-5906.2009.01441.x

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science*, *8*(1), 160–179. https://doi.org/10.1111/tops.12186

Corner, A., & Hahn, U. (2009). Evaluating science arguments: Evidence, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied*, *15*(3), 199–212. https://doi.org/10.1037/a0016533

Corriveau, K. H., Chen, E. E., & Harris, P. L. (2015). Judgments about fact and fiction by children from religious and nonreligious backgrounds. *Cognitive Science*, *39*(2), 353–382. https://doi.org/10.1111/cogs.12138

Costa, P. T., & McCrae, R. R. (1992). *NEO personality inventory–Revised (NEO PI-R) and NEO five-factor inventory (NEO-FFI): Professional manual.* Psychological Assessment Resources.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., … Zhou, X. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-018-0400-9

Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., & Mayberg, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, *33*, 1914–1928. https://doi.org/10.1002/hbm.21333

Cremers, H. R., Wager, T. D., & Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *PloS ONE*, *12*, e0184923. https://doi.org/10.1371/journal.pone.0184923

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test-retest reliability in infant speech perception tasks. *Infancy*, *21*(5), 648–667. https://doi.org/10.1111/infa.12127

Crüwell, S., van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., Orben, A., Parsons, S., & Schulte-Mecklenbeck, M. (2019). Seven easy steps to open science. *Zeitschrift für Psychologie*, *227*(4), 237–248. https://doi.org/10.1027/2151-2604/a000387

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. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Cutright, K. M. (2011). The beauty of boundaries: When and why we seek structure in consumption. *Journal of Consumer Research*, *38*, 775–790. https://doi.org/10.1086/661563

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*, *9*, 179–194.

Daws, R. E., & Hampshire, A. (2017). The negative relationship between reasoning and religiosity is underpinned by a bias for intuitive responses specifically when intuition and logic are in conflict. *Frontiers in Psychology*, *8*, 2191–2191. https://doi.org/10.3389/fpsyg.2017.02191

de Rijcke, S., & Penders, B. (2018). Resist calls for replicability in the humanities. *Nature*, *560*, 29. https://doi.org/10.1038/d41586-018-05845-z

de Vrieze, J. (2018). The metawars. *Science*, *361*(6408), 1184–1188. https://doi.org/10.1126/science.361.6408.1184

Decety, J., Cowell, J. M., Lee, K., Mahasneh, R., Malcolm-Smith, S., Selcuk, B., & Zhou, X. (2015). RETRACTED: The negative association between religiousness and children's altruism across the world. *Current Biology*, *25*(22), 2951–2955. https://doi.org/10.1016/j.cub.2015.09.056

Dechesne, M., Pyszczynski, T., Arndt, J., Ransom, S., Sheldon, K. M., van Knippenberg, A., & Janssen, J. (2003). Literal and symbolic immortality: The effect of evidence of literal immortality on self-esteem striving in response to mortality salience. *Journal of Personality and Social Psychology*, *84*(4), 722–737. https://doi.org/10.1037/0022-3514.84.4.722

DeHaven, A. C. (2017). Preregistration: A plan, not a prison.

Dell'Acqua, R., Job, R., Peressotti, F., & Pascali, A. (2007). The picture-word interference effect is not a Stroop effect. *Psychonomic Bulletin & Review*, *14*(4), 717–722. https://doi.org/10.3758/bf03196827

DellaVigna, S., & Vivalt, E. (2019). Social Science Prediction Platform.

Dennett, D. C. (2006). *Breaking the spell: Religion as a natural phenomenon*. Penguin.

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*(6), 1278–1290. https://doi.org/10.1037/a0024402

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

Dovidio, J. F. (2016). Commentary: A big problem requires a foundational change. *Journal of Experimental Social Psychology*, *66*, 159–165. https://doi.org/10.1016/j.jesp.2016.01.008

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081. https://doi.org/10.1371/journal.pone.0029081

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & 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*(50), 15343–15347. https://doi.org/10.1073/pnas.1516179112

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*(2), 195–224. https://doi.org/10.1007/s11013-007-9046-2

Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI. *Trends in Cognitive Sciences*, *20*, 425–443. https://doi.org/10.1016/j.tics.2016.03.014

Dutilh, G., Vandekerckhove, J., Ly, A., Matzke, D., Pedroni, A., Frey, R., Rieskamp, J., & Wagenmakers, E.-J. (2017). A test of the diffusion model explanation for the worst performance rule using preregistration and blinding. *Attention, Perception, & Psychophysics*, *79*, 713–725. https://doi.org/10.3758/s13414-017-1304-y

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A.-M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y.-S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., van Maanen, L., … Donkin, C. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, *26*(4), 1051–1069. https://doi.org/10.3758/s13423-017-1417-2

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

Duyvendak, J. W. (2019). On Slow Science.

Ebersole, C., Atherton, O., Belanger, A., Skulborstad, H., Allen, J., Banks, J., Baranski, E., Bernstein, M., Bonfiglio, D., Boucher, L., Brown, E., Budiman, N., Cairo, A., Capaldi, C., Chartier, C., Chung, J., Cicero, D., Coleman, J., Conway, J., … Nosek, B. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82.

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), 1–6. https://doi.org/10.1037/pspa0000187

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. *Commentary in MARP special issue.*

Egner, T., Ely, S., & Grinband, J. (2010). Going, going, gone: Characterizing the time-course of congruency sequence effects. *Frontiers in Psychology*, *1*, 154. https://doi.org/10.3389/fpsyg.2010.00154

Egner, T., Etkin, A., Gale, S., & Hirsch, J. (2008). Dissociable neural systems resolve conflict from emotional versus nonemotional distracters. *Cerebral Cortex*, *18*(6), 1475–1484. https://doi.org/10.1093/cercor/bhm179

Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of Task–Relevant information. *Nature Neuroscience*, *8*(12), 1784–1790. https://doi.org/10.1038/nn1594

Emmons, N. A., & Kelemen, D. (2014). The development of children's prelife reasoning: Evidence from two cultures. *Child Development*, *85*(4), 1617–1633. https://doi.org/10.1111/cdev.12220

Eriksson, K. (2012). The nonsense math effect. *Judgment and Decision Making*, *7*(6), 746–749.

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS ONE*, *12*(9), e0184661. https://doi.org/10.1371/journal.pone.0184661

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Sitek, K. R., Gomez, D. E. P., Lurie, D. J., Ye, Z., Poldrack, R. A., & Gorgolewski, K. J. (2018). Poldracklab/fmriprep: 1.0.15. https://doi.org/10.5281/zenodo.1248927

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., et al. (2019). FMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*, 111. https://doi.org/10.1038/s41592-018-0235-4

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*, 219–234.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34.

Etz, A., Bartlema, A., Vanpaemel, W., Wagenmakers, E.-J., & Morey, R. D. (2019). An exploratory survey of student and researcher intuitions about statistical evidence.

Evans, A. M., Sleegers, W., & Mlakar, Ž. (2020). Individual differences in receptivity to scientific bullshit. *Judgment and Decision Making*, *15*, 401–412. https://doi.org/10.31234/osf.io/2r65q

Evans, E. M. (2001). Cognitive and contextual factors in the emergence of diverse belief systems: Creation versus evolution. *Cognitive Psychology*, 217–266. https://doi.org/10.1006/cogp.2001.0749

Eylon, D., & Allison, S. T. (2005). The "frozen in time" effect in evaluations of the dead. *Personality and Social Psychology Bulletin*, *31*(12), 1708–1717. https://doi.org/10.1177/0146167205277806

Farias, M., Newheiser, A.-K., Kahane, G., & de Toledo, Z. (2013). Scientific faith: Belief in science increases in the face of stress and existential anxiety. *Journal of Experimental Social Psychology*, *49*, 1210–1213. https://doi.org/10.1016/j.jesp.2013.05.008

Farr, C. (2020). Stanford Medical faculty lambaste former colleague and Trump coronavirus advisor Dr. Scott Atlas.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Fayard, J. V., Bassi, A. K., Bernstein, D. M., & Roberts, B. W. (2009). Is cleanliness next to godliness? Dispelling old wives' tales: Failure to replicate Zhong and Liljenquist (2006). *Journal of Articles in Support of the Null Hypothesis*, *6*(2), 22–30.

Fernandez-Duque, D., Evans, J., Christian, C., & Hodges, S. D. (2014). Superfluous neuroscience information makes explanations of psychological phenomena more appealing. *Journal of Cognitive Neuroscience*, *27*(5), 926–944. https://doi.org/10.1162/jocn_a_00750

Feynman, R. P. (1974). Cargo Cult Science.

Fillard, P., Descoteaux, M., Goh, A., Gouttard, S., Jeurissen, B., Malcolm, J., Ramirez-Manzanares, A., Reisert, M., Sakaie, K., Tensaouti, F., Yo, T., Mangin, J.-F., & Poupon, C. (2011). Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. *NeuroImage*, *56*(1), 220–234. https://doi.org/10.1016/j.neuroimage.2011.01.032

First, M. B., Gibbon, M., Spitzer, R. L., & Benjamin, L. S. (1997). *User's guide for the structured clinical interview for DSM-IV axis II personality disorders: SCID-II*. American Psychiatric Pub.

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology*, *10*.

Fiske, S. T., & Dupree, C. (2014). Gaining trust as well as respect in communicating to motivated audiences about science topics. *Proceedings of the National Academy of Sciences*, *111*(Supplement 4), 13593–13597. https://doi.org/10.1073/pnas.1317505111

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

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*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102. https://doi.org/10.1016/S1053-8119(09)70884-5

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., & Dreber, A. (2018). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*. https://doi.org/10.1016/j.joep.2018.10.009

Forstmann, M., & Burgmer, P. (2015). Adults are intuitive mind-body dualists. *Journal of Experimental Psychology: General*, *144*(1), 222–235. https://doi.org/10.1037/xge0000045

Forstmann, M., Burgmer, P., & Mussweiler, T. (2012). "The mind is willing, but the flesh is weak": The effects of mind-body dualism on health behavior. *Psychological Science*, *23*(10), 1239–1245. https://doi.org/10.1177/0956797612442392

&

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151–156. https://doi.org/10.3758/s13423-012-0227-9

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*, 153–169. https://doi.org/10.1016/j.jmp.2013.02.003

Francis, L. J. (1997). The psychology of gender differences in religion: A review of empirical research. *Religion*, *27*, 81–96. https://doi.org/10.1006/reli.1996.0066

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Friesen, J. P., Campbell, T. H., & Kay, A. C. (2015). The psychological advantage of unfalsifiability: The appeal of untestable religious and political ideologies. *Journal of Personality and Social Psychology*, *108*(3), 515–529. https://doi.org/10.1037/pspp0000018

Frith, U. (2020). Fast lane to slow science. *Trends in Cognitive Sciences*, *24*(1), 1–2. https://doi.org/10.1016/j.tics.2019.10.007

Fu, D. Y., & Hughey, J. J. (2019). Releasing a preprint is associated with more attention and citations for the peer-reviewed article (P. Rodgers & O. Amaral, Eds.). *eLife*, *8*, e52646. https://doi.org/10.7554/eLife.52646

Funk, C. (2020). Key findings about Americans' confidence in science and their views on scientists' role in society.

Funk, C., Kennedy, B., & Johnson, C. (2020). Trust in medical scientists has grown in U.S., but mainly among democrats.

Funk, C., Tyson, A., Kennedy, B., & Johnson, C. (2020). Scientists are among the most trusted groups in society, though many value practical experience over expertise.

Gabry, J., Goodrich, B., & Lysy, M. (2020). Rstantools: Tools for developing r packages interfacing with Stan.

Galen, L. W., Smith, C. M., Knapp, N., & Wyngarden, N. (2011). Perceptions of religious and nonreligious targets: Exploring the effects of perceivers' religious fundamentalism. *Journal of Applied Social Psychology*, *41*(9), 2123–2143. https://doi.org/10.1111/j.1559-1816.2011.00810.x

Galinsky, A. D., Magee, J. C., Gruenfeld, D. H., Whitson, J. A., & Liljenquist, K. A. (2008). Power reduces the press of the situation: Implications for creativity, conformity, and dissonance. *Journal of Personality and Social Psychology*, *95*, 1450–1466. https://doi.org/10.1037/a0012633

Gallup. (2016). Gallup Poll Social Series: Values and Beliefs. Retrieved from: https://news.gallup.com/poll/193271/americans-believe-god.aspx.

Gallup. (2019). Wellcome gobal monitor – first wave findings.

Garssen, B., Visser, A., & Pool, G. (2020). Does spirituality or religion positively affect mental health? Meta-analysis of longitudinal studies. *The International Journal for the Psychology of Religion*, *31*(1), 4–20. https://doi.org/10.1080/10508619.2020.1729570

Gasquoine, P. G. (2013). Localization of function in anterior cingulate cortex: From psychosurgery to functional neuroimaging. *Neuroscience & Biobehavioral Reviews*, *37*, 340–348. https://doi.org/10.1016/j.neubiorev.2013.01.002

Gauchat, G. (2011). The cultural authority of science: Public trust and acceptance of organized science. *Public Understanding of Science*, *20*(6), 751–770. https://doi.org/10.1177/0963662510365246

Gauchat, G. (2012). Politicization of science in the public sphere: A study of public trust in the United States, 1974 to 2010. *American Sociological Review*, *77*(2), 167–187. https://doi.org/10.1177/0003122412438225

Gazzaniga, M., & Ivry, R. B. (2013). *Cognitive neuroscience: The biology of the mind: Fourth international student edition*. WW Norton.

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*(3), e18–e39. https://doi.org/10.1037/pspp0000104

Gelfand, A. E., Smith, A. F. M., & Lee, T. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*, 523–532. https://doi.org/10.2307/2290286

&

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis (3rd ed.)* Chapman & Hall/CRC.

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician*, *73*(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, *40*, 530–543. https://doi.org/10.3102/1076998615606113

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.1080/01621459.1993.10476353

George, L. K., Ellison, C. G., & Larson, D. B. (2002). Explaining the relationships between religious involvement and health. *Psychological Inquiry*, *13*(3), 190–200. https://doi.org/10.1207/S15327965PLI1303_04

Gervais, W. M., Shariff, A. F., & Norenzayan, A. (2011). Do you believe in atheists? Distrust is central to anti-atheist prejudice. *Journal of Personality and Social Psychology*, *101*, 1189–1206.

Gervais, W. M., McKee, S. E., & Malik, S. (2020). Do religious primes increase risk taking? Evidence against "anticipating divine protection" in two preregistered direct replications of Kupor, Laurin, and Levav (2015). *Psychological Science*, *31*(7), 858–864. https://doi.org/10.1177/0956797620922477

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, *336*(6080), 493–496. https://doi.org/10.1126/science.1215647

Gervais, W. M., van Elk, M., Xygalatas, D., McKay, R. T., Aveyard, M., Buchtel, E. E., Dar-Nimrod, I., Klocová, E. K., Ramsay, J. E., Riekki, T., et al. (2018). Analytic atheism: A cross-culturally weak and fickle phenomenon? *Judgment and Decision Making*, *13*, 268–274.

Gervais, W. M., Willard, A. K., Norenzayan, A., & Henrich, J. (2011). The cultural transmission of faith: Why innate intuitions are necessary, but insufficient, to explain religious belief. *Religion*, *41*(3), 389–410. https://doi.org/10.1080/0048721X.2011.604510

Gervais, W. M., Xygalatas, D., McKay, R. T., van Elk, M., Buchtel, E. E., Aveyard, M., Schiavone, S. R., Dar-Nimrod, I., Svedholm-Häkkinen, A. M., Riekki, T., Klocová, E. K., Ramsay, J. E., & Bulbulia, J. (2017). Global evidence of extreme intuitive moral prejudice against atheists. *Nature Human Behaviour*, *1*(8), 0151. https://doi.org/10.1038/s41562-017-0151

Gewin, V. (2016). Data sharing: An open mind on open data. *Nature*, *529*(7584), 117–119. https://doi.org/10.1038/nj7584-117a

Geyer, C. J. (2011). Introduction to Markov Chain Monte Carlo. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. CRC Press. https://doi.org/10.1201/b10905-3

&

Gilder, T. S. E., & Heerey, E. A. (2018). The role of experimenter belief in social priming. *Psychological Science*, *29*(3), 403–417. https://doi.org/10.1177/0956797617737128

Gilmore, R. O., Kennedy, J. L., & Adolph, K. E. (2018). Practical solutions for sharing data and materials from psychological research. *Advances in Methods and Practices in Psychological Science*, *1*, 121–130. https://doi.org/10.1177/2515245917746500

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. https://doi.org/10.1177/1745691612457576

Gligorić, V., & Vilotijević, A. (2020). "Who said it?" How contextual information influences perceived profundity of meaningful quotes and pseudo-profound bullshit. *Applied Cognitive Psychology*, *34*(2), 535–542. https://doi.org/10.1002/acp.3626

Glover, G. H. (1999). Deconvolution of impulse response in Event–Related BOLD fMRI. *NeuroImage*, *9*(4), 416–429. https://doi.org/10.1006/nimg.1998.0419

Gneezy, U., Keenan, E. A., & Gneezy, A. (2014). Avoiding overhead aversion in charity. *Science*, *346*, 632–635. https://doi.org/10.1126/science.1253932

Gomes, C. M., & McCullough, M. E. (2015). The effects of implicit religious primes on dictator game allocations: A preregistered replication experiment. *Journal of Experimental Psychology: General*, *144*(6), e94–e104. https://doi.org/10.1037/xge0000027

Good, M., Inzlicht, M., & Larson, M. J. (2015). God will forgive: Reflecting on God's love decreases neurophysiological responses to errors. *Social Cognitive and Affective Neuroscience*, *10*(3), 357–363. https://doi.org/10.1093/scan/nsu096

Goode, C., Keefer, L. A., & Molina, L. E. (2014). A compensatory control account of meritocracy. *Journal of Social and Political Psychology*, *2*(1), 313–334. https://doi.org/10.5964/jspp.v2i1.372

Gorgolewski, K. J., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13. https://doi.org/10.3389/fninf.2011.00013

Gorgolewski, K. J., Esteban, O., Ellis, D. G., Notter, M. P., Ziegler, E., Johnson, H., Hamalainen, C., Yvernault, B., Burns, C., Manhães-Savio, A., Jarecka, D., Markiewicz, C. J., Salo, T., Clark, D., Waskom, M., Wong, J., Modat, M., Dewey, B. E., Clark, M. G., … Ghosh, S. (2017). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python. 0.13.1. https://doi.org/10.5281/zenodo.581704

Göritz, A. S., & Moser, K. (2006). Web-based mood induction. *Cognition and Emotion*, *20*, 887–896.

Gorsuch, R. L., & Smith, C. S. (1983). Attributions of responsibility to God: An interaction of religious beliefs and outcomes. *Journal for the Scientific Study of Religion*, *22*, 340–352. https://doi.org/10.2307/1385772

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*, 93–104. https://doi.org/10.1037/0003-066x.59.2.93

Gould, D. (1996). Using vignettes to collect data for nursing research studies: How valid are the findings? *Journal of Clinical Nursing*, *5*(4), 207–212. https://doi.org/10.1111/j.1365-2702.1996.tb00253.x

Grafman, J., Cristofori, I., Zhong, W., & Bulbulia, J. (2020). The neural basis of religious cognition. *Current Directions in Psychological Science*, *29*(2), 126–133. https://doi.org/10.1177/0963721419898183

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science (New York, N.Y.)*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Gray, K., Anne Knickman, T., & Wegner, D. M. (2011). More dead than dead: Perceptions of persons in the persistent vegetative state. *Cognition*, *121*(2), 275–280. https://doi.org/10.1016/j.cognition.2011.06.014

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Wiley.

Greenaway, K. H., Louis, W. R., & Hornsey, M. J. (2013). Loss of control increases belief in precognition and belief in precognition increases control. *PLoS ONE*, *8*(8), e71327–e71327. https://doi.org/10.1371/journal.pone.0071327

Greenberg, J., Porteus, J., Simon, L., Pyszczynski, T., & Solomon, S. (1995). Evidence of a Terror Management function of cultural icons: The effects of mortality salience on the inappropriate use of cherished cultural symbols. *Personality and Social Psychology Bulletin*, *21*(11), 1221–1228. https://doi.org/10.1177/01461672952111010

Greenberg, J., Pyszczynski, T., Solomon, S., Simon, L., & Breus, M. (1994). Role of consciousness and accessibility of death-related thoughts in mortality salience effects. *Journal of Personality and Social Psychology*, *67*(4), 627–637. https://doi.org/10.1037/0022-3514.67.4.627

Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the effects of removing "Too fast" responses and respondents from web surveys. *Public Opinion Quarterly*, *79*(2), 471–503. https://doi.org/10.1093/poq/nfu058

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, *48*, 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Grim, B. (2008). Religion in China on the eve of the 2008 Beijing Olympics.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). Bridgesampling: An R Package for Estimating Normalizing Constants. *Journal of Statistical Software*, *92*.

Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*, 123–138. https://doi.org/10.31222/osf.io/heamz

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2021). A primer on Bayesian model-averaged meta-analysis. *Advances in Methods and Practices in Psychological Science*, *4*(3), 25152459211031256. https://doi.org/10.1177/25152459211031256

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2019). Informed Bayesian t-tests. *The American Statistician*, 1–13. https://doi.org/10.1080/00031305.2018.1562983

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. https://doi.org/10.31222/osf.io/m8ujg

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). Bridgesampling: Bridge sampling for marginal likelihoods and Bayes factors.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, *15*(5), 1243–1255. https://doi.org/10.1177/1745691620921521

Gruber, D., & Dickerson, J. A. (2012). Persuasive images in popular science: Testing judgments of scientific reasoning and credibility. *Public Understanding of Science*, *21*, 938–948. https://doi.org/10.1177/0963662512454072

Gut, A., Lambert, A., Gorbaniuk, O., & Mirski, R. (2021). Folk beliefs about soul and mind: Cross-cultural comparison of folk intuitions about the ontology of the person. *Journal of Cognition and Culture*, *21*(3-4), 346–369. https://doi.org/10.1163/15685373-12340116

Haaf, J. M., Hoogeveen, S., Berkhout, S., Gronau, Q. F., & Wagenmakers, E.-J. (2020). A Bayesian multiverse analysis of Many Labs 4: Quantifying the evidence against mortality salience. https://doi.org/10.31234/osf.io/cb9er

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*(4), 779–798. https://doi.org/10.31234/osf.io/ktjnq

Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, *26*(3), 772–789. https://doi.org/10.3758/s13423-018-1522-x

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*(1), 43–55. https://doi.org/10.1111/1468-5906.t01-1-00160

Hahn, U., Harris, A. J. L., & Corner, A. (2016). Public reception of climate science: Coherence, reliability, and independence. *Topics in Cognitive Science*, *8*, 180–195.

Hahn, U., Harris, A. J. L., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*(4), 337–367. https://doi.org/10.22329/il.v29i4.2903

Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished manuscript, University of Virginia*, 191–221.

Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior and Development*, *21*, 167–179. https://doi.org/10.1016/s0163-6383(98)90001-7

Hajcak, G., Moser, J. S., Yeung, N., & Simons, R. F. (2005). On the ERN and the significance of errors. *Psychophysiology*, *42*(2), 151–160. https://doi.org/10.1111/j.1469-8986.2005.00270.x

Hanel, P. H. P., & Zarzeczna, N. (2022). From multiverse analysis to multiverse operationalisations: 262,143 ways of measuring well-being. *Commentary in MARP special issue.*

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., et al. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*, 180448. https://doi.org/10.31222/osf.io/39cfb

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. https://doi.org/10.1017/S0140525X18000675

Hardy, A. (1981). *The spiritual nature of man. A study of contemporary religious experience.*

Hariri, A. R., Bookheimer, S. Y., & Mazziotta, J. C. (2000). Modulating emotional responses: Effects of a neocortical network on the limbic system. *NeuroReport*, *11*(1), 43. https://doi.org/10.1097/00001756-200001170-00009

Harris, A. J., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The appeal to expert opinion: Quantitative support for a Bayesian network approach. *Cognitive Science*, *40*(6), 1496–1533.

Harris, J. I., Usset, T., Voecks, C., Thuras, P., Currier, J., & Erbes, C. (2018). Spiritually integrated care for PTSD: A randomized controlled trial of “Building Spiritual Strength”. *Psychiatry Research*, *267*, 420–428. https://doi.org/10.1016/j.psychres.2018.06.045

Harris, P. L. (2011a). Conflicting Thoughts about Death. *Human Development*, *54*(3), 160–168. https://doi.org/10.1159/000329133

Harris, P. L. (2011b). Death in Spain, Madagascar, and beyond. *Children’s Understanding of Death: From Biological to Religious Conceptions* (pp. 19–40). Cambridge University Press.

Harris, P. L. (2012). *Trusting what you’re told: How children learn from others*. Harvard University Press.

Harris, P. L., & Giménez, M. (2005). Children’s acceptance of conflicting testimony: The case of death. *Journal of Cognition and Culture*, *5*(1), 143–164. https://doi.org/10.1163/1568537054068606

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology*, *69*(1), 251–273. https://doi.org/10.1146/annurev-psych-122216-011710

Harris, S., Kaplan, J. T., Curiel, A., Bookheimer, S. Y., Iacoboni, M., & Cohen, M. S. (2009). The neural correlates of religious and nonreligious belief. *PLoS ONE*, *4*(10), e7272. https://doi.org/10.1371/journal.pone.0007272

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.

Hayward, R. D., & Elliott, M. (2014). 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*(1), 23–43. https://doi.org/10.1007/s13644-013-0135-0

Heck, D. W., & Gronau, Q. F. (2017). metaBMA: Bayesian model averaging for random- and fixed-effects meta-analysis [R Package].

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology*, *82*(6), 903–918. https://doi.org/10.1037/0022-3514.82.6.903

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.*

Heisig, J. P., & Schaeffer, M. (2019). Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. *European Sociological Review*, *35*(2), 258–279. https://doi.org/10.1093/esr/jcy053

Henrich, J. (2009). The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*, *30*(4), 244–260. https://doi.org/10.1016/j.evolhumbehav.2009.03.005

Henrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter.* Princeton University Press.

Henrich, J. (2020). *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous.* Penguin UK.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, *22*, 165–196. https://doi.org/10.1016/s1090-5138(00)00071-4

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

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. *Commentary in MARP special issue.*

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215. https://doi.org/10.1177/2515245919898657

Hodge, K. M. (2008). Descartes' mistake: How afterlife beliefs challenge the assumption that humans are intuitive Cartesian substance dualists. *Journal of Cognition and Culture*, *8*(3-4), 387–415. https://doi.org/10.1163/156853708X358236

Hodge, K. M. (2011a). On imagining the afterlife. *Journal of Cognition and Culture*, *11*(3-4), 367–389.

Hodge, K. M. (2011b). Why immortality alone will not get me to the afterlife. *Philosophical Psychology*, *24*(3), 395–410. https://doi.org/10.1080/09515089.2011.559620

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Chapman & Hall/CRC. https://doi.org/10.1201/b11158

Holcombe, A. O., Kovacs, M., Aust, F., & Aczel, B. (2020). Documenting contributions to scholarly articles using CRediT and tenzing. *PLoS One*, *15*(12), e0244611.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., Coles, M. G., & Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neuroscience*, *7*, 497. https://doi.org/10.1038/nn1238

Hone, L. S. E., & McCullough, M. E. (2015). Does religious cognition really down-regulate hand grip endurance in men? A failure to replicate. *Evolution and Human Behavior*, *36*(2), 81–85. https://doi.org/10.1016/j.evolhumbehav.2014.08.007

Hood, B., Gjersoe, N. L., & Bloom, P. (2012). Do children think that duplicating the body also duplicates the mind? *Cognition*, *125*(3), 466–474. https://doi.org/10.1016/j.cognition.2012.07.005

Hoogeveen, S., Altay, S., Bendixen, T., Berniūnas, R., Bulbulia, J. A., Cheshin, A., Gentili, C., Georgescu, R., Haaf, J. M., Hagel, K., Kavanagh, C. M., Levy, N., McKay, R., Neely, A., Qiu, L., Rabelo, A., Ramsay, J. E., Ross, R. M., Turpin, H., … van Elk, M. (2022). *Does she still love and feel hungry? Afterlife beliefs, mind-body dualism, and religion across 24 countries.*

Hoogeveen, S., Haaf, J. M., Bulbulia, J. A., Ross, R. M., McKay, R., Altay, S., Bendixen, T., Berniūnas, R., Cheshin, A., Gentili, C., Georgescu, R., Gervais, W. M., Hagel, K., Kavanagh, C. M., Levy, N., Neely, A., Qiu, L., Rabelo, A., Ramsay, J. E., … van Elk, M. (2022). The Einstein effect provides global evidence for scientific source credibility effects and the influence of religiosity. *Nature Human Behaviour*, *6*(4), 523–535. https://doi.org/10.1038/s41562-021-01273-8

Hoogeveen, S., Sarafoglou, A., Aczel, B., Aditya, Y., Alayan, A., Allen, P., Altay, S., Alzahawi, S., Amir, Y., Anthony, F.-V., Appiah, O., Atkinson, Q. D., Baimel, A., Balkaya-Ince, M., Balsamo, M., Banker, S., Bartoš, F., Becerra, M., Beffara, B., … Wagenmakers, E.-J. (2022). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior*. https://doi.org/10.31234/osf.io/pbfye

Hoogeveen, S., Sarafoglou, A., van Elk, M., & Wagenmakers, E.-J. (2022). Many-Analysts Religion Project: Reflection and conclusion. *Religion, Brain & Behavior*.

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. https://doi.org/10.1177/2515245920919667

Hoogeveen, S., Schjoedt, U., & van Elk, M. (2018). Did I do that? Expectancy effects of brain stimulation on error-related negativity and sense of agency. *Journal of Cognitive Neuroscience*, 1–14. https://doi.org/10.1162/jocn_a_01297

Hoogeveen, S., Snoek, L., & van Elk, M. (2020). Religious belief and cognitive conflict sensitivity: A preregistered fMRI study. *Cortex*, *129*, 247–265. https://doi.org/10.1016/j.cortex.2020.04.011

Hoogeveen, S., & van Elk, M. (2021). Advancing the cognitive science of religion through replication and open science. *Journal for the Cognitive Science of Religion*, *6*(1-2), 158–190. https://doi.org/10.1558/jcsr.39039

Hoogeveen, S., Wagenmakers, E.-J., Kay, A. C., & van Elk, M. (2018). Compensatory control and religious beliefs: A registered replication report across two countries. *Comprehensive Results in Social Psychology*, *3*(3), 240–265. https://doi.org/10.1080/23743603.2019.1684821

Hove, M. J., Stelzer, J., Nierhaus, T., Thiel, S. D., Gundlach, C., Margulies, D. S., Van Dijk, K. R., Turner, R., Keller, P. E., & Merker, B. (2015). Brain network reconfiguration and perceptual decoupling during an absorptive state of consciousness. *Cerebral Cortex*, *26*, 3116–3124. https://doi.org/10.1093/cercor/bhv137

Howell, E. L., Wirz, C. D., Scheufele, D. A., Brossard, D., & Xenos, M. A. (2020). Deference and decision-making in science and society: How deference to scientific authority goes beyond confidence in science and scientists to become authoritarianism. *Public Understanding of Science*, *29*(8), 800–818. https://doi.org/10.1177/0963662520962741

Hox, J. J. C. M., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods*, *6*, 87–93. https://doi.org/10.18148/srm/2012.v6i2.5033

Hu, A. (2016). Ancestor worship in contemporary China: An empirical investigation. *China Review*, *16*(1), 169–186.

Huang, J., Cheng, L., & Zhu, J. (2013). Intuitive conceptions of dead persons' mentality: A cross-cultural replication and more. *International Journal for the Psychology of Religion*, *23*(1), 29–41. https://doi.org/10.1080/10508619.2013.735493

Huber, D. E., Potter, K. W., & Huszar, L. D. (2019). Less "Story" and more "Reliability" in cognitive neuroscience. *Cortex; a journal devoted to the study of the nervous system and behavior*, *113*, 347. https://doi.org/10.1016/j.cortex.2018.10.030

Huijts, T., & Kraaykamp, G. (2011). Religious involvement, religious context, and self-assessed health in Europe. *Journal of Health and Social Behavior*, *52*(1), 91–106. https://doi.org/10.1177/0022146510394950

Huntenburg, J. M. (2014). *Evaluating nonlinear coregistration of BOLD EPI and T1w images* (Doctoral dissertation). Freie Universität Berlin.

Huntenburg, J. M., Bazin, P.-L., & Margulies, D. S. (2018). Large-scale gradients in human cortical organization. *Trends in Cognitive Sciences*, *22*, 21–31. https://doi.org/10.1016/j.tics.2017.11.002

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, *59*(3), 944–960. https://doi.org/10.1111/ecin.12992

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184. https://doi.org/10.1177/2515245919882903

Ibrahim, J. G., & Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, *15*(1), 46–60.

Inglehart, R. (2006). Mapping global values. *Comparative Sociology*, *5*, 115–136. https://doi.org/10.1163/156913306778667401

Inzlicht, M., & Tullett, A. M. (2010). Reflecting on God: Religious primes can reduce neurophysiological response to errors. *Psychological Science*, *21*(8), 1184–1190. https://doi.org/10.1177/0956797610375451

Inzlicht, M., McGregor, I., Hirsh, J. B., & Nash, K. (2009). Neural markers of religious conviction. *Psychological Science*, *20*(3), 385–392. https://doi.org/10.1111/j.1467-9280.2009.02305.x

Inzlicht, M., Tullett, A. M., & Good, M. (2011). The need to believe: A neuroscience account of religion as a motivated process. *Religion, Brain & Behavior*, *1*(3), 192–212. https://doi.org/10.1080/2153599x.2011.647849

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701. https://doi.org/10.1080/09332480.2005.10722754

Irons, W. (2001). Religion as a hard-to-fake sign of commitment. In R. Nesse (Ed.), *Evolution and the capacity for commitment* (pp. 292–309). Russell Sage Foundation.

Irons, W. (2008). Why people believe (what other people see as) crazy ideas. In J. A. Bulbulia, R. Sosis, R. Genet, C. Genet, & K. Wyman (Eds.), *The Evolution of Religion: Studies, Theories, and Critiques* (pp. 51–57). Collins Foundation Press.

Islam, C.-G., & Lorenz, J. (2022). How to increase the robustness of survey studies. *Commentary in MARP special issue.*

Jallais, C., & Gilet, A.-L. (2010). Inducing changes in arousal and valence: Comparison of two mood induction procedures. *Behavior Research Methods*, *42*, 318–325. https://doi.org/10.3758/brm.42.1.318

JASP Team. (2019). JASP (version 0.11.1)[Computer software].

Jaswal, V. K., & Neely, L. A. (2006). Adults don't always know best: Preschoolers use past reliability over age when learning new words. *Psychological Science*, *17*(9), 757–758. https://doi.org/10.1111/j.1467-9280.2006.01778.x

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, *31*, 203–222. https://doi.org/10.1017/s030500410001330x

Jeffreys, H. (1939). *Theory of Probability* (First). Oxford University Press.

Jeffreys, H. (1961). *Theory of Probability* (Third). Oxford University Press. https://doi.org/10.2307/2530899

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*, 825–841. https://doi.org/10.1006/nimg.2002.1132

Jern, A., Chang, K.-m. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224. https://doi.org/10.1037/a0035941

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for Truth–Telling. *Psychological Science*, *23*, 524–532. https://doi.org/10.1037/e632032012-001

Johnson, C. N. (1990). If you had my brain, where would I be? Children's understanding of the brain and identity. *Child Development*, *61*(4), 962–972.

Johnson, D. D. (2020). *Strategic instincts: The adaptive advantages of cognitive biases in international politics* (Vol. 172). Princeton University Press.

Johnson, D. D., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, *477*(7364), 317–320.

Johnson, K. A., Okun, M. A., Cohen, A. B., Sharp, C. A., & Hook, J. N. (2019). Development and validation of the five-factor LAMBI measure of God representations. *Psychology of Religion and Spirituality*, *11*(4), 339–349. https://doi.org/10.1037/rel0000207

Johnson, M. K., Rowatt, W. C., Barnard-Brak, L. M., Patock-Peckham, J. A., LaBouff, J. P., & Carlisle, R. D. (2011). A mediational analysis of the role of right-wing authoritarianism and religious fundamentalism in the religiosity–prejudice link. *Personality and Individual Differences*, *50*(6), 851–856. https://doi.org/10.1016/j.paid.2011.01.010

Johnston, A. M., Mills, C. M., & Landrum, A. R. (2015). How do children weigh competence and benevolence when deciding whom to trust? *Cognition*, *144*, 76–90. https://doi.org/10.1016/j.cognition.2015.07.015

Jonas, E., & Fischer, P. (2006). Terror management and religion: Evidence that intrinsic religiousness mitigates worldview defense following mortality salience. *Journal of Personality and Social Psychology*, *91*(3), 553–567. https://doi.org/10.1037/0022-3514.91.3.553

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxsom, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, C., … Coles, N. A. (2021). To which world regions does the valence–dominance model of social perception apply? *Nature Human Behaviour*, *5*(1), 159–169. https://doi.org/10.1038/s41562-020-01007-2

Jong, J. (2021). Death anxiety and religion. *Current Opinion in Psychology*, *40*, 40–44. https://doi.org/10.1016/j.copsyc.2020.08.004

Jong, J., Halberstadt, J., & Bluemke, M. (2012). Foxhole atheism, revisited: The effects of mortality salience on explicit and implicit religious belief. *Journal of Experimental Social Psychology*, *48*(5), 983–989. https://doi.org/10.1016/j.jesp.2012.03.005

Kahan, D. M., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research*, *14*(2), 147–174. https://doi.org/10.1080/13669877.2010.511246

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*, 773–795. https://doi.org/10.2307/2291091

Kavanagh, C. M., Jong, J., McKay, R., & Whitehouse, H. (2018). Positive experiences of high arousal martial arts rituals are linked to identity fusion and costly pro-group actions. *European Journal of Social Psychology*, *0*(0). https://doi.org/10.1002/ejsp.2514

Kavanagh, C. M., & Kapitany, R. (2017). Promoting the benefits and clarifying misconceptions about preregistration, preprints, and open science for cognitive science of religion. *Journal for the Cognitive Science of Religion*, *5*, 461–481.

Kay, A. C., Gaucher, D., McGregor, I., & Nash, K. (2010). Religious belief as compensatory control. *Personality and Social Psychology Review*, *14*(1), 37–48.

Kay, A. C., Gaucher, D., Napier, J. L., Callan, M. J., & Laurin, K. (2008). God and the government: Testing a compensatory control mechanism for the support of external systems. *Journal of Personality and Social Psychology*, *95*(1), 18–35. https://doi.org/10.1037/0022-3514.95.1.18

Kay, A. C., Moscovitch, D. A., & Laurin, K. (2010). Randomness, attributions of arousal, and belief in God. *Psychological Science*, *21*(2), 216–218. https://doi.org/10.1177/0956797609357750

Kay, A. C., Shepherd, S., Blatz, C. W., Chua, S. N., & Galinsky, A. D. (2010). For God (or) country: The hydraulic relation between government instability and belief in religious sources of control. *Journal of Personality and Social Psychology*, *99*, 725–739. https://doi.org/10.1037/a0021140

Kay, A. C., Whitson, J. A., Gaucher, D., & Galinsky, A. D. (2009). Compensatory control: Achieving order through the mind, our institutions, and the heavens. *Current Directions in Psychological Science*, *18*, 264–268. https://doi.org/10.1111/j.1467-8721.2009.01649.x

Kelemen, D. (1999). Why are rocks pointy? Children's preference for teleological explanations of th natural world. *Developmental psychology*, *35*(6), 1440.

Kelemen, D. (2004). Are children "intuitive theists"? Reasoning about purpose and design in nature. *Psychological science*, *15*(5), 295–301.

Kelley, J., & de Graaf, N. D. (1997). National context, parental socialization, and religious belief: Results from 15 nations. *American Sociological Review*, *62*(4), 639–659. https://doi.org/10.2307/2657431

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science*, *342*, 377–380. https://doi.org/10.1126/science.1239918

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L., Kennett, C., Slowik, A., Sonnleitner, C., Hess–Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low cost, effective method for increasing transparency. *PLoS Biology*, *14*, e1002456. https://doi.org/10.1371/journal.pbio.1002456

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Neto, E. C., et al. (2017). Mindboggling morphometry of human brains. *PLoS Computational Biology*, *13*, e1005350. https://doi.org/10.1371/journal.pcbi.1005350

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Hofelich Mohr, A., IJzerman, H., Nilsonne, G., Vanpaemel, W., & Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*(1), 1–15. https://doi.org/10.1525/collabra.158

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J., Cromar, R., Vidamuerte, D., Gardiner, G., Gosnell, C., Grahe, J. E., Hall, C., Joy-Gaba, J. A., Legg, A. M., Levitan, C., ... Ratliff, K. A. (2019). Many Labs 4:

&

Failure to replicate mortality salience effect with and without original author involvement. https://doi.org/10.31234/osf.io/vef2c

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahńik, S., Bernstein, M., Bocian, K., Brandt, M., Brooks, B., Brumbaugh, C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*, 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., et al. (2014). Theory building through replication: Response to commentaries on the "Many Labs" replication project. *Social Psychology*, *45*(4), 299–311.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J., Bahník, S., Batra, R., Berkics, M., Bernstein, M., Berry, D., Bialobrzeska, O., Binan, E., Bocian, K., Brandt, M., Busching, R., Redei, A., … Nosek, B. (2018). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69. https://doi.org/10.1111/j.1467-9574.2005.00279.x

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). Springer Verlag.

Koenig, H. G. (2009). Research on religion, spirituality, and mental health: A review. *The Canadian Journal of Psychiatry*, *54*(5), 283–291. https://doi.org/10.1177/070674370905400502

Koenig, H. G., Al-Zaben, F., & VanderWeele, T. J. (2020). Religion and psychiatry: Recent developments in research. *BJPsych Advances*, *26*(5), 262–272. https://doi.org/10.1192/bja.2019.81

Koenig, H. G., Hill, T. D., Pirutinsky, S., & Rosmarin, D. H. (2021). Commentary on "Does spirituality or religion positively affect mental health?" *The International Journal for the Psychology of Religion*, *31*(1), 27–44. https://doi.org/10.1080/10508619.2020.1766868

Koenig, H. G., & Larson, D. B. (2001). Religion and mental health: Evidence for an association. *International Review of Psychiatry*, *13*(2), 67–78. https://doi.org/10.1080/09540260124661

Konvalinka, I., Xygalatas, D., Bulbulia, J., Schjødt, U., Jegindø, E.-M., Wallot, S., Van Orden, G., & Roepstorff, A. (2011). Synchronized arousal between performers and related spectators in a fire-walking ritual. *Proceedings of the National Academy of Sciences*, *108*, 8514–8519. https://doi.org/10.1073/pnas.1016955108

Kossowska, M., Szwed, P., Wronka, E., Czarnek, G., & Wyczesany, M. (2016). Anxiolytic function of fundamentalist beliefs: Neurocognitive evidence. *Personality and Individual Differences*, *101*, 390–395. https://doi.org/10.1016/J.PAID.2016.06.039

Krause, N. M., Brossard, D., Scheufele, D. A., Xenos, M. A., & Franke, K. (2019). Trends—Americans' trust in science and scientists. *Public Opinion Quarterly*, *83*(4), 817–836. https://doi.org/10.1093/poq/nfz041

Kregting, J., Scheepers, P., Vermeer, P., & Hermans, C. (2018). Why God has left the Netherlands: Explanations for the decline of institutional Christianity in the Netherlands between 1966 and 2015. *Journal for the Scientific Study of Religion*, *57*(1), 58–79. https://doi.org/10.1111/jssr.12499

Kreps, S. E., & Kriner, D. L. (2020). Model uncertainty, political contestation, and public trust in science: Evidence from the COVID-19 pandemic. *Science Advances*, *6*(43), eabd4563. https://doi.org/10.1126/sciadv.abd4563

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Kruglanski, A. W., Raviv, A., Bar-Tal, D., Raviv, A., Sharvit, K., Ellis, S., Bar, R., Pierro, A., & Mannetti, L. (2005). Says who? Epistemic authority effects in social judgment. *Advances in Experimental Social Psychology* (pp. 345–392). San Diego.

Krypotos, A.-M., Klein, R., & Jong, J. (2022). Resolving religious debates through a multiverse approach. *Commentary in MARP special issue.*

Kupor, D. M., Laurin, K., & Levav, J. (2015). Anticipating divine protection? Reminders of God can increase nonmoral risk taking. *Psychological Science*, *26*(4), 374–384. https://doi.org/10.1177/0956797614563108

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lachapelle, E., Montpetit, É., & Gauvin, J.-P. (2014). Public perceptions of expert credibility on policy issues: The role of expert framing and political worldviews - expert framing and political worldviews. *Policy Studies Journal*, *42*, 674–697. https://doi.org/10.1111/psj.12073

Ladd, K. L., & Messick, K. J. (2016). A brief history of the psychological study of the role(s) of religion. In W. Woody, R. Miller, & W. Wozniak (Eds.), *Psychological specialties in historical context: Enriching the classroom experience for teachers and students.* (pp. 204–216). Society for the Teaching of Psychology.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*, 1–14.

Lammers, J., Dubois, D., Rucker, D. D., & Galinsky, A. D. (2017). Ease of retrieval moderates the effects of power: Implications for the replicability of power recall effects. *Social Cognition*, *35*, 1–17. https://doi.org/10.1521/soco.2017.35.1.1

Landau, M. J., Kay, A. C., & Whitson, J. A. (2015). Compensatory control and the appeal of a structured world. *Psychological Bulletin*, *141*(3), 694–722. https://doi.org/10.1037/a0038703

Landy, J. F., Jia, M. (, Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., ... Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479. https://doi.org/10.1037/bul0000220

Lang, M., Mitkidis, P., Kundt, R., Nichols, A., Krajčíková, L., & Xygalatas, D. (2016). Music as a sacred cue? Effects of religious music on moral behavior. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00814

Laurin, K., Kay, A. C., & Moscovitch, D. A. (2008). On the belief in God: Towards an understanding of the emotional substrates of compensatory control. *Journal of Experimental Social Psychology*, *44*(6), 1559–1562. https://doi.org/10.1016/j.jesp.2008.07.007

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Legare, C. H., Evans, E. M., Rosengren, K. S., & Harris, P. L. (2012). The coexistence of natural and supernatural explanations across cultures and development. *Child Development*, *83*(3), 779–793. https://doi.org/10.1111/j.1467-8624.2012.01743.x

Legare, C. H., & Souza, A. L. (2014). Searching for control: Priming randomness increases the evaluation of ritual efficacy. *Cognitive Science*, *38*(1), 152–161.

Lerner, M. J. (1980). *The Belief in a Just World: A Fundamental Delusion*. Plenum. https://doi.org/10.2307/2067083

Leshinskaya, A., Contreras, J. M., Caramazza, A., & Mitchell, J. P. (2017). Neural representations of belief concepts: A representational similarity approach to social semantics. *Cerebral Cortex*, *27*, 344–357. https://doi.org/10.1093/cercor/bhw401

Levy, N. (2017). Religious beliefs are factual beliefs: Content does not correlate with context sensitivity. *Cognition*, *161*, 109–116. https://doi.org/10.1016/j.cognition.2017.01.012

Levy, N. (2019). Due deference to denialism: Explaining ordinary people's rejection of established scientific findings. *Synthese*, *196*(1), 313–327. https://doi.org/10.1007/s11229-017-1477-x

Levy, N. (2020). Belie the belief? Prompts and default states. *Religion, Brain & Behavior*, *10*(1), 35–48.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Lewis, C. A., & Cruise, S. M. (2006). Religion and happiness: Consensus, contradictions, comments and concerns. *Mental Health, Religion & Culture*, *9*(3), 213–225. https://doi.org/10.1080/13694670600615276

Liittschwager, J. (1994). *Children's reasoning about identity across transformations* (Doctoral dissertation). Stanford University. San Francisco, CA.

Lillqvist, O., & Lindeman, M. (1998). Belief in astrology as a strategy for self-verification and coping with negative life-events. *European Psychologist*, *3*(3), 202–208. https://doi.org/10.1027/1016-9040.3.3.202

Lim, C., & Putnam, R. D. (2010). Religion, social networks, and life satisfaction. *American Sociological Review*, *75*(6), 914–933. https://doi.org/10.1177/0003122410386686

Lindeman, M., Svedholm-Hakkinen, A. M., & Lipsanen, J. (2015). Ontological confusions but not mentalizing abilities predict religious belief, paranormal belief, and belief in supernatural purpose. *Cognition*, *134*, 63–76. https://doi.org/10.1016/j.cognition.2014.09.008

Lindeman, M., van Elk, M., Lipsanen, J., Marin, P., & Schjødt, U. (2019). Religious unbelief in three western European countries: Identifying and characterizing unbeliever types using latent class analysis. *The International Journal for the Psychology of Religion*, *29*(3), 184–203. https://doi.org/10.1080/10508619.2019.1591140

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science*, *26*, 1827–1832. https://doi.org/10.1177/0956797615616374

Liquin, E. G., Metz, S. E., & Lombrozo, T. (2020). Science demands explanation, religion tolerates mystery. *Cognition*, *204*, 104398. https://doi.org/10.1016/j.cognition.2020.104398

Lobato, E. J. C., Tabatabaeian, S., Fleming, M., Sulzmann, S., & Holbrook, C. (2019). Religiosity predicts evidentiary standards. *Social Psychological and Personality Science*, *11*, 546–551. https://doi.org/10.1177/1948550619869613

Lodder, P. (2022). Why researchers should not ignore measurement error and skewness in questionnaire item scores. *Commentary in MARP special issue.*

Lopez, J., & Hillygus, D. S. (2018). *Why so serious? Survey trolls and misinformation* (SSRN Scholarly Paper No. 3131087). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3131087

Luhrmann, T. M. (2005). The art of hearing God: Absorption, dissociation, and contemporary American spirituality. *Spiritus: A Journal of Christian Spirituality*, *5*(2), 133–157. https://doi.org/10.1353/scs.2006.0014

Luhrmann, T. M. (2012). *When God talks back: Understanding the American evangelical relationship with God.* Vintage Books.

Luhrmann, T. M., Padmavati, R., Tharoor, H., & Osei, A. (2015). Hearing voices in different cultures: A social kindling hypothesis. *Topics in Cognitive Science*, *7*(4), 646–663. https://doi.org/10.1111/tops.12158

Lun, V. M.-C., & Bond, M. H. (2013). Examining the relation of religion and spirituality to subjective well-being across national cultures. *Psychology of Religion and Spirituality*, *5*(4), 304–315. https://doi.org/10.1037/a0033641

Luo, W., & Chen, F. (2021). The salience of religion under an atheist state: Implications for subjective well-being in contemporary China. *Social Forces*, *100*(2), 852–878. https://doi.org/10.1093/sf/soab049

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*, 187–189.

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.* John Wiley and Sons.

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). MIT Press.

Macdonald, K., Germine, L., Anderson, A., Christodoulou, J., & McGrath, L. M. (2017). Dispelling the myth: Training in education or neuroscience decreases but does not eliminate beliefs in neuromyths. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01314

MacKay, D. J. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*, 590–604. https://doi.org/10.1162/neco.1992.4.4.590

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163. https://doi.org/10.1037/0033-2909.109.2.163

MacLeod, C. M., & MacDonald, P. A. (2000). Interdimensional interference in the Stroop effect: Uncovering the cognitive and neural anatomy of attention. *Trends in Cognitive Sciences*, *4*(10), 383–391. https://doi.org/10.1016/s1364-6613(00)01530-8

Mahoney, A. (2008). Theological expressions as costly signals of religious commitment. In J. A. Bulbulia, R. Sosis, R. Genet, C. Genet, E. Harris, & K. Wyman (Eds.), *The Evolution of Religion: Studies, Theories, and Critiques* (pp. 161–166). Collins Foundation Press.

Maier, M. E., & Steinhauser, M. (2016). Error significance but not error expectancy predicts error-related negativities for different error types. *Behavioural Brain Research*, *297*, 259–267. https://doi.org/10.1016/j.bbr.2015.10.031

Maier, S. F., & Seligman, M. E. (1976). Learned helplessness: Theory and evidence. *Journal of Experimental Psychology: General*, *105*(1), 3–46. https://doi.org/10.1037//0096-3445.105.1.3

Maier-Hein, K. H., Neher, P. F., Houde, J.-C., Côté, M.-A., Garyfallidis, E., Zhong, J., Chamberland, M., Yeh, F.-C., Lin, Y.-C., Ji, Q., Reddick, W. E., Glass, J. O., Chen, D. Q., Feng, Y., Gao, C., Wu, Y., Ma, J., He, R., Li, Q., … Descoteaux, M. (2017). The challenge of mapping the human connectome based on diffusion tractography. *Nature Communications*, *8*(1), 1349. https://doi.org/10.1038/s41467-017-01285-x

Maij, D. L. R., van Harreveld, F., Gervais, W., Schrag, Y., Mohr, C., & van Elk, M. (2017). Mentalizing skills do not differentiate believers from non-believers, but credibility enhancing displays do. *PLoS ONE*, *12*(8), e0182764. https://doi.org/10.1371/journal.pone.0182764

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Makris, N., & Pnevmatikos, D. (2007). Children's understanding of human and supernatural mind. *Cognitive Development*, *22*(3), 365–375. https://doi.org/10.1016/j.cogdev.2006.12.003

Malka, A., Lelkes, Y., Srivastava, S., Cohen, A. B., & Miller, D. T. (2012). The association of religiosity and political conservatism: The role of political engagement. *Political Psychology*, *33*(2), 275–299. https://doi.org/10.1111/j.1467-9221.2012.00875.x

Marcus, Z. J., & McCullough, M. E. (2021). Does religion make people more self-controlled? A review of research from the lab and life. *Current Opinion in Psychology*, *40*, 167–170. https://doi.org/10.1016/j.copsyc.2020.12.001

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl,

E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), 654–660. https://doi.org/10.1038/s41586-022-04492-9

Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., Bezgin, G., Eickhoff, S. B., Castellanos, F. X., Petrides, M., et al. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, *113*, 12574–12579. https://doi.org/10.1073/pnas.1608282113

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*(3), 367–380. https://doi.org/10.1016/j.cognition.2009.05.012

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. https://doi.org/10.1007/s10943-015-0099-1

Mayo, R. (2019). The skeptical (ungullible) mindset. *The Social Psychology of Gullibility: Conspiracy Theories, Fake News and Irrational Beliefs*, *140*.

Mayr, E. (1961). Cause and effect in biology: Kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science*, *134*(3489), 1501–1506.

McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition*, *107*, 343–352. https://doi.org/10.1016/j.cognition.2007.07.017

McClintock, C. H., Lau, E., & Miller, L. (2016). Phenotypic dimensions of spirituality: Implications for mental health in China, India, and the United States. *Frontiers in Psychology*, *7*, 1600. https://doi.org/10.3389/fpsyg.2016.01600

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 636–649. https://doi.org/10.1037/0278-7393.9.4.636

McCullough, M. E., Carter, E. C., DeWall, C. N., & Corrales, C. M. (2012). Religious cognition down-regulates sexually selected, characteristically male behaviors in men, but not in women. *Evolution and Human Behavior*, *33*(5), 562–568. https://doi.org/10.1016/j.evolhumbehav.2012.02.004

McCullough, M. E., Swartwout, P., Shaver, J. H., Carter, E. C., & Sosis, R. (2016). Christian religious badges instill trust in Christian and non-Christian perceivers. *Psychology of Religion and Spirituality*, *8*(2), 149–163. https://doi.org/10.1037/rel0000045

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (Second). Chapman & Hall/CRC Press. https://doi.org/10.1201/9781315372495

McGinnies, E., & Ward, C. D. (1980). Better liked than right: Trustworthiness and expertise as factors in credibility. *Personality and Social Psychology Bulletin*, *6*(3), 467–472. https://doi.org/10.1177/014616728063023

McKay, R., & Whitehouse, H. (2015). Religion and morality. *Psychological Bulletin*, *141*(2), 447. https://doi.org/10.1037/a0038455

McNamara, A. A. (2022). The impact (or lack thereof) of analysis choice on conclusions with Likert data from the Many Analysts Religion Project. *Commentary in MARP special issue.*

McPhetres, J., & Zuckerman, M. (2017). Religious people endorse different standards of evidence when evaluating religious versus scientific claims. *Social Psychological and Personality Science*, *8*, 836–842. https://doi.org/10.1177/1948550617691098

McPhetres, J., & Zuckerman, M. (2018). Religiosity predicts negative attitudes towards science and lower levels of science literacy. *PLoS One*, *13*, e0207125. https://doi.org/10.1371/journal.pone.0207125

McPhetres, J. (2018). Religiosity and Confirmation Bias or: How I learned to stop worrying and love preregistered direct replications. *OSF Preprints*. https://doi.org/10.31219/osf.io/g7apd

McPhetres, J., Jong, J., & Zuckerman, M. (2020). Religious Americans have less positive attitudes toward science, but this does not extend to other cultures. *Social Psychological and Personality Science*, 1–9. https://doi.org/10.1177/1948550620923239

Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700. https://doi.org/10.1016/j.tics.2016.07.001

Mercier, H. (2020). *Not Born Yesterday: The science of who we trust and what we believe*. Princeton University Press.

Michael, R. B., Newman, E. J., Vuorre, M., Cumming, G., & Garry, M. (2013). On the (non) persuasive power of a brain image. *Psychonomic Bulletin & Review*, *20*(4), 720–725. https://doi.org/10.3758/s13423-013-0391-6

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., & Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31. https://doi.org/10.1126/science.1245317

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. https://doi.org/10.1073/pnas.1317511111

Miller, A. S., & Hoffmann, J. P. (1995). Risk and religion: An explanation of gender differences in religiosity. *Journal for the Scientific Study of Religion*, *34*, 63–75. https://doi.org/10.2307/1386523

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032. https://doi.org/10.1080/01621459.1988.10478694

Mitkidis, P., Xygalatas, D., Buttrick, N., Porubanova, M., & Lienard, P. (2015). The impact of authority on cooperation: A cross-cultural examination of systemic trust. *Adaptive Human Behavior and Physiology*, *1*(3), 341–357.

Moore, A., & Malinowski, P. (2009). Meditation, mindfulness and cognitive flexibility. *Consciousness and Cognition*, *18*, 176–186. https://doi.org/10.1016/j.concog.2008.12.008

Morey, R. D. (2015). Multiple Comparisons with BayesFactor, Part 2 – order restrictions.

Morey, R. D., & Rouder, J. N. (2015). {BayesFactor} 0.9.11-1.

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes factors for common designs.

Morey, R. D., & Rouder, J. N. (2021). BayesFactor: Computation of Bayes factors for common designs.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … 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. https://doi.org/10.1177/2515245918797607

Mummolo, J., & Peterson, E. (2019). Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, *113*(2), 517–529. https://doi.org/10.1017/S0003055418000837

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. https://doi.org/10.1038/s41562-016-0021

Munafò, M. R., & Smith, G. D. (2018). Robust research needs many lines of evidence. *Nature*, *553*, 399–401. https://doi.org/10.1038/d41586-018-01023-3

Munro, G. D., & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, *23*(6), 636–653. https://doi.org/10.1177/0146167297236007

Murphy, J., & Martinez, N. (2022). Quantifying religiosity: A comparison of approaches based on categorical self-identification and multidimensional measures of religious activity. *Commentary in MARP special issue.*

Nakagawa, S., & Hauber, M. E. (2011). Great challenges with few subjects: Statistical strategies for neuroscientists. *Neuroscience & Biobehavioral Reviews*, *35*(3), 462–473. https://doi.org/10.1016/j.neubiorev.2010.06.003

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nelson, T. A., Abeyta, A. A., & Routledge, C. (2020). Does meaning motivate magical thinking among theists and atheists? *Social Psychological and Personality Science*, *11*(2), 176–184. https://doi.org/10.1177/1948550619829063

Neubauer, R. L. (2014). Prayer as an interpersonal relationship: A neuroimaging study. *Religion, Brain & Behavior*, *4*(2), 92–103. https://doi.org/10.1080/2153599X.2013.768288

Newheiser, A.-K., Farias, M., & Tausch, N. (2011). The functional nature of conspiracy beliefs: Examining the underpinnings of belief in the Da Vinci Code conspiracy. *Personality and Individual Differences*, *51*(8), 1007–1011. https://doi.org/10.1016/j.paid.2011.08.011

Newman, G. E., Blok, S. V., & Rips, L. J. (2006). Beliefs in afterlife as a by-product of persistence judgments. *Behavioral and Brain Sciences*, *29*(5), 480–481.

Newton, C., Feeney, J., & Pennycook, G. (2021). On the disposition to think analytically: Four distinct intuitive-analytic thinking styles. https://doi.org/10.31234/osf.io/r5wez

&

Nichols, A. D., Lang, M., Kavanagh, C., Kundt, R., Yamada, J., Ariely, D., & Mitkidis, P. (2020). Replicating and extending the effects of auditory religious cues on dishonest behavior. *PLOS ONE*, *15*(8), e0237007. https://doi.org/10.1371/journal.pone.0237007

Norenzayan, A. (2013). *Big gods: How religion transformed cooperation and conflict*. Princeton University Press. https://doi.org/10.1515/9781400848324

Norenzayan, A., & Gervais, W. M. (2013). The origins of religious disbelief. *Trends in Cognitive Sciences*, *17*(1), 20–25. https://doi.org/10.1016/j.tics.2012.11.006

Norenzayan, A., Gervais, W. M., & Trzesniewski, K. H. (2012). Mentalizing deficits constrain belief in a personal God. *PLoS ONE*, *7*(5), e36880. https://doi.org/10.1371/journal.pone.0036880

Norenzayan, A., & Lee, A. (2010). It was meant to happen: Explaining cultural variations in fate attributions. *Journal of Personality and Social Psychology*, *98*, 702–720. https://doi.org/10.1037/a0019141

Norenzayan, A., Shariff, A. F., Gervais, W. M., Willard, A. K., McNamara, R. A., Slingerland, E., & Henrich, J. (2016). The cultural evolution of prosocial religions. *Behavioral and Brain Sciences*, *39*. https://doi.org/10.1017/S0140525X14001356

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., & Bar–Anan, Y. (2012). Scientific Utopia: I. Opening Scientific Communication. *Psychological Inquiry*, *23*, 217–243. https://doi.org/10.1080/1047840x.2012.692215

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

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. https://doi.org/10.1177/1745691612459058

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in cognitive sciences*, *23*, 815–818.

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

Nosek, B. A., & Errington, T. M. (2017). Making sense of replications. *eLife*, *6*, e23383. https://doi.org/10.7554/eLife.23383

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

NPR. (2018). In psychology and other social sciences, many studies fail the reproducibility test.

O'Brien, T. L., & Noy, S. (2018). Cultural authority in comparative context: A multilevel analysis of trust in science and religion. *Journal for the Scientific Study of Religion*, *57*(3), 495–513. https://doi.org/10.1111/jssr.12537

Okulicz-Kozaryn, A. (2010). Religiosity and life satisfaction across nations. *Mental Health, Religion & Culture*, *13*(2), 155–169. https://doi.org/10.1080/13674670903273801

Oosterwijk, S. (2017). Choosing the negative: A behavioral demonstration of morbid curiosity. *PLoS ONE*, *12*(7), e0178399. https://doi.org/10.1371/journal.pone.0178399

Open Knowledge Foundation. (2020). Brits demand openness from government in tackling coronavirus.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660. https://doi.org/10.1177/1745691612462588

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. https://doi.org/10.1126/science.aac4716

Osarchuk, M., & Tatz, S. J. (1973). Effect of induced fear of death on belief in afterlife. *Journal of Personality and Social Psychology*, *27*(2), 256–260. https://doi.org/10.1037/h0034769

Padgett, V. R., & Jorgenson, D. O. (1982). Superstition and economic threat: Germany, 1918-1940. *Personality and Social Psychology Bulletin*, *8*(4), 736–741. https://doi.org/10.1177/0146167282084021

Paloutzian, R. F., & Park, C. L. (2014). *Handbook of the psychology of religion and spirituality*. Guilford Publications.

&

Paloutzian, R. (2017). Invitation to the psychology of religion. New York.

Pan, D. (2017). Is Chinese culture dualist? An answer to Edward Slingerland from a medical philosophical viewpoint. *Journal of the American Academy of Religion*, *85*, 1017–1031. https://doi.org/10.1093/jaarel/lfx028

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Park, C. L. (2005). Religion as a meaning-making framework in coping with life stress. *Journal of Social Issues*, *61*(4), 707–729.

Pashler, H., Rohrer, D., & Harris, C. R. (2013). Can the goal of honesty be primed? *Journal of Experimental Social Psychology*, *49*, 959–964. https://doi.org/10.1016/j.jesp.2013.05.011

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

Payne, G., & Williams, M. (2005). Generalization in qualitative research. *Sociology-the Journal of The British Sociological Association*, *39*(2), 295–314. https://doi.org/10.1177/0038038505050540

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. https://doi.org/10.1093/biomet/82.4.669

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60. https://doi.org/10.1145/3241036

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. *Commentary in MARP special issue.*

Peels, R., & Bouter, L. (2018a). The possibility and desirability of replication in the humanities. *Palgrave Communications*, *4*(1), 95. https://doi.org/10.1057/s41599-018-0149-x

Peels, R., & Bouter, L. (2018b). Replication in humanities just as desirable as in sciences.

Peirce, C. S. P. (1992). The fixation of belief (1877). *The Essential Peirce, Volume 1: Selected Philosophical Writings (1867–1893)*. Indiana University Press.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, *123*(3), 335–346. https://doi.org/10.1016/j.cognition.2012.03.003

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition*, *42*(1), 1–10. https://doi.org/10.3758/s13421-013-0340-7

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2020). On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs. *Judgment and Decision making*, *15*(4), 476.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, *10*(6), 549–563.

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2016). Atheists and agnostics are more reflective than religious believers: Four empirical studies and a meta-analysis. *PLOS ONE*, *11*(4), e0153039. https://doi.org/10.1371/journal.pone.0153039

Pessoa, L., Gutierrez, E., Bandettini, P. A., & Ungerleider, L. G. (2002). Neural correlates of visual working memory: fMRI amplitude predicts task performance. *Neuron*, *35*(5), 975–987. https://doi.org/10.1016/S0896-6273(02)00817-6

Petitmengin, C. (2006). Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenology and the Cognitive sciences*, *5*, 229–269. https://doi.org/10.1007/s11097-006-9022-2

Petitmengin, C., Van Beek, M., Bitbol, M., Nissou, J.-M., et al. (2017). What is it like to meditate? Methods and issues for a micro-phenomenological description of meditative experience. *Journal of Consciousness Studies*, *24*, 170–198.

Petitmengin, C., Van Beek, M., Bitbol, M., Nissou, J.-M., & Roepstorff, A. (2018). Studying the experience of meditation through micro-phenomenology. *Current opinion in psychology*. https://doi.org/10.1016/j.copsyc.2018.10.009

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change* (pp. 1–24). Springer-Verlag. https://doi.org/10.1007/978-1-4612-4964-1_1

Pew Research Center. (2012). The global religious landscape.

Pew Research Center. (2015). The future of world religions: Population growth projections, 2010-2050.

Pew Research Center. (2018). Why do levels of religious observance vary by age and country?

Pfadt, J. M., & van den Bergh, D. (2020). Bayesrel: Bayesian reliability estimation.

Plante, T. G., & Sherman, A. C. (2001). *Faith and health: Psychological perspectives*. Guilford Press.

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.

Poloma, M. M., & Pendleton, B. F. (1989). Exploring types of prayer and quality of life: A research note. *Review of Religious Research*, *31*(1), 46–53. https://doi.org/10.2307/3511023

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of applied social psychology*, *34*, 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

Potter, K. W., Huszar, L. D., & Huber, D. E. (2018). Does inhibition cause forgetting after selective retrieval? A reanalysis and failure to replicate. *Cortex*, *104*, 26–45. https://doi.org/10.1016/j.cortex.2018.03.026

Poushter, J., & Fetterolf, J. (2019). A changing world: Global views on diversity, gender equality, family life and the importance of religion.

Power, E. A. (2017). Discerning devotion: Testing the signaling theory of religion. *Evolution and Human Behavior*, *38*(1), 82–91. https://doi.org/10.1016/j.evolhumbehav.2016.07.003

Proulx, T., Inzlicht, M., & Harmon-Jones, E. (2012). Understanding all inconsistency compensation as a palliative response to violated expectations. *Trends in Cognitive Sciences*, *16*(5), 285–291. https://doi.org/10.1016/j.tics.2012.04.002

Purzycki, B. G., Apicella, C., Atkinson, Q. D., Cohen, E., McNamara, R. A., Willard, A. K., Xygalatas, D., Norenzayan, A., & Henrich, J. (2016). Moralistic gods, supernatural punishment and the expansion of human sociality. *Nature*, *530*(7590), 327–330. https://doi.org/10.1038/nature16980

Purzycki, B. G., & Arakchaa, T. (2013). Ritual behavior and trust in the Tyva Republic. *Current Anthropology*, *54*(3), 381–388. https://doi.org/10.1086/670526

Purzycki, B. G., Ross, C. T., Apicella, C., Atkinson, Q. D., Cohen, E., McNamara, R. A., Willard, A. K., Xygalatas, D., Norenzayan, A., & Henrich, J. (2018). Material security, life history, and moralistic religions: A cross-cultural examination. *PLoS ONE*, *13*(3), e0193856. https://doi.org/10.1371/journal.pone.0193856

Pyle, R. E. (2006). Trends in religious stratification: Have religious group socioeconomic distinctions declined in recent decades? *Sociology of Religion*, *67*, 61–79. https://doi.org/10.1093/socrel/67.1.61

Pyysiäinen, I. (2006). No evidence of a specific adaptation. *Behavioral and Brain Sciences*, *29*(5), 483–484.

Qualtrics. (2019). Online survey sofware qualtrics.

R Development Core Team. (2004). *R: A Language and Environment for Statistical Computing*.

&

Raaijmakers, Q. A., Van Hoof, J., t Hart, H., Verbogt, T., & Vollebergh, W. A. (2000). Adolescents' midpoint responses on Likert-type scale items: Neutral or missing values? *International Journal of Public Opinion Research*, *12*, 208–216.

Raven, J., Raven, J. C., & Court, J. H. (1998). Manual for raven's progressive matricesand vocabulary scales. Oxford Psychologists Press.

Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*(1), 1–48. https://doi.org/10.1006/cogp.1999.0735

Reitsma, J., Pelzer, B., Scheepers, P., & Schilderman, H. (2012). Believing and belonging in Europe. *European Societies*, *14*(4), 611–632. https://doi.org/10.1080/14616696.2012.726367

Reynolds, T. W., Bostrom, A., Read, D., & Morgan, M. G. (2010). Now what do people know about global climate change? Survey studies of educated laypeople. *Risk Analysis: An International Journal*, *30*(10), 1520–1538. https://doi.org/10.1111/j.1539-6924.2010.01448.x

Rice, T. W. (2003). Believe it or not: Religious and other paranormal beliefs in the United States. *Journal for the Scientific Study of Religion*, *42*(1), 95–106. https://doi.org/10.1111/1468-5906.00163

Riekki, T., Lindeman, M., & Lipsanen, J. (2013). Conceptions about the mind-body problem and their relations to afterlife beliefs, paranormal beliefs, religiosity, and ontological confusions. *Advances in Cognitive Psychology*, *9*(3), 112–120. https://doi.org/10.2478/v10053-008-0138-5

Rips, L. J., Blok, S., & Newman, G. (2006). Tracing the identity of objects. *Psychological Review*, *113*(1), 1.

Risen, J. L. (2016). Believing what we do not believe: Acquiescence to superstitious beliefs and other powerful intuitions. *Psychological Review*, *123*(2), 182–207. https://doi.org/10.1037/rev0000017

Ritzema, R. J., & Young, C. (1983). Causal schemata and the attribution of supernatural causality. *Journal of Psychology and Theology*, *11*, 36–43. https://doi.org/10.1177/009164718301100106

Robins, R. W., Trzesniewski, K. H., Tracy, J. L., Gosling, S. D., & Potter, J. (2002). Global self-esteem across the life span. *Psychology and Aging*, *17*, 423–434. https://doi.org/10.1037//0882-7974.17.3.423

Robinson, C. (2010). Cross-cutting messages and political tolerance: An experiment using evangelical protestants. *Political Behavior*, *32*(4), 495–515. https://doi.org/10.1007/s11109-010-9118-9

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

Rosmarin, D. H., Pargament, K. I., Pirutinsky, S., & Mahoney, A. (2010). 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*(7), 799–808. https://doi.org/10.1016/j.janxdis.2010.05.014

Ross, R. M., Sulik, J., Buczny, J., & Schivinski, B. (2022). Many analysts and few incentives. *Commentary in MARP special issue.*

&

Roth, L. M., & Kroll, J. C. (2007). Risky business: Assessing risk preference explanations for gender differences in religiosity. *American Sociological Review*, *72*(2), 205–220. https://doi.org/10.1177/000312240707200204

Rothbaum, F., Weisz, J. R., & Snyder, S. S. (1982). Changing the world and changing the self: A two-process model of perceived control. *Journal of Personality and Social Psychology*, *42*(1), 5–37. https://doi.org/10.1037//0022-3514.42.1.5

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877–903. https://doi.org/10.1080/00273171.2012.734737

Rouder, J. N., Haaf, J. M., & Aust, F. (2018). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, *85*(1), 41–56. https://doi.org/10.1080/03637751.2017.1394581

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. https://doi.org/10.1037/met0000216

Rouder, J. N., Kumar, A., & Haaf, J. M. (2019). Why most studies of individual differences with inhibition tasks are bound to fail. https://doi.org/10.31234/osf.io/3cjr5

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/J.JMP.2012.08.001

Rowhani-Farid, A., Aldcroft, A., & Barnett, A. G. (2020). Did awarding badges increase data sharing in *BMJ Open* ? A randomized controlled trial. *Royal Society Open Science*, *7*(3), 191818. https://doi.org/10.1098/rsos.191818

Ruiter, S., & van Tubergen, F. (2009). Religious attendance in cross-national perspective: A multilevel analysis of 60 countries. *American Journal of Sociology*, *115*(3), 863–895. https://doi.org/10.1086/603536

Rutjens, B. T., Sengupta, N., van der Lee, R., van Koningsbruggen, G. M., Martens, J. P., Rabelo, A., & Sutton, R. M. (2021). Science skepticism across 24 countries. *Social Psychological and Personality Science*, 19485506211001329. https://doi.org/10.1177/19485506211001329

Rutjens, B. T., & van der Lee, R. (2020). Spiritual skepticism? Heterogeneous science skepticism in the Netherlands. *Public Understanding of Science*, *29*, 335–352. https://doi.org/10.1177/0963662520908534

Rutjens, B. T., van der Pligt, J., & van Harreveld, F. (2010). Deus or Darwin: Randomness and belief in theories about the origin of life. *Journal of Experimental Social Psychology*, *46*(6), 1078–1080. https://doi.org/10.1016/j.jesp.2010.07.009

Rutjens, B. T., van Harreveld, F., van der Pligt, J., Kreemers, L. M., & Noordewier, M. K. (2013). Steps, stages, and structure: Finding compensatory order in scientific theories. *Journal of Experimental Psychology: General*, *142*(2), 313–318. https://doi.org/10.1037/a0028716

Sales, S. M. (1972). Economic threat as a determinant of conversion rates in authoritarian and nonauthoritarian churches. *Journal of Personality and Social Psychology*, *23*(3), 420–428. https://doi.org/10.1037/h0033157

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., … McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117

Sanchez, C., Sundermeier, B., Gray, K., & Calin-Jageman, R. J. (2017). Direct replication of Gervais & Norenzayan (2012): No evidence that analytic thinking decreases religious belief. *PLoS ONE*, *12*(2), e0172636. https://doi.org/10.1371/journal.pone.0172636

Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E.-J. (2022). Comparing analysis blinding with preregistration in the many-analysts religion project. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.31234/osf.io/6dn8f

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*. https://doi.org/10.31234/osf.io/6k5gr

Scheibehenne, B., Gronau, Q. F., Jamil, T., & Wagenmakers, E.-J. (2017). Fixed or random? A resolution through model-averaging. Reply to Carlsson, Schimmack, Williams, and Burkner. *Psychological Science*, *28*, 1698–1701.

Schjoedt, U. (2009). The religious brain: A general introduction to the experimental neuroscience of religion. *Method & Theory in the Study of Religion*, *21*(3), 310–339. https://doi.org/10.1163/157006809x460347

Schjoedt, U., & Bulbulia, J. (2011). The need to believe in conflicting propositions. *Religion, Brain & Behavior*, *1*(3), 236–239. https://doi.org/10.1080/2153599X.2011.647857

Schjoedt, U., Sørensen, J., Nielbo, K. L., Xygalatas, D., Mitkidis, P., & Bulbulia, J. (2013). Cognitive resource depletion in religious interactions. *Religion, Brain and Behavior*, *3*(1), 39–55. https://doi.org/10.1080/2153599X.2012.736714

Schjoedt, U., Stødkilde-Jørgensen, H., Geertz, A. W., Lund, T. E., & Roepstorff, A. (2011). The power of charisma-perceived charisma inhibits the frontal executive network of believers in intercessory prayer. *Social Cognitive and Affective Neuroscience*, *6*(1), 119–127. https://doi.org/10.1093/scan/nsq023

Schjoedt, U., Stødkilde-Jørgensen, H., Geertz, A. W., & Roepstorff, A. (2009). Highly religious participants recruit areas of social cognition in personal prayer. *Social Cognitive and Affective Neuroscience*, *4*(2), 199–207. https://doi.org/10.1093/scan/nsn050

Schjoedt, U., & van Elk, M. (2019). The Neuroscience of Religion. In J. Barrett (Ed.), *Oxford Handbook of the Cognitive Science of Religion*. Oxford Univerity Press.

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. https://doi.org/10.1037/a0015108

Schönbrodt, F. D. (2017). BFDA: An R package for Bayes factor design analysis (Version 0.2)
Programmers: _:n41361.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schott, E., Rhemtulla, M., & Byers-Heinlein, K. (2018). Should I test more babies? Solutions for transparent data peeking. *Infant Behavior and Development*. https://doi.org/10.31234/osf.io/gxfaj

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. *Commentary in MARP special issue.*

Schwadel, P. (2016). Does higher education cause religious decline? A longitudinal analysis of the within- and between-person effects of higher education on religiosity. *The Sociological Quarterly*, *57*, 759–786. https://doi.org/10.1111/tsq.12153

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C. M., van Assen, M. A. L. M., Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson, E., Snellman, K., Amy Sommer, S., Otner, S. M. G., Robinson, D., … 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. https://doi.org/10.1016/j.obhdp.2021.02.003

Scurich, N., & Shniderman, A. (2014). The selective allure of neuroscientific explanations. *PLoS One*, *9*(9), e107529. https://doi.org/10.1371/journal.pone.0107529

Sedikides, C. (2010). Why does religiosity persist? *Personality and Social Psychology Review*, *14*(1), 3–6.

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.

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*. https://doi.org/10.3758/s13428-019-01231-3

Seybold, K. S., & Hill, P. C. (2001). The role of religion and spirituality in mental and physical health. *Current Directions in Psychological Science*, *10*(1), 21–24. https://doi.org/10.1111/1467-8721.00106

Shah, A. K., Mullainathan, S., & Shafir, E. (2012). Some consequences of having too little. *Science*, *338*, 682–685. https://doi.org/10.1126/science.1222426

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., Kavvadia, F., & Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE*, *8*, e56515. https://doi.org/10.1371/journal.pone.0056515

Shariff, A. F., & Norenzayan, A. (2007). God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game. *Psychological Science*, *18*(9), 803–809. https://doi.org/10.1111/j.1467-9280.2007.01983.x

Shariff, A. F., Willard, A. K., Andersen, T., & Norenzayan, A. (2016). Religious priming: A meta-analysis with a focus on prosociality. *Personality and Social Psychology Review*, *20*(1), 27–48. https://doi.org/10.1177/1088868314568811

Shariff, A. F., Willard, A. K., Muthukrishna, M., Kramer, S. R., & Henrich, J. (2016). What is the association between religious affiliation and children's altruism? *Current Biology*, *26*(15), R699–R700. https://doi.org/10.1016/j.cub.2016.06.031

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423–428. https://doi.org/10.1037/a0025391

Sherwood, H. (2018). Religion: Why faith is becoming more and more popular. *The Guardian*.

Sibley, C. G., Greaves, L. M., Satherley, N., Wilson, M. S., Overall, N. C., Lee, C. H. J., Milojev, P., Bulbulia, J., Osborne, D., Milfont, T. L., Houkamau, C. A., Duck, I. M., Vickers-Jones, R., & Barlow, F. K. (2020). Effects of the COVID-19 pandemic and nationwide lockdown on trust, attitudes toward government, and well-being. *American Psychologist*, *75*(5), 618. https://doi.org/10.1037/amp0000662

Siedlecka, E., & Denson, T. F. (2019). Experimental methods for inducing basic emotions: A qualitative review. *Emotion Review*, *11*, 87–97.

Silberzahn, R., & Uhlmann, E. L. (2015). Crowdsourced research: Many hands make tight work. *Nature*, *526*(7572), 189–191. https://doi.org/10.1038/526189a

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahnik, Š., Bai, F., Bannard, C., Bonnier, E., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simkin, H. (2020). The centrality of events, religion, spirituality, and subjective well-being in Latin American Jewish immigrants in Israel. *Frontiers in Psychology*, *11*, 576402. https://doi.org/10.3389/fpsyg.2020.576402

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False–Positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1037/e519702015-014

Simonsohn, U. (2015a). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569. https://doi.org/10.1177/0956797614567341

Simonsohn, U. (2015b). The default Bayesian test is prejudiced against small effects.

Singelis, T. M., Triandis, H. C., Bhawuk, D. P. S., & Gelfand, M. J. (1995). Horizontal and vertical dimensions of individualism and collectivism: A theoretical and measurement refinement. *Cross-Cultural Research*, *29*(3), 240–275. https://doi.org/10.1177/106939719502900302

Slingerland, E., & Chudek, M. (2011). The prevalence of mind–body dualism in early China. *Cognitive Science*, *35*, 997–1007. https://doi.org/10.1111/j.1551-6709.2011.01186.x

Slingerland, E., & Collard, M. (2011). *Creating consilience: Integrating the sciences and the humanities*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199794393.001.0001

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

Smith, C., & Faris, R. (2005). Socioeconomic inequality in the American religious system: An update and assessment. *Journal for the Scientific Study of Religion*, *44*, 95–104. https://doi.org/10.1111/j.1468-5906.2005.00267.x

Smith, C. T., De Houwer, J., & Nosek, B. A. (2013). Consider the source: Persuasion of implicit evaluations is moderated by source credibility. *Personality and Social Psychology Bulletin*, *39*(2), 193–205. https://doi.org/10.1177/0146167212472374

Smith, E. (2022). Individual-level versus country-level moderation. *Commentary in MARP special issue.*

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, *44*, 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

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*(4), 614–636. https://doi.org/10.1037/0033-2909.129.4.614

Snoek, L., van der Miesen, M. M., Beemsterboer, T., van der Leij, A., Eigenhuis, A., & Scholte, H. S. (2020). The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses. *bioRxiv*. https://doi.org/10.1101/2020.06.16.155317

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour*, *5*(8), 990–997. https://doi.org/10.1038/s41562-021-01142-4

Soler, M. (2012). Costly signaling, ritual and cooperation: Evidence from Candomblé, an Afro-Brazilian religion. *Evolution and Human Behavior*, *33*(4), 346–356. https://doi.org/10.1016/j.evolhumbehav.2011.11.004

Sosis, R. (2006). Religious behaviors, badges, and bans: Signaling theory and the evolution of religion. *Where God and Man Meet. How the Brain and Evolutionary Sciences are Revolutionizing Our Understanding of Religion and Spirituality.*, *1*, 61–86.

Sosis, R., & Alcorta, C. (2003). Signaling, solidarity, and the sacred: The evolution of religious behavior. *Evolutionary Anthropology: Issues, News, and Reviews*, *12*(6), 264–274. https://doi.org/10.1002/evan.10120

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science*, *30*(5), 711–727. https://doi.org/10.1177/0956797619831612

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*, 886–899. https://doi.org/10.1177/1745691615609918

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic viligance. *Mind & Language*, *25*, 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Sperber, D. (1997). Intuitive and reflective beliefs. *Mind & Language*, *12*(1), 67–83. https://doi.org/10.1111/j.1468-0017.1997.tb00062.x

Sperber, D. (2010). The Guru effect. *Review of Philosophy and Psychology*, *1*(4), 583–592. https://doi.org/10.1007/s13164-010-0025-0

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, *84*(5), 1041–1053. https://doi.org/10.1037/0022-3514.84.5.1041

Starns, J. J., Cataldo, A. M., Rotello, C. M., Annis, J., Aschenbrenner, A., Bröder, A., Cox, G., Criss, A., Curl, R. A., Dobbins, I. G., Dunn, J., Enam, T., Evans, N. J., Farrell, S., Fraundorf, S. H., Gronlund, S. D., Heathcote, A., Heck, D. W., Hicks, J. L., … Wilson, J. (2019). Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Advances in Methods and Practices in Psychological Science*, *2*(4), 335–349. https://doi.org/10.1177/2515245919869583

Starreveld, P. A., & La Heij, W. (2017). Picture-word interference is a Stroop effect: A theoretical analysis and new empirical findings. *Psychonomic Bulletin & Review*, *24*(3), 721–733. https://doi.org/10.3758/s13423-016-1167-6

Stavrova, O. (2015). Religion, self-rated health, and mortality: Whether religiosity delays death depends on the cultural context. *Social Psychological and Personality Science*, *6*(8), 911–922. https://doi.org/10.1177/1948550615593149

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*(1), 90–105. https://doi.org/10.1016/j.ssresearch.2012.07.002

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

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. https://doi.org/10.3758/s13428-018-01189-8

Sterelny, K. (2018). Religion re-explained. *Religion, Brain & Behavior*, *8*(4), 406–425. https://doi.org/10.1080/2153599X.2017.1323779

Strack, F., Schwarz, N., & Gschneidinger, E. (1985). Happiness and reminiscing: The role of time perspective, affect, and mode of thinking. *Journal of Personality and Social Psychology*, *49*, 1460–1469. https://doi.org/10.1037//0022-3514.49.6.1460

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59–71. https://doi.org/10.1177/1745691613514450

&

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Sturgis, P., Roberts, C., & Smith, P. (2014). Middle alternatives revisited: How the neither/nor response acts as a way of saying "I don't know"? *Sociological Methods & Research*, *43*(1), 15–38. https://doi.org/10.1177/0049124112452527

Swigart, K. L., Anantharaman, A., Williamson, J. A., & Grandey, A. A. (2020). Working while liberal/conservative: A review of political ideology in organizations. *Journal of Management*, *46*(6), 1063–1091. https://doi.org/10.1177/0149206320909419

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Tamminen, K. A., & Poucher, Z. A. (2018). Open science in sport and exercise psychology: Review of current approaches and considerations for qualitative inquiry. *Psychology of Sport and Exercise*, *36*, 17–28. https://doi.org/10.1016/j.psychsport.2017.12.010

Tang, J., Critchley, H. D., Glaser, D., Dolan, R. J., & Butterworth, B. (2006). Imaging informational conflict: A functional magnetic resonance imaging study of numerical Stroop. *Journal of Cognitive Neuroscience*, *18*(12), 2049–2062. https://doi.org/10.1162/jocn.2006.18.12.2049

Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*, 401–409. https://doi.org/10.1037/h0058700

Taves, A. (1999). *Fits, trances, & visions: Experiencing religion and explaining experience from Wesley to James*. Princeton University Press. https://doi.org/10.2307/2674486

Taves, A. (2010). No field is an island: Fostering collaboration between the academic study of religion and the sciences. *Method & Theory in the Study of Religion*, *22*(2), 170–188. https://doi.org/10.1163/157006810X512356

Taylor, M., Cartwright, B. S., & Bowden, T. (1991). Perspective taking and theory of mind: Do children predict interpretive diversity as a function of differences in observers' knowledge? *Child Development*, *62*(6), 1334–1351. https://doi.org/10.1111/j.1467-8624.1991.tb01609.x

TechCrunch. (2019). Online translation sofware DeepL.

Teper, R., & Inzlicht, M. (2012). Meditation, mindfulness and executive control: The importance of emotional acceptance and brain-based performance monitoring. *Social Cognitive and Affective Neuroscience*, *8*, 85–92. https://doi.org/10.1093/scan/nss045

The ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference: *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245919900809

The White House Press Briefing. (2020). Remarks by President Trump in Press Briefing [Statements & Releases].

Thoresen, C. E. (1999). Spirituality and health: Is there a relationship? *Journal of Health Psychology*, *4*(3), 291–300. https://doi.org/10.1177/135910539900400314

&

Tierney, W., Hardy, J., III, Ebersole, C., Viganola, D., Clemente, E., Gordon, E., Hoogeveen, S., Haaf, J. M., Dreber, A., Johannesson, M., Pfeiffer, T., Chapman, H., Gantman, A., Vanaman, M., DeMarree, K., Igou, E., Wylie, J., Storbeck, J., Andreychik, M., … Uhlmann, E. L. (2021). A creative destruction approach to replication: Implicit work and sex morality across cultures. *Journal of Experimental Social Psychology*, *93*, 104060.

Tiryakian, E. A. (1991). L'exceptionnelle vitalité religieuse aux Etats-Unis: Une relecture de Protestant-Catholic-Jew. *Social Compass*, *38*(3), 215–238. https://doi.org/10.1177/003776891038003002

Treiber, J. M., White, N. S., Steed, T. C., Bartsch, H., Holland, D., Farid, N., McDonald, C. R., Carter, B. S., Dale, A. M., & Chen, C. C. (2016). Characterization and correction of geometric distortions in 814 diffusion weighted images. *PLoS ONE*, *11*, e0152472. https://doi.org/10.1371/journal.pone.0152472

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging*, *29*, 1310. https://doi.org/10.1109/tmi.2010.2046908

Tversky, A., & Gati, I. (1978). Studies of similarity. *Cognition and Categorization*, *1*(1978), 79–98.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., McCarthy, R. J., Riegelman, A., Silberzahn, R., & Nosek, B. A. (2019). Scientific Utopia III: Crowdsourcing science. *Perspectives on Psychological Science*, *14*(5), 711–733. https://doi.org/10.1177/1745691619850561

Uhlmann, E. L., Poehlman, T. A., & Bargh, J. A. (2009). American moral exceptionalism. *Social and psychological bases of ideology and system justification* (pp. 27–52). Oxford University Press.

Uhlmann, E. L., Poehlman, T. A., Tannenbaum, D., & Bargh, J. A. (2011). Implicit Puritanism in American moral cognition. *Journal of Experimental Social Psychology*, *47*(2), 312–320. https://doi.org/10.1016/j.jesp.2010.10.013

Vail, K. E., Arndt, J., & Abdollahi, A. (2012). Exploring the existential function of religion and supernatural agent beliefs among christians, muslims, atheists, and agnostics. *Personality and Social Psychology Bulletin*, *38*(10), 1288–1300. https://doi.org/10.1177/0146167212449361

Valtonen, J., Ahn, W.-k., & Cimpian, A. (2021). Neurodualism: People assume that the brain affects the mind more than the mind affects the brain. *Cognitive Science*, *45*(9), e13034. https://doi.org/10.1111/cogs.13034

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. *Commentary in MARP special issue.*

van der Miesen, M. M., van der Lande, G. J., Hoogeveen, S., Schjoedt, U., & van Elk, M. (2022). The effect of source credibility on the evaluation of statements in a spiritual and scientific context: A registered report study. *Comprehensive Results in Social Psychology.* https://doi.org/10.1080/23743603.2022.2041984

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2020). Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and

Spearman's $\rho$. *Journal of Applied Statistics*, *47*(16), 2984–3006. https://doi.org/10.1080/02664763.2019.1709053

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Bayesian inference for Kendall's rank correlation coefficient. *The American Statistician*, *72*, 303–308. https://doi.org/10.1080/00031305.2016.1264998

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., et al. (2020). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 1–14.

van Elk, M. (2019). Replication and open science in the psychology of religion: Editorial to the special issue. *The International Journal for the Psychology of Religion*, *29*(4), 227–229. https://doi.org/10.1080/10508619.2019.1687189

van Elk, M., & Lodder, P. (2018). Experimental manipulations of personal control do not increase illusory pattern perception. *Collabra: Psychology*, *4*(1), 19. https://doi.org/10.1525/collabra.155http://doi.org/10.1525/collabra.155

van Elk, M. (2021). 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), 1–3. https://doi.org/10.1080/10508619.2021.1877029

van Elk, M., & Aleman, A. (2017). Brain mechanisms in religion and spirituality: An integrative predictive processing framework. *Neuroscience & Biobehavioral Reviews*, *73*, 359–378. https://doi.org/10.1016/J.NEUBIOREV.2016.12.031

van Elk, M., Groenendijk, E., & Hoogeveen, S. (2020). Placebo brain stimulation affects subjective but not neurocognitive measures of error processing. *Journal of Cognitive Enhancement*, *4*(4), 389–400. https://doi.org/10.1007/s41465-020-00172-6

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. https://doi.org/10.3389/fpsyg.2015.01365

van Elk, M., Rutjens, B. T., van der Pligt, J., & van Harreveld, F. (2016). Priming of supernatural agent concepts and agency detection. *Religion, Brain & Behavior*, *6*, 4–33.

van Elk, M., & Snoek, L. (2020). The relationship between individual differences in grey matter volume and religiosity and mystical experiences: A pre-registered voxel-based morphometry study. *European Journal of Neuroscience*, *51*, 850–865. https://doi.org/10.1111/ejn.14563

van Elk, M., & Wagenmakers, E.-J. (2017). Can the experimental study of religion be advanced using a Bayesian predictive framework? *Religion, Brain & Behavior*, *7*(4), 331–334. https://doi.org/10.1080/2153599X.2016.1249915

van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-.-J. (2017). Estimates of Between-Study Heterogeneity for 705 Meta-Analyses Reported in *Psychological Bulletin* From 1990–2013. *Journal of Open Psychology Data*, *5*.

Van Leeuwen, N. (2014). Religious credence is not factual belief. *Cognition*, *133*, 698–715. https://doi.org/10.1016/j.cognition.2014.08.015

van Lissa, C. J. (2022). Complementing preregistered confirmatory analyses with rigorous, reproducible exploration using machine learning. *Commentary in MARP special issue*.

van Lissa, C. J., Peikert, A., & Brandmaier, A. M. (2021). *Worcs: Workflow for open reproducible code in science*. Manual.

van Maanen, L., van Rijn, H., & Borst, J. P. (2009). Stroop and picture—word interference are two sides of the same coin. *Psychonomic Bulletin & Review*, *16*(6), 987–999. https://doi.org/10.3758/pbr.16.6.987

van Mulukom, V. (2017). Remembering religious rituals: Autobiographical memories of high-arousal religious rituals considered from a narrative processing perspective. *Religion, Brain & Behavior*, *7*, 191–205. https://doi.org/10.1080/2153599x.2016.1232304

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. https://doi.org/10.31234/osf.io/4frms

Van Tongeren, D. R., Pennington, A. R., McIntosh, D. N., Newton, T., Green, J. D., Davis, D. E., & Hook, J. N. (2017). Where, O death, is thy sting? The meaning-providing function of beliefs in literal immortality. *Mental Health, Religion & Culture*, *20*(5), 413–427. https://doi.org/10.1080/13674676.2017.1355358

van Veen, V., & Carter, C. S. (2005). Separating semantic conflict and response conflict in the Stroop task: A functional MRI study. *Neuroimage*, *27*(3), 497–504.

van Dongen, N. N. N., van Doorn, J. B., Gronau, Q. F., van Ravenzwaaij, D., Hoekstra, R., Haucke, M. N., Lakens, D., Hennig, C., Morey, R. D., Homer, S., Gelman, A., Sprenger, J., & Wagenmakers, E.-J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician*, *73*(sup1), 328–339. https://doi.org/10.1080/00031305.2019.1565553

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.31234/osf.io/2yphf

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R^ for assessing convergence of MCMC. *Bayesian Analysis*, 1–28. https://doi.org/10.1214/20-BA1221

Veldkamp, C., Bakker, M., van Assen, M., Crompvoets, E., Ong, H., Soderberg, C., Mellor, D., Nosek, B., & Wicherts, J. (2017). Restriction of opportunistic use of researcher degrees of freedom in pre-registrations on the Open Science Framework. *The Human Fallibility of Scientists: Dealing with error and bias in academic research* (pp. 106–133).

Verburg, M., Huzarevich, J., Hughes, K., Thompson, J., Figueira, R., Conner Koreis, C., Morse, S., & Riordan, C. (2016). Who's in control: You, God, or government?

Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian ests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475. https://doi.org/10.1037/a0036731

&

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*, 1525. https://doi.org/10.3389/fpsyg.2019.01525

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. *Commentary in MARP special issue.*

Vogt, B. A. (2005). Pain and emotion interactions in subregions of the cingulate gyrus. *Nature Reviews Neuroscience*, *6*, 533. https://doi.org/10.1038/nrn1704

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., Alquist, J. L., Baker, M. D., Brizi, A., Bunyi, A., Butschek, G. J., Campbell, C., Capaldi, J., Cau, C., Chambers, H., Chatzisarantis, N. L. D., Christensen, W. J., Clay, S. L., Curtis, J., … Albarracín, D. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science*, *32*(10), 1566–1581. https://doi.org/10.1177/0956797621989733

Vuorre, M. (2016). Meta-analysis is a special case of Bayesian multilevel modeling.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin–Hudon, E., Bulnes, L. C., Caldwell, T. L., Calin–Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., … Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*, 917–928. https://doi.org/10.1177/1745691616674458

Wagenmakers, E.-J., Etz, A., Gronau, Q., & Dablander, F. (2018). The single most prevalent misinterpretation of Bayes' rule [Blog Post].

Wagenmakers, E.-J., Gronau, Q. F., & Vandekerckhove, J. (2019). Five Bayesian intuitions for the stopping rule principle. https://doi.org/10.31234/osf.io/5ntkd

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432. https://doi.org/10.1037/a0022790

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 627–633. https://doi.org/10.1177/1745691612463078

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

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*(11), 1473–1480. https://doi.org/10.1038/s41562-021-01211-8

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, *605*(7910), 423–425. https://doi.org/10.1038/d41586-022-01332-8

Bandiera_abtest: a Cg_type: Comment Subject_term: Publishing, Research management

Wagge, J. R., Brandt, M. J., Lazarevic, L. B., Legate, N., Christopherson, C., Wiggins, B., & Grahe, J. E. (2019). Publishing research with undergraduate students via replication work: The Collaborative Replications and Education Project. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00247

Wald, K. D., & Calhoun-Brown, A. (2014). *Religion and Politics in the United States*. Rowman & Littlefield.

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*(2), 337–354. https://doi.org/10.1037/0022-3514.92.2.337

Wang, C. S., Whitson, J. A., & Menon, T. (2012). Culture, control, and illusory pattern perception. *Social Psychological and Personality Science*, *3*(5), 630–638. https://doi.org/10.1177/1948550611433056

Wang, S., Peterson, D. J., Gatenby, J. C., Li, W., Grabowski, T. J., & Madhyastha, T. M. (2017). Evaluation of field map and nonlinear registration methods for correction of susceptibility artifacts in diffusion MRI. *Frontiers in Neuroinformatics*, *11*, 17. https://doi.org/10.3389/fninf.2017.00017

Warner, R. S. (1993). Work in progress toward a new paradigm for the sociological study of religion in the United States. *American Journal of Sociology*, *98*(5), 1044–1093. https://doi.org/10.1086/230139

Watson-Jones, R. E., Busch, J. T. A., Harris, P. L., & Legare, C. H. (2017). Does the body survive death? Cultural variation in beliefs about life everlasting. *Cognitive Science*, *41*, 455–476. https://doi.org/10.1111/cogs.12430

Weber, E. U., & Stern, P. C. (2011). Public understanding of climate change in the United States. *American Psychologist*, *66*(4), 315–328. https://doi.org/10.1037/a0023253

Weber, M. (1930). *The Protestant ethic and the spirit of capitalism*. Routledge. https://doi.org/10.4324/9780203995808

Weber, S. R., & Pargament, K. I. (2014). The role of religion and spirituality in mental health. *Current Opinion in Psychiatry*, *27*(5), 358–363. https://doi.org/10.1097/YCO.0000000000000080

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*(3), 470–477. https://doi.org/10.1162/jocn.2008.20040

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, *114*(43), 11374–11379. https://doi.org/10.1073/pnas.1704347114

Weisman, K., Legare, C. H., Smith, R. E., Dzokoto, V. A., Aulino, F., Ng, E., Dulin, J. C., Ross-Zehnder, N., Brahinsky, J. D., & Luhrmann, T. M. (2021). Similarities and differences in concepts of mental life among adults and children in five cultures. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01184-8

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for ANOVA designs. *The American Statistician*, *66*(2), 104–111. https://doi.org/10.1080/00031305.2012.695956

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*, 475–482. https://doi.org/10.1007/bf03395630

White, C. J. M., Norenzayan, A., & Schaller, M. (2018). The content and correlates of belief in karma across cultures. *Personality and Social Psychology Bulletin*, 18. https://doi.org/10.1177/0146167218808502

White, C. (2015). Establishing personal identity in reincarnation: Minds and bodies reconsidered. *Journal of Cognition and Culture*, *15*(3-4), 402–429.

Whitehouse, H., & Lanman, J. A. (2014). The ties that bind us: Ritual, fusion, and identification. *Current Anthropology*, *55*(6), 674–695. https://doi.org/10.1086/678698

Whitson, J. A., & Galinsky, A. D. (2008). Lacking control increases illusory pattern perception. *Science*, *322*(5898), 115–117. https://doi.org/10.1126/science.1159845

WHOQOL Group. (1998). Development of the World Health Organization WHOQOL-BREF Quality of Life Assessment. *Psychological Medicine*, *28*(3), 551–558. https://doi.org/10.1017/s0033291798006667

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.31219/osf.io/umq8d

Wiebe, D. (2009). Religious biases in funding religious studies research? *Religio: Revue pro Religionistiku*, *17*(2), 125–140.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.18

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. https://doi.org/10.1177/2515245918767122

Wilson, M. S., Bulbulia, J., & Sibley, C. G. (2013). Differences and similarities in religious and paranormal beliefs: A typology of distinct faith signatures. *Religion, Brain & Behavior*, *4*(2), 104–126. https://doi.org/10.1080/2153599x.2013.779934

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience*, *18*(4), 582–589. https://doi.org/10.1038/nn.3973

Wingen, T., Berkessel, J. B., & Englich, B. (2020). No replication, no trust? How low replicability influences trust in psychology. *Social Psychological and Personality Science*, *11*(4), 454–463. https://doi.org/10.1177/1948550619877412

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage*, *92*, 381–397. https://doi.org/10.1016/j.neuroimage.2014.01.060

Wissenschaft im Dialog. (2020). Science barometer special edition on Corona.

World Bank Group. (2017). World Bank Group - International Development, Poverty, & Sustainability.

World Values Survey. (2010). Wave 6 Official Aggregate v. 20150418.

Wright, R. (2010). *The evolution of God: The origins of our beliefs*. Hachette UK.

Wulff, D. M. (1998). Rethinking the rise and fall of the psychology of religion. *Religion in the Making* (pp. 181–202). Brill.

Xygalatas, D., Mitkidis, P., Fischer, R., Reddish, P., Skewes, J., Geertz, A. W., Roepstorff, A., & Bulbulia, J. (2013). Extreme rituals promote prosociality. *Psychological Science*, *24*(8), 1602–1605. https://doi.org/10.1177/0956797612472910

Xygalatas, D., Schjoedt, U., Bulbulia, J., Konvalinka, I., Jegindø, E.-M., Reddish, P., Geertz, A. W., & Roepstoff, A. (2013). Autobiographical memory in a fire-walking ritual. *Journal of Cognition and Culture*, *13*, 1–16. https://doi.org/10.1163/15685373-12342081

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*, 665. https://doi.org/10.1038/nmeth.1635

Yeung, A. W. K. (2017). Do neuroscience journals accept replications? A survey of literature. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00468

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931–959. https://doi.org/10.1037/0033-295x.111.4.939

Yong, E. (2012). A failed replication draws a scathing personal attack from a psychology professor. *Discover Magazine*.

Yonker, J. E., Edman, L. R. O., Cresswell, J., & Barrett, J. L. (2016). Primed analytic thought and religiosity: The importance of individual characteristics. *Psychology of Religion and Spirituality*, *8*(4), 298–298. https://doi.org/10.1037/rel0000095

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*, 45–57. https://doi.org/10.1109/42.906424

Zhong, C.-B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*, 1451–1452. https://doi.org/10.1126/science.1130726

Zimmer, Z., Jagger, C., Chiu, C.-T., Ofstedal, M. B., Rojo, F., & Saito, Y. (2016). Spirituality, religiosity, aging and health in global perspective: A review. *SSM - Population Health*, *2*, 373–381. https://doi.org/10.1016/j.ssmph.2016.04.009

Zinnbauer, B. J., Pargament, K. I., Cole, B., Rye, M. S., Butter, E. M., Belavich, T. G., Hipp, K. M., Scott, A. B., & Kadar, J. L. (1997). Religion and spirituality: Unfuzzying the fuzzy. *Journal for the Scientific Study of Religion*, *36*(4), 549–564. https://doi.org/10.2307/1387689

Zmigrod, L., Rentfrow, P. J., Zmigrod, S., & Robbins, T. W. (2019). Cognitive flexibility and religious disbelief. *Psychological Research*, *83*, 1749–1759. https://doi.org/10.1007/s00426-018-1034-3

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*(4), 579–597. https://doi.org/10.1037/a0016399

Zuckerman, M., Silberman, J., & Hall, J. A. (2013). The relation between intelligence and religiosity: A meta-analysis and some proposed explanations. *Personality and Social Psychology Review*, *17*, 325–354. https://doi.org/10.1177/1088868313497266

# A

# Religious Displays and Perceived Trustworthiness

## A.1 Introduction

In the interest of transparency and opening up the file drawer, I will use this space to shortly discuss a final study of the cross-cultural religious replication project (CCRRP) that hasn't been written up yet.[1] In this subproject, we investigated the much-debated relation between religion and morality (e.g., Gervais, Shariff, et al., 2011; McKay & Whitehouse, 2015; Norenzayan et al., 2016; Shariff & Norenzayan, 2007). Many theories posit that religion fosters in-group cooperation, due to the influence of moralizing supernatural entities (Atkinson & Bourrat, 2011; Henrich, 2015; Norenzayan, 2013; Norenzayan et al., 2016; Purzycki et al., 2016) or communal rituals (Alcorta & Sosis, 2005; Sosis & Alcorta, 2003; Sterelny, 2018; Xygalatas, Mitkidis, et al., 2013).

A central question in the literature is: are religious individuals more moral, prosocial, generous, trustworthy? It is generally claimed that (costly) religious behaviours (e.g., wearing specific clothing, fasting etc.) can be used as a reliable signal of a person's trustworthiness and commitment to the religious community (Sosis, 2006). Indeed, the positive effect of religious signals on perceived trustworthiness and reputation has been established across multiple cultures (Power, 2017; Purzycki & Arakchaa, 2013; Soler, 2012). In addition to behavioral signals, various studies extended this effect to more subtle symbolic signals of religious commitment (Bailey & Garrou, 1983; Chia & Jih, 1994; Galen et al., 2011; McCullough et al., 2016). For instance, it has been found that participants (both religious and nonreligious) perceived a target to be more trustworthy when the target displayed a religious badge (i.e., a visual display of one's religious identity, such as specific clothing or adornments) associated with Christianity (McCullough et al., 2016). However, previous studies on the effect of religious symbolic displays have all been conducted in the US – a country where religiosity is highly socially desirable (e.g., Gervais, Shariff, et al., 2011; Kelley & de Graaf, 1997). Thus, in the CCRRP, we experimentally tested whether a (subtle) religious display enhanced trustworthiness ratings of a pictured target person across 24 countries.

[1] This project will eventually also be reported in full; any details missing here will be included in a full report in the near future.

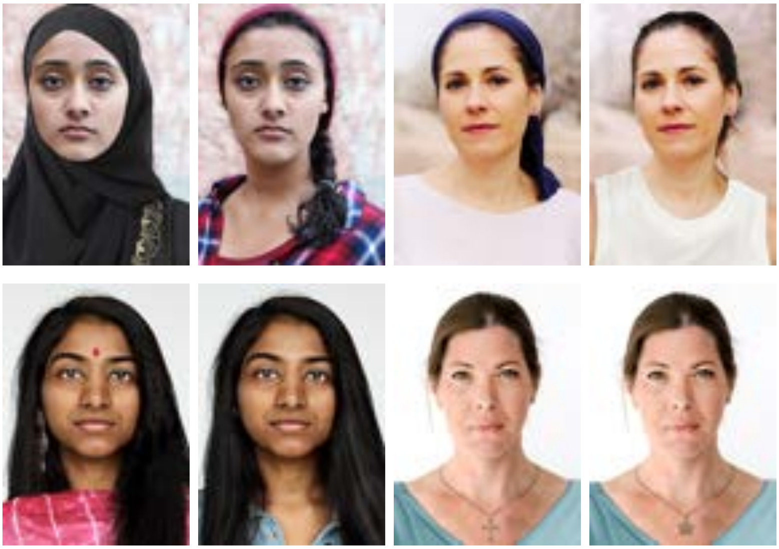


**Figure A.1:** Four sets of stimuli as used in the religious display study. For each pair, the picture on the left features the religious display and the picture on the right the control display. The first set was used in Muslim countries, the second in Israel, the third in India, and the fourth in North America and Northern Europe.

## A.2 Methods

Participants were presented with a photo of a woman with or without a religious badge (between-subjects; see Figure A.1 for four sets of example stimuli). Participants first rated the perceived religiosity of the target on a 7-point Likert scale. The main outcome variable was a trustworthiness rating, operationalized as a concrete scenario, namely the estimated probability that the target person would return a certain amount of borrowed money, i.e., for the US: "Imagine you lent this person $50, how likely do you think it is that she will give it back to you?".[2]

For the main research question on whether religious displays influence trustworthiness perceptions, we tested three hypotheses, varying in breadth of the effect. For all three hypotheses, we applied Bayes factor models comparison, comparing the predictive adequacy of the null-model (no effect), the common-effect model (effect of same size across countries), the positive-effects model (effect varying in size across countries), and the unconstrained model (effect varying in direction across countries). A noteworthy extension of these hierarchical models is presented by Haaf and Rouder (2019): in addition to the four models mentioned above, we could include the prediction that in some countries, the effect is truly zero, whereas in others it is truly positive. This type of mixture between the null-model and the positive effects model is often called a spike-and-slab model, where the spike refers to the absence of an effect (a spike at zero) and the slab refers to the distribution of positive effect sizes

[2] The amount of money was adjusted to the local currency in each county and equate to approximately 50 US dollars or somewhat lower if that amount would constitute a lot of money for an average person in a particular country.

(E. I. George & McCulloch, 1993; Mitchell & Beauchamp, 1988).[3]

## A.3 Results

Aggregating across countries and levels of participant religiosity, trustworthiness ratings were slightly higher in the religious display condition ($M = 4.79$, $SD = 1.45$, $n = 5109$) than in the control display condition ($M = 4.77$, $SD = 1.47$, $n = 5050$). Figure A.2 shows the effect of display condition on trustworthiness ratings for each level of religiosity and for each country separately.

For the manipulation check, the Bayes factor model comparison provided evidence approaching $\infty$: targets wearing a religious display were perceived as more religious than the control targets. The unstandardized size of the display condition effect on perceived religiosity is 1.43 95% credible interval [1.16, 1.66].

First, for the general comparison between religious display versus control across all subjects, we find most evidence for the null-model. This model outperforms the common-effect model by a factor of 16.80, the spike-and-slab model by a factor of 7.10, and the unconstrained model by a factor of 31.62. Second, we tested a stricter hypothesis, namely whether a religious display versus a control display increases perceived trustworthiness only for religious raters, while it does not affect perceived trustworthiness for non-religious raters. Note that we still use the full dataset for this comparison, yet a different coding of the effect (i.e., religious participants in the religious display condition vs. everyone else). For this religious-participants-only hypothesis, we obtained most evidence for the spike-and-slab model, assuming that in some countries there is no effect of religious display, whereas in others there is a positive effect. The data fit the spike-and-slab model 2,505 times better than the null-model, 71.5 times better than the common-effect model, 163 times better than the positive-effects model, and 4.07 times better than the unconstrained model. Finally, the strictest version of the hypothesis tested whether the religious display effect occurs exclusively for co-religionists (i.e., raters belonging to the same religion as the depicted target): a religious display versus a control display increases perceived trustworthiness only for raters matching the depicted religion, while it does not affect perceived trustworthiness for all other raters. For this matching-religions-only hypothesis, the data provided most evidence for the spike-and-slab model again: this model outperforms the null-model by a factor of 142,584, common-effect model by a factor of 197, the positive-effects model by a factor of 27.6, and the unconstrained model by a factor of 1.90. Using this strictest inclusion criterion, the religious display effect may be taken to occur for religious ingroups in Australia, Canada, Croatia, Ireland, Israel, Morocco, Singapore, Spain, Turkey, and the UK, though still only convincingly in Turkey. The observed and estimated effect for each country, as well as the relative Bayes factors for the models of interest are depicted in Figure A.3. The descriptives per condition for each of the three different sets of comparisons are given in Table A.1.

[3] The spike-and-slab model is typically applied in the context of variable selection, such as the selection of predictors in a regression model. Here, it is applied to selecting the "predictors" for each individual unit of the level one variable in the hierarchical model, e.g., the effect of a particular predictor in each country.

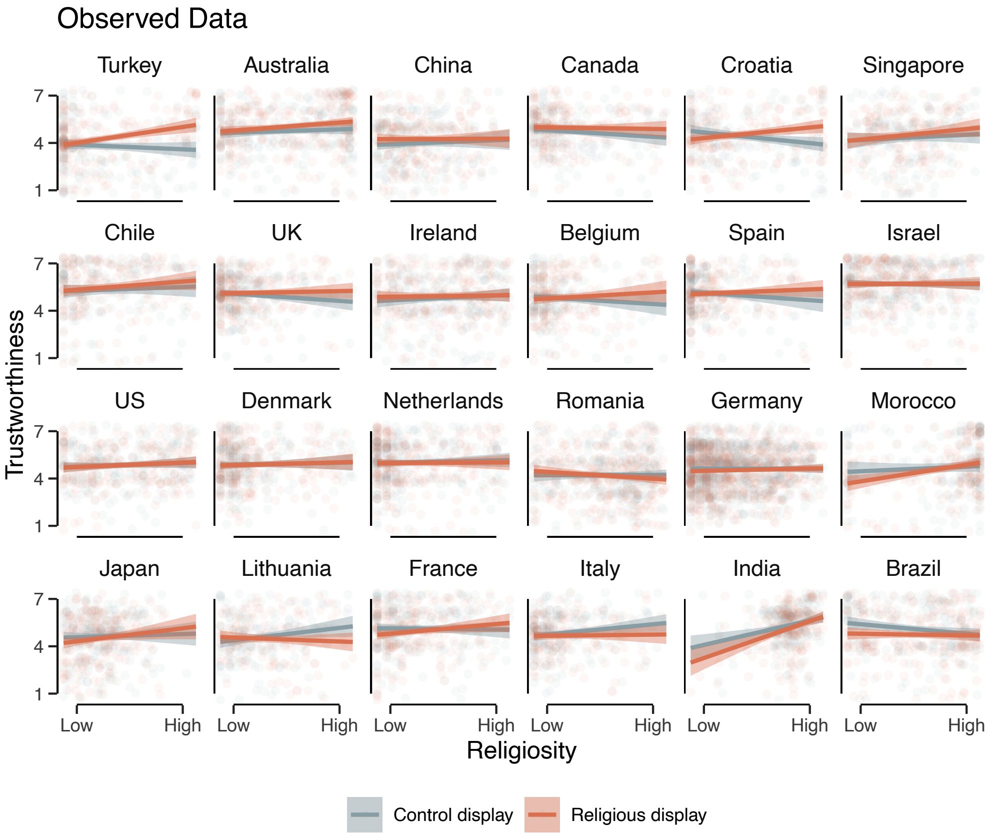


**Figure A.2:** Descriptive pattern of results per country. Countries are ordered by the overall difference in trustworthiness ratings for the religious display versus the control display condition (from left to right, top to bottom). Red lines and points denote ratings for the religious display condition and grey lines and points denote ratings for the contro display condition. The shaded bands around the lines denote the 95% confidence interval. Data points are jittered to enhance visibility. Trustworthiness was measured on a 7-point Likert scale.

**Table A.1:** Descriptives for the different sets of comparisons of the display effect

| Display effect | Condition | *M* | *SD* | *n* |
|---|---|---|---|---|
| All subjects | control | 4.77 | 1.47 | 5050 |
| | religious | 4.79 | 1.45 | 5109 |
| Religious subjects only | control | 4.76 | 1.45 | 8402 |
| | religious | 4.87 | 1.47 | 1757 |
| Religious ingroup only | control | 4.76 | 1.45 | 8955 |
| | religious | 4.91 | 1.48 | 1204 |

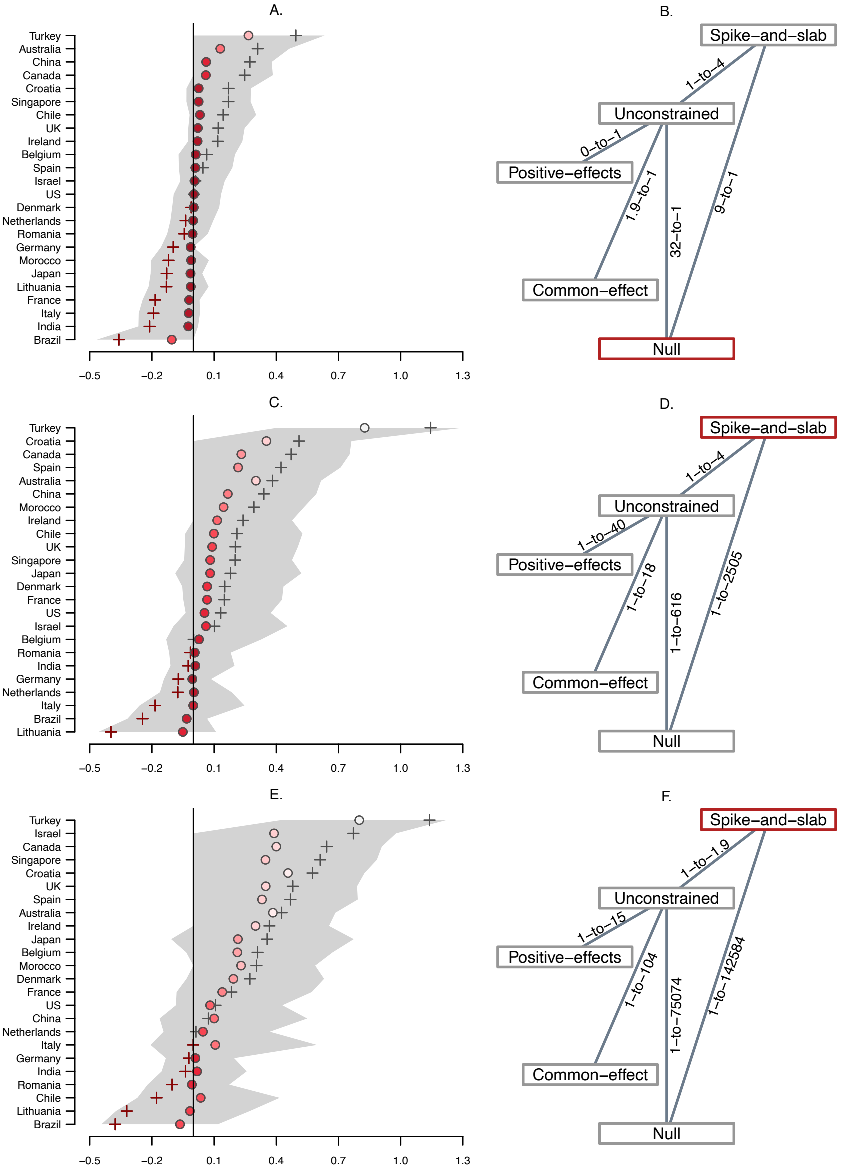


**Figure A.3:** Model estimates (on the left) and Bayesian model comparison results (on the right) for A./B. the general religious display effect; C./D. religious-participants-only religious display effect; E./F. matching-religions-only religious display effect. Left column: Crosses reflect observed effects with red crosses indicating negative effects. Points reflect model estimates with lighter shading indicating larger posterior weights of being in the slab. Right column: Bayes factors for all five models. The red frames indicate the winning model. Figure adjusted from Haaf and Rouder (2019).

# English Summary

Religion is ubiquitous; the majority of the world's population practices a religion, wars have been fought over religious disputes, and hospitals and charities have been established through religious institutions. While the perceived influence of religion seems declining in the West, globally, religion is actually on the rise. To many people, religion remains an important factor in their daily lives.

Over the last decades, scholars in the psychology and cognitive science of religion have been trying to understand the origin, the function, and the consequences of the fascinating phenomenon that religion is. In the current dissertation, we aimed to contribute to the scientific inquiry of religion, not by proposing new theories or adding hypotheses, but by rigorously testing influential existing ones. Specifically, we tried to add to previous research by reexamining previously reported effects while (1) including large and diverse samples, (2) applying open science practices and Bayesian modeling techniques (3) conducting replications of key effects plus potential alternative explanations or effect (e.g., correlational instead of experimental effects), (4) critically assessing and visualizing patterns in the raw data, and (5) applying new tools to assess robustness (e.g., a many-analysts approach, analysis blinding).

In Part I we set the stage and introduced key concepts of the replication crisis and the resulting open science movement. In Chapter 2, we translated the problems and suggested solutions from psychological science to the cognitive science of religion, with a particular focus on neuroscience, developmental research, and qualitative research. We provided a 'replication script' and a glimpse behind the scenes of the cross-cultural religious replication project (CCRRP) described in Part III. In Chapter 3, we explored the role of the intuitive plausibility of research outcomes in the context of the replication crisis. By asking laypeople to predict replication outcomes we aimed to address the question: could we have known if we had simply listened more to common sense? The study suggests that laypeople's predictions of replication outcomes contain useful information about replicability of social science studies, especially when the forecasters are unanimous in their verdict.

Part II addressed three specific replication studies. Chapter 4 reported a direct replication of *compensatory control theory* (CCT), which postulates that religion can serve as an external source of control that can substitute a perceived lack of personal control (Kay et al., 2008). We found that neither in the Netherlands, nor in the US did an experimental manipulation threatening personal control increase belief in a controlling God. However, the evidence indicated that personal control experienced in one's life is negatively related to belief in a controlling God in the US, suggesting that for some people, belief in a controlling supernatural entity might compensate for a lack of personal control. In the Netherlands, however, this negative correlation was absent.

In Chapter 5 we described an fMRI study on the relation between religiosity and behavioral and neural conflict processing, exploring the theory that religious beliefs are characterized by a lower tendency for skepticism and error monitoring. This work involves a conceptual replication of the study by Inzlicht et al. (2009). Contrary to the original study, however, we found no evidence that individual differences in religiosity were related to performance on the Stroop task as measured in accuracy and interference effects, nor to neural markers of response conflict (correct responses vs. errors) or informational conflict (congruent vs. incongruent stimuli).

In Chapter 6 we reported a Bayesian reanalysis of the Many Labs 4 replication study (R. A. Klein et al., 2019) on the mortality salience effect from Terror Management Theory (Greenberg et al., 1995; Greenberg et al., 1994). We conducted a multiverse analysis across theoretically or statistically-motivated data inclusion criteria and prior settings. The results largely converged to the conclusion that the data provide evidence against the mortality salience effect: reminders of one's own death do not seem to strengthen one's cultural identity.

In Part III, we described the results of a cross-cultural data collection effort involving 10,195 participants from 24 countries. Chapter 7 reported an experimental study on source credibility effects at play in the context of science and spirituality. We found evidence for what we call the 'Einstein effect': people tend to confer more credibility to incomprehensible claims when attributed to a scientist than when the very same claims are attributed to a spiritual guru. This Einstein effect differed for religious versus non-religious participants: individuals scoring low on religiosity considered the statement from the guru less credible than the statement from the scientist, while this difference was less pronounced for highly religious individuals.

In Chapter 8 we investigated mind-body dualism and the relation with religiosity. Following previous work, we used a vignette describing the passing of the person and subsequently inquired the continuation or cessation of bodily states (e.g., hunger) and mental states (e.g., love). We replicated previous work showing that people tend to reason dualistically as they consider mental states more likely to continue after death than bodily states. While individual religiosity was associated with both overall continuity judgments and mind-body dualism (i.e., the difference between mental and bodily states), a context manipulation emphasising religion did enhance overall continuity but not mind-body dualism. Contrary to intuitive dualism accounts, however, the pattern of results suggested that cessation rather than continuation is the default response, even for high-level mental processes.

Chapter 9 introduced the many-analysts religion project (MARP), in which we recruited 120 analysis teams to investigate the robustness of the relation between religiosity and well-being in the CCRRP data. Results on the positive association between religiosity and self-reported well-being were remarkably consistent: all but 3 teams reported positive effect sizes with credible/confidence intervals excluding zero. Somewhat more variability was observed for question 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), though a 2/3 majority of analysis teams again reported positive effect sizes with confidence/credible intervals excluding zero.

In Chapter 10, we reflected on the outcomes of the MARP and put both the answers to the research questions as well as our experiences with a many-analysts approach in a broader perspective. We addressed the issue of theoretical specificity (e.g., how to best operationalize religiosity?), highlighted some in-depth observations beyond the primary research questions (e.g., the relation between religiosity and psychological versus physical well-being, the role of objective versus subjective cultural norms of religion), considered methodological concerns (e.g., the issue of measurement invariance, treating the Likert scale data as ordinal or continuous), and discussed our experience of organizing a many-analysts project.

Chapter 11 describes the results of an experimental manipulation applied to the MARP. We assigned all analysis teams participating in the MARP to either a preregistration or an analysis blinding condition. After the teams proposed an analysis based on their assigned preparation method, we compared the teams' experiences and efficiency. We found that subjective experiences and workload are comparable between methods, but that blinding may lead to fewer deviations from the planned analysis. We discussed when each method may be most useful and argue for a combined approach, which prevents bias while retaining flexibility.

Based on the findings in this dissertation, the –perhaps unsurprising– conclusion we can draw is: some effects are and some aren't replicable. We found that religiosity indeed seems positively related to self-reported well-being; religiosity seems predictive of the tendency to make post-mortem continuity judgments of psychological states, in particular for mental states (e.g., love) compared to bodily states (e.g., hunger); religiosity seems related to credibility ratings for gobbledegook statements, and a reduced relative difference for those from a scientist compared to a spiritual guru. At the same time, we obtained convincing evidence for the absence of other effects: an experimental attenuation of personal control does not seem to activate a compensatory mechanism of belief in a controlling God; neural markers of cognitive conflict and error processing do not seem to be associated with religiosity; focusing on one's death does not seem to strengthen one's cultural identity.

Finally, in addition to the theoretical contributions, I hope this dissertation may have shown the value of replication research and of adopting (new) practices such as preregistration, analysis blinding, crowd-sourced analyses, Bayesian hierarchical modeling, and cross-cultural collaborations.

# Nederlandse Samenvatting

Religie is een wijdverbreid fenomeen; de meerderheid van de wereldbevolking hangt een religie aan, oorlogen zijn uitgevochten om religieuze conflicten, en ziekenhuizen en liefdadigheidsorganisaties zijn opgericht door religieuze instanties. Hoewel de invloed van religie in het Westen op z'n retour lijkt, neemt religiositeit wereldwijd juist toe. Voor veel mensen blijft godsdienst een belangrijke factor in hun dagelijks leven.

Gedurende de afgelopen decennia hebben wetenschappers in de psychologie van de religie geprobeerd antwoorden te vinden op vragen over de oorsprong, de functie en de gevolgen van dit fascinerende fenomeen. In dit proefschrift hebben we geprobeerd een bijdrage te leveren aan het wetenschappelijk onderzoek naar religie, niet door nieuwe theorieën of hypotheses op te werpen, maar door invloedrijke bestaande theorieën grondig te testen. Het doel was om bij te dragen aan voorgaand onderzoek door eerdere bevindingen opnieuw onder de loep te nemen en daarbij (1) grote en gevarieerde steekproeven te gebruiken, (2) *open science* principes en Bayesiaanse modellen toe te passen, (3) replicatiestudies uit te voeren van belangrijke effecten plus mogelijke alternatieve verklaringen of effecten (bijv. correlationele in plaats van experimentele effecten), (4) patronen in de ruwe data kritisch te beoordelen en te visualiseren, en (5) nieuwe technieken toe te passen om de robuustheid van resultaten te beoordelen (bijv. een 'meerdere data-analisten' methode of 'geblindeerde analyses').

In deel I schetsten we de achtergrond en introduceerden sleutelbegrippen uit de replicatiecrisis en de daaruit voortvloeiende *open science* beweging. In hoofdstuk 2 vertaalden we de problemen en de voorgestelde oplossingen uit de psychologie naar de cognitieve wetenschap van religie, met een specifieke focus op neurowetenschap, ontwikkelingsonderzoek en kwalitatief onderzoek. We hebben een 'replicatiescript' voorgesteld en een blik achter de schermen gegeven van het cross-culturele religieuze replicatieproject (CCRRP) dat in deel III werd beschreven. In hoofdstuk 3 onderzochten we de rol van de intuïtieve plausibiliteit van onderzoeksresultaten in de context van de replicatiecrisis. Door leken te vragen replicatieresultaten te voorspellen, probeerden we antwoord te vinden op de vraag: hadden we het kunnen weten als we simpelweg meer naar ons boerenverstand hadden geluisterd? Het onderzoek suggereert dat de voorspellingen van replicatie-uitkomsten door leken nuttige informatie bevatten over de repliceerbaarheid van sociaalwetenschappelijke studies, vooral wanneer de voorspellers unaniem zijn in hun oordeel.

Deel II behandelde drie specifieke replicatiestudies. In hoofdstuk 4 werd een directe replicatie gerapporteerd van de *compensatoire controle theorie* (CCT), die stelt dat religie kan dienen als een externe bron van controle die een subjectief ervaren gebrek aan persoonlijke controle kan vervangen. De resultaten lieten zien dat noch in Nederland, noch in de VS een experimentele manipulatie die persoonlijke controle bedreigde, het geloof in een controlerende God deed toenemen. Wel bleek in de VS dat ervaren persoonlijke controle negatief samenhangt met geloof in een control-

S

erende God, wat suggereert dat voor sommige mensen het geloof in een controlerende bovennatuurlijke entiteit een gebrek aan persoonlijke controle kan compenseren. In Nederland was deze negatieve correlatie echter afwezig.

In hoofdstuk 5 beschreven we een fMRI studie naar de relatie tussen religiositeit en gedrags- en neurale conflictverwerking, waarbij we de theorie onderzochten dat religieuze overtuigingen worden gekenmerkt door een verminderde neiging tot scepsis en het monitoren van onjuistheden. Dit werk betreft een conceptuele replicatie van de studie van Inzlicht et al. (2009). In tegenstelling tot de oorspronkelijke studie vonden wij echter geen bewijs dat individuele verschillen in religiositeit samenhingen met prestaties op de *Stroop* taak, noch met neurale markers van responsconflict (correcte reacties vs. fouten) of informatieconflict (congruente vs. incongruente stimuli).

In hoofdstuk 6 beschreven we een Bayesiaanse heranalyse van de Many Labs 4 replicatiestudie (R. A. Klein et al., 2019) naar het *sterfelijkheid-saillantie effect* uit Terror Management Theory (Greenberg et al., 1995; Greenberg et al., 1994). We voerden een 'multiversum-analyse' uit over theoretisch of statistisch gemotiveerde data-inclusiecriteria en *prior* instellingen. De resultaten leidden grotendeels tot de conclusie dat de data bewijs leveren tegen het sterfelijkheid-saillantie effect: herinneringen aan de je sterfelijkheid lijken de je eigen culturele identiteit niet te versterken.

In deel III beschreven we de resultaten van een cross-culturele dataverzameling waarbij 10,195 deelnemers uit 24 landen betrokken waren. In hoofdstuk 7 rapporteerden we een experimentele studie over de effecten van de geloofwaardigheid van bronnen in de context van wetenschap en spiritualiteit. We vonden bewijs voor wat we het 'Einstein effect' noemen: mensen zijn geneigd meer geloofwaardigheid toe te kennen aan onbegrijpelijke beweringen wanneer die worden toegeschreven aan een wetenschapper dan wanneer diezelfde beweringen worden toegeschreven aan een spirituele goeroe. Dit Einstein effect verschilde voor religieuze en niet-religieuze deelnemers: mensen die laag scoorden op religiositeit vonden de bewering van de goeroe minder geloofwaardig dan de bewering van de wetenschapper, terwijl dit verschil minder sterk aanwezig was voor zeer religieuze mensen.

In hoofdstuk 8 onderzochten we lichaam-geest dualisme en de relatie met religiositeit. In navolging van eerder onderzoek gebruikten we een scenario over een overleden persoon en vroegen vervolgens naar het wel of niet voortbestaan van lichamelijke condities (bijv. honger) en mentale condities (bijv. liefde). We repliceerden eerder werk dat aantoonde dat mensen geneigd zijn dualistisch te redeneren, aangezien ze vaker aangeven dat mentale processen zullen voortduren na de dood dan lichamelijke processen. Individuele religiositeit was zowel geassocieerd met een sterkere neiging te oordelen dat processen voortbestaan in het algemeen, als met lichaam-geest dualisme in het bijzonder (d.w.z. het verschil tussen mentale en lichamelijke processen). Een context-manipulatie waarbij een religieuze benadering van de dood werd benadrukt, versterkte wel de neiging om algemeen voortbestaan van processen te indiceren, maar niet specifiek lichaam-geest dualisme. In tegenstelling tot de theorie van 'intuïtief dualisme' suggereert het patroon van resultaten dat het standaardoordeel vergaan in plaats van voortbestaan is, zelfs voor hogere mentale processen.

In hoofdstuk 9 werd het meerdere-data-analisten religie project (MARP) geïntroduceerd, waarin we 120 analyseteams inschakelden om de robuustheid van de relatie tussen religiositeit en welzijn in de CCRRP-data te onderzoeken. De resultaten over de positieve associatie tussen religiositeit en zelf-gerapporteerd welzijn waren opmerkelijk consistent: op 3 na rapporteerden alle teams positieve effectgroottes waarbij nul buiten het geloofwaardigheids-/ betrouwbaarheidsinterval viel. Iets meer variabiliteit werd waargenomen voor de vraag of de relatie tussen religiositeit en zelf-gerapporteerd 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), hoewel een $^2/_3$ meerderheid van de analyseteams opnieuw positieve effectgroottes rapporteerden met geloofwaardigheids/betrouwbaarheidsinterval dat nul uitsluit.

In hoofdstuk 10 reflecteerden we op de uitkomsten van het MARP en plaatsten zowel de antwoorden op de onderzoeksvragen als onze ervaringen met een meerdere-data-analisten-aanpak in een breder perspectief. We besproken de kwestie van theoretische specificiteit (bijv., hoe kan religiositeit het best worden geoperationaliseerd?), belichtten enkele genuanceerde observaties die verder gingen dan de primaire onderzoeksvragen (bijv., de relatie tussen religiositeit en psychologisch versus lichamelijk welzijn, de rol van objectieve versus subjectieve culturele normen van religie), behandelden methodologische kwesties (bijv., de kwestie van meetinvariantie, het behandelen van de Likert-schaaldata als ordinaal of continu), en bespraken onze ervaringen met het organiseren van een meerdere-data-analisten-project.

In hoofdstuk 11 worden de resultaten beschreven van een experimentele manipulatie die in het MARP is toegepast. Alle analyseteams die aan het MARP deelnamen werden toegewezen aan een preregistratie- of een geblindeerde analyse conditie. Nadat de teams een analyse hadden voorgesteld op basis van de hun toegewezen voorbereidingsmethode, zijn de ervaringen en efficiëntie van de teams vergeleken. We vonden dat subjectieve ervaringen en werklast vergelijkbaar zijn tussen de methoden, maar dat analyseblindering kan leiden tot minder vaak hoeven afwijken van de geplande analyse. We bespraken wanneer elk van beide methodes het nuttigst kan zijn en pleitten voor een gecombineerde aanpak, die vooringenomenheid voorkomt en toch flexibiliteit behoudt.

Op basis van de bevindingen in dit proefschrift is de –wellicht weinig verrassende– conclusie die we kunnen trekken: sommige effecten zijn wel en andere zijn niet repliceerbaar. We vonden dat religiositeit inderdaad positief gerelateerd lijkt aan zelf-gerapporteerd welzijn; religiositeit lijkt voorspellend voor de neiging om te oordelen dat psychologische processen post-mortem voortbestaan, in het bijzonder mentale processen (bijv. liefde) in vergelijking met lichamelijke processen (bijv. honger); religiositeit lijkt gerelateerd aan de geloofwaardigheid van onzinuitspraken, en een verminderd relatief verschil voor die afkomstig van een wetenschapper in vergelijking met een spirituele goeroe. Tegelijkertijd hebben we overtuigend bewijs verkregen voor de afwezigheid van andere effecten: een experimentele afname van persoonlijke controle lijkt geen compenserend mechanisme van geloof in een controlerende God te activeren; neurale markers van cognitief conflict en foutverwerking lijken niet geassocieerd te zijn met religiositeit; focus op je eigen dood lijkt je culturele identiteit niet te versterken.

Tenslotte hoop ik dat dit proefschrift, naast de theoretische bijdragen, de waarde heeft aangetoond van replicatieonderzoek en van het toepassen van (nieuwe) onderzoekspraktijken zoals preregistratie, analyseblindering, *crowd-sourced* analyses, Bayesiaanse hiërarchische modellering, en cross-culturele samenwerkingsverbanden.

*You'll never walk alone.*

Richard Rodgers & Oscar Hammerstein II<br>
Lee Towers

# Contributions

Contributorship was documented with CRediT taxonomy using tenzing (Holcombe et al., 2020).

Chapter 2:

**Conceptualization:** S.H. and M.v.E.
**Funding acquisition:** M.v.E.
**Supervision:** M.v.E.
**Writing - original draft:** S.H.
**Writing - review & editing:** M.v.E.

Chapter 3:

**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. and A.S.
**Methodology:** S.H., A.S., and E.-J.W.
**Supervision:** E.-J.W.
**Visualization:** S.H. and A.S.
**Writing - Original Draft:** S.H. and A.S.
**Writing - Review & Editing:** E.-J.W.

We thank Job Bank, Chiel van 't Hoff, and Philip Kleinbrahm for their support in creating the study materials and collecting the data.

Chapter 4:

**Conceptualization:** S.H., E.-J.W., A.C.K., and M.v.E.
**Data curation:** S.H.
**Formal analysis:** S.H.
**Funding acquisition:** M.v.E.
**Investigation:** S.H.
**Methodology:** S.H.
**Project administration:** M.v.E.
**Supervision:** E.-J.W. and M.v.E.
**Visualization:** S.H.
**Writing - original draft:** S.H.
**Writing - review & editing:** E.-J.W., A.C.K., and M.v.E.

Chapter 5:

**Conceptualization:** S.H. and M.v.E.
**Data curation:** S.H. and L.S.
**Formal analysis:** S.H. and L.S.
**Funding acquisition:** M.v.E.
**Investigation:** S.H. and L.S.
**Methodology:** S.H. and L.S.
**Project administration:** M.v.E.
**Supervision:** M.v.E.
**Visualization:** S.H. and L.S.
**Writing - original draft:** S.H. and L.S.
**Writing - review & editing:** M.v.E.

Chapter 6:

**Conceptualization:** J.M.H, S.H. and E.-J.W.
**Formal analysis:** J.M.H, S.H. and S.B.
**Methodology:** J.M.H, S.H., S.B and Q.F.G.
**Supervision:** J.M.H. and E.-J.W.
**Visualization:** J.M.H, S.H. and S.B.
**Writing - original draft:** J.M.H.
**Writing - review & editing:** S.H., S.B., Q.F.G and E.-J.W.

Chapter 7:

**Conceptualization:** S.H. and M.v.E.
**Data curation:** S.H.
**Formal analysis:** S.H., J.M.H., and J.A.B.
**Funding acquisition:** R.M.R., R.M., S.A., N.L., and M.v.E.
**Investigation:** S.H., R.M.R., R.M., S.A., T.B., R.B., A.C., C.G., R.G., K.H., C.K., A.N., L.Q., A.R., J.E.R., H.T., F.U., R.W., and D.X.
**Methodology:** S.H.
**Project administration:** S.H. and M.v.E.
**Supervision:** M.v.E.
**Visualization:** S.H.
**Writing - original draft:** S.H.
**Writing - review & editing:** J.M.H., J.A.B., R.M.R., R.M., S.A., T.B., R.B., A.C., C.G., R.G., W.M.G., K.H., C.K., N.L., A.N., L.Q., A.R., J.E.R., B.T.R., H.T., F.U., R.W., D.X., and M.v.E.

Chapter 8:

**Conceptualization:** S.H. and M.v.E.
**Data curation:** S.H.
**Formal analysis:** S.H., J.A.B. and J.M.H.
**Funding acquisition:** S.A., N.L., R.M., R.M.R., and M.v.E.
**Investigation:** S.H., S.A., T.B., R.B., A.C., C.G., R.G., K.H., C.K., R.M., A.N., L.Q., A.R., J.E.R., R.M.R., H.T., R.W., and D.X.

**Methodology:** S.H.
**Project administration:** S.H. and M.v.E.
**Supervision:** M.v.E.
**Visualization:** S.H.
**Writing - original draft:** S.H.
**Writing - review & editing:** S.A., T.B., R.B., J.A.B., A.C., C.G., R.G., J.M.H., K.H., C.K., N.L., R.M., A.N., L.Q., A.R., J.E.R., R.M.R., H.T., R.W., D.X., and M.v.E.

Chapter 9:

Author List Suzanne Hoogeveen, Alexandra Sarafoglou, Balazs Aczel, Yonathan Aditya, Alexandra J. Alayan, Peter J Allen, Sacha Altay, Shilaan Alzahawi, Yulmaida Amir, Francis-Vincent Anthony, Obed Kwame Appiah, Quentin D. Atkinson, Adam Baimel, Merve Balkaya-Ince, Michela Balsamo, Sachin Banker, František Bartoš, Mario Becerra, Bertrand Beffara, Julia Beitner, Theiss Bendixen, Jana B. Berkessel, Renatas Berniūnas, Matthew I. Billet, Joseph Billingsley, Tiago Bortolini, Heiko Breitsohl, Amélie Bret, Faith L Brown, Jennifer Brown, Claudia C. Brumbaugh, Jacek Buczny, Joseph Bulbulia, Saúl Caballero, Leonardo Carlucci, Cheryl L. Carmichael, Marco E. G. V. Cattaneo, Sarah J. Charles, Scott Claessens, Maxinne C. Panagopoulos, Angelo Brandelli Costa, Damien L. Crone, Stefan Czoschke, Christian Czymara, E. Damiano D'Urso, Örjan Dahlström, Anna Dalla Rosa, Henrik Danielsson, Jill De Ron, Ymkje Anna de Vries, Kristy K. Dean, Bryan J. Dik, David J. Disabato, Jaclyn K. Doherty, Tim Draws, Lucas Drouhot, Marin Dujmovic, Yarrow Dunham, Tobias Ebert, Peter A. Edelsbrunner, Anita Eerland, Christian T. Elbaek, Shole Farahmand, Hooman Farahmand, Miguel Farias, Abrey A. Feliccia, Kyle Fischer, Ronald Fischer, Donna Fisher-Thompson, Zoë Francis, Susanne Frick, Lisa K. Frisch, Diogo Geraldes, Emily Gerdin, Linda Geven, Omid Ghasemi, Erwin Gielens, Vukašin Gligorić, Kristin Hagel, Nandor Hajdu, Hannah R. Hamilton, Imaduddin Hamzah, Paul H. P. Hanel, Christopher E. Hawk, Karel K. Himawan, Benjamin C. Holding, Lina E. Homman, Moritz Ingendahl, Hilla Inkilä, Mary L. Inman, Chris-Gabriel Islam, Ozan Isler, David Izydorczyk, Bastian Jaeger, Kathryn A. Johnson, Jonathan Jong, Johannes A. Karl, Erikson Kaszubowski, Benjamin A. Katz, Lucas A Keefer, Stijn Kelchtermans, John M. Kelly, Richard A. Klein, Bennett Kleinberg, Megan L. Knowles, Marta Kołczyńska, Dave Koller, Julia Krasko, Sarah Kritzler, Angelos-Miltiadis Krypotos, Thanos Kyritsis, Todd L. Landes, Ruben Laukenmann, Guy A. Lavender Forsyth, Aryeh Lazar, Barbara J. Lehman, Neil Levy, Ronda F. Lo, Paul Lodder, Jennifer Lorenz, Paweł Łowicki, Albert L. Ly, Esther Maassen, Gina M Magyar-Russell, Maximilian Maier, Dylan R. Marsh, Nuria Martinez, Marcellin Martinie, Ihan Martoyo, Susan E. Mason, Anne Lundahl Mauritsen, Phil McAleer, Thomas McCauley, Michael McCullough, Ryan McKay, Camilla M. McMahon, Amelia A. McNamara, Kira K. Means, Brett Mercier, Panagiotis Mitkidis, Benoît Monin, Jordan W. Moon, David Moreau, Jonathan Morgan, James Murphy, George Muscatt, Christof Nägel, Tamás Nagy, Ladislas Nalborczyk, Gustav Nilsonne, Pamina Noack, Ara Norenzayan, Michèle B. Nuijten, Anton Olsson-Collentine, Lluis Oviedo, Yuri G. Pavlov, James O. Pawelski, Hannah I. Pearson, Hugo Pedder, Hannah K. Peetz, Michael Pinus, Steven Pirutinsky, Vince Polito, Michaela Porubanova, Michael J. Poulin, Jason M Prenoveau, Mark A. Prince, John Protzko, Campbell Pryor, Benjamin G. Purzycki, Lin Qiu, Julian Quevedo Pütter, André Rabelo, Milen L. Radell, Jonathan E. Ramsay, Graham Reid, Andrew J. Roberts, Lindsey M. Root Luna, Robert M. Ross, Piotr Roszak, Nirmal Roy, Suvi-Maria K. Saarelainen, Joni Y. Sasaki, Catherine Schaumans, Bruno Schivinski, Marcel C. Schmitt, Sarah A. Schnitker, Martin Schnuerch, Marcel R. Schreiner, Victoria Schüttengruber, Simone Sebben, Suzanne C. Segerstrom, Berenika Seryczyńska, Uffe Shjoedt, Müge Simsek, Willem W. A. Sleegers, Eliot R. Smith, Walter J. Sowden, Marion Späth, Christoph Spörlein, William Stedden, Andrea H. Stoevenbelt, Simon Stuber, Justin Sulik, Christiany Suwartono, Stylianos Syropoulos, Barnabas Szaszi, Peter Szecsi, Ben M. Tappin, Louis Tay, Robert T. Thibault, Burt Thompson, Christian M. Thurn, Josefa Torralba, Shelby D. Tuthill, Ann-Marie Ullein, Robbie C. M. Van Aert, Marcel A.L.M. van Assen, Patty Van Cappellen, Olmo R. Van den Akker, Ine Van der Cruyssen, Jolanda Van der Noll, Noah N. N. van Dongen, Caspar J. van Lissa, Valerie van Mulukom, Don van Ravenzwaaij, Casper J. J. van Zyl, Leigh Ann Vaughn, Bojana Većkalov, Bruno Verschuere, Michelangelo Vianello, Felipe Vilanova, Allon Vishkin, Vera Vogel, Leonie V.D.E. Vogelsmeier, Shoko Watanabe, Cindel J. M. White, Kristina Wiebels, Sera Wiechert, Zachary Z. Willett, Maciej Witkowiak, Charlotte V. O. Witvliet, Dylan Wiwad, Robin Wuyts, Dimitris Xygalatas, Xin Yang, Darren J. Yeo, Onurcan Yilmaz, Natalia Zarzeczna, Yitong Zhao, Josjan Zijlmans, Michiel van Elk, Eric-Jan Wagenmakers.

**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., B.M.T., 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 10:

**Conceptualization:** S.H., A.S., M.v.E., and E.-J.W.
**Supervision:** M.v.E. and E.-J.W.
**Writing - original draft:** S.H., and A.S.
**Writing - review & editing:** M.v.E., and E.-J.W.

Chapter 11:

**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:** 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.

*It takes a village to raise a thesis.*

A grateful PhD-candidate

# Dankwoord

Vijf jaar later, 335 pagina's verder, en ontelbare ervaringen rijker en nu is het opeens klaar. Het voelt als een kleine opluchting maar ook een soort weemoedig afscheid. Want ondanks de sporadische momenten van stress en frustratie heb ik ontzettend genoten van dit PhD-avontuur. En dat dit traject een geweldige ervaring was heb ik grotendeels aan een aantal fantastische mensen te danken. In de eerste plaats natuurlijk mijn geweldige begeleiders. EJ en Michiel, ik wil jullie ontzettend bedanken voor de motivatie, inspiratie, jullie betrokkenheid en tegelijkertijd de vrijheid die ik tijdens dit traject heb ervaren. Ik heb me de afgelopen 5 jaar enorm bevoorrecht gevoeld met twee zulke diverse begeleiders wat betreft expertise en supervisiestijl maar met een gedeelde passie voor Goede Wetenschap. Ik had me geen betere en vruchtbaardere combinatie van promotoren voor kunnen stellen!

EJ, ik bewonder jouw passie voor wetenschap, je optimisme, en je enorme drive om de wetenschap vooruit te helpen. Ik vond het geweldig hoe je altijd weer hetzelfde enthousiasme toonde wanneer ik je soms opnieuw moest bijpraten over lopende projecten. Ik heb ontzettend veel van je geleerd, van de oneindige superieuriteit van Bayesiaanse statistiek tot de Oxford-komma en het belang van mooie data-plotjes. Bedankt dat ik als groentje toch lid mocht worden van jouw *Bayesian family* (aka Bayesian army). Daarnaast ook bedankt voor de grote levenslessen: er gaat niks boven een pistoletje kipkerrie-salade en vakanties met kinderen zijn vermoeiender dan werken (dat laatste begin ik steeds meer in te zien).

Michiel, dankzij een leuke stage bij jou tijdens mijn master, ontdekte ik de fascinerende "kracht van geloven" en uiteindelijk mondde onze geslaagde samenwerking uit in dit AiO-traject. Ik ben ontzettend dankbaar voor de kans die je me bood en de prettige begeleiding die volgde. Ik waardeer je altijd snelle en nuttige feedback, ondersteuning waar nodig en vrijheid waar mogelijk. Ik herinner me hoe je tijdens een lab-lunch in mijn eerste maand als AiO ons op het hart drukte ook een leven naast het werk te behouden. Ik vond dat toen een enigszins onverwacht advies, maar ik ben het tijdens mijn PhD het belang van dat advies steeds meer gaan inzien, evenals het belang van geven van zo'n boodschap als begeleider. Jouw eerlijkheid en goede balans tussen nieuwsgierigheid, passie, nuchterheid en een kritische blik zijn mij tijdens dit PhD-avontuur blijven inspireren. Ik ben blij te zien hoe succesvol jij bent in het volgen van je passies, afwijken van de gebaande paden en de verwachtingen van je wetenschappelijke omgeving. Ik vind het heel gaaf hoe je nu weer een nieuwe weg in slaat van de psychedelica, ik blijf je uiteraard volgen. Daarnaast bedankt voor de gezelligheid tijdens labuitjes en mijn introductie in de wereld van de *cognitive science of religion* tijdens de workshop in het klooster op Sicilië.

I'd also like to thank my excellent committee members: Frenk, Han, Ryan, Julia, and Olga. I feel honoured to have my thesis evaluated by such an inspiring group of scientists; thank you very much for finding the time in your busy lives to read this monstrously thick book. I hope I get to work with each of you (again) at some point!

I also want to express my gratitude to all the inspiring colleagues at the psychology department. Yes, imposter syndrome is (still) very real, but through the years I've enormously valued all informal chats in the pantry, brownbag talks and lab meetings. I've had the luxury of having been part of two program groups at the psychology department, which has greatly enriched my scientific knowledge and views, and also proved a fascinating source of observing cultural differences (which is obviously interesting; we're psychologists after all). I still haven't figured out where I truly belong – I'd rather not choose at all – but I've felt at home in both groups; thanks a lot to everyone for that.

The methods crew: Quentin, Don, Angelika, Frantisek, thanks for all the fun lunches we had over the years. The discussions on coin tossing records, chicken-diets, and politically incorrect remarks never cease to amaze me (read: entertain and infuriate at the same time). Julia, thanks so much for supervising me in my current post-doc. I very much enjoy working together, and I'm grateful that you also push me when necessary (=often), show how to kick ass in this (still) somewhat male-dominated culture, and just are an awesome female role-model in science. I look forward to more collaborations and toddler playdates!

Alexandra, I'm so glad we ended up working together on the laypeople project, which proved the beginning of a number of fruitful and –importantly– always fun collaborations. Our projects together were often the most fun to work on; I've highly enjoyed our bitching and swearing, our MARP-hustling skills, our collective brightest moments (sqrt(-2) and see footnote 2 in Chapter 3), and our fancy coffee + chocolate feasts. I can hardly believe it somehow resulted in scientific output, but the process was always a lot of fun. Let's keep up with this tradition! And I'm glad to have you as a paranymph at my side during the defence (in this case, I could actually toss some questions to you, so be prepared!).

My SP-family and roomies on the second floor: Maria, Kunalan, Tiarah, YongQi, Aidan, Enzo; my dear office bitches, thank you so much for brightening my life with some gossip, cute dog-pictures, endless coffee breaks, and some well-dosed scientific discussions from time to time. Although we were all kind of doing our own thing, I'm grateful for the relaxed, helpful, and friendly vibe in the office and the opportunity to have an uplifting chat whenever anyone got stuck on stupid R code or annoying email. I miss you guys!

Lisanne, na onze ontmoeting op het vliegveld in Atlanta en wandelingen over het kerkhof was onze vriendschap wat mij betreft een feit. Bedankt dat je me meegenomen hebt in de wereld van de sociale psychologie, zowel op sociaal als inhoudelijk gebied. Ik vond het heel leuk dat we ondanks onze verschillende interesses en expertise toch nog een keer hebben kunnen samenwerken aan een artikel! En bovenal ben ik heel dankbaar voor de gezelligheid en goede gesprekken over de grote en kleine zaken des levens. (En nu je Duitsland weer ingeruild hebt voor het prachtige Utrecht wordt het weer hoog tijd voor koffiedates en eindeloos lullen over van alles en nog wat!)

Lieve kibbelingen: Emma, Rosa, Yvette, Frederica, Sterre, Anne, Julia, Vera. Ik ben enorm dankbaar dat ik naast het wetenschappelijke geneuzel bij jullie gewoon lekker Suus kan zijn. Het doet mij enorm goed bij jullie te kunnen kletsen over alles behalve werk; etentjes, vakanties, wijntjes, sporten, spelletjes, het is altijd dikke lol. Ik hoop dat we onze tradities van maandelijkse diners en jaarlijkse weekendjes weg ondanks drukke banen, gillende kids en veeleisende mannen vast kunnen houden. Ik voel me vereerd dat jullie me als Schiedammer hebben opgenomen in jullie Utrecht-gang. Bedankt voor jullie liefde, lol en interesse (ondanks mijn terughoudendheid om inhoudelijk over m'n werk te praten; booooring). Roos en Veer, ik ben blij dat we nog steeds een studieclub app hebben, zo blijven we ons toch nog een beetje jong voelen ondanks hypotheken en baby's (sorry Roos!).

Eli, na onze succesvolle samenwerking tijdens de minor en bachelorscriptie volgde een tijd van gezamenlijke activiteiten: samen beginnen aan dezelfde master (leuke tripjes naar het NIN!!), samenwonen op de Croesie, *presentation*-scripts maken, spelletjesavonden, kooksessies en kennismaking met de wonderen van de academische wereld. Hoewel jij er na een paar jaar wel weer klaar mee was, heb ik toch veel gehad aan onze gezamenlijk opdrachten en discussies over wetenschap. Heel fijn hoe ik met jou goed over die gekke wetenschappelijke wereld kan praten maar daarna vooral ook over de echte belangrijke dingen in het leven (bier, mannen, films, spelletjes). Sas, Lena, en Linds, hoewel we elkaar niet heel vaak meer zien, denk ik met plezier terug aan onze (post)studententijd als roomies. Altijd fun op de Croese!

Lukas, Marc, Liz, Celine: de psychologiecrew (+ASW, I know, Lukas): als ik terug denk aan m'n bachelor zijn de beste herinneringen vooral van onze gezamenlijke (niet-studie-gerelateerde) avonturen; kameel-tochten in Tunesië, wodkaproeverijen om 11u 's morgens in Sint-Petersburg, kuttiën, weerwolven en andere niet nader te benoemen spellen. Bedankt voor alle fun in het Langeveldgebouw en ver daarbuiten.

Sophie en Juud: ik denk dat tijdens het maken van onze Oscar-waardige postmoderne profielwerkstukfilm (Quentin Tarantino. Punt.) het zaadje voor onze ambitie geplant is. Juud, ik ben blij dat jij de culturele talenten voor ons alledrie glansrijk vervult. Sop, ik ben trots dat we onze *nerdheid* lang genoeg hebben kunnen vasthouden om nu allebei onze PhD behaald te hebben (meneer Tromp-Meesters zou trots op ons zijn!). Hoewel we de laatste jaren vooral via onze moeders van elkaars voortgang op de hoogte werden gehouden, ben ik ontzettend blij dat we elkaar nu toch zo mogen bijstaan tijdens de laatste loodjes.

Jits & Juul: wij starten ons psychologie-avontuur in dezelfde werkgroep, werden huisgenoten en vriendinnen voor het leven. Douchetreinen, verbroken Homeland pacts, biertjes in Otje en dansjes in de Woo (my favorites!!), vakanties, wat hebben we niet samen overleefd? De gesprekken en dilemma's zijn gedurende de jaren wat veranderd; van wel of niet naar college om 9 uur, wel of niet de 12e aflevering van Friends kijken op een avond, wijn van 2 of 3 euro bij de Appie, naar wel of geen koophuis, babyshowers en borstvoedingsperikelen. Van slapeloze nachten door bier, dansjes en onbereikbare mannen naar slapeloze nachten door hongerige baby's en behoeftige dreumesen. Ik ben ontzettend blij en dankbaar dat we al deze fases naar volwassenheid (haha, who are we kidding) samen hebben mogen doorlopen. Op naar nog vele dates met en zonder koters!

Gi & Caroline, Ies en Veer, ik kan me geen leukere schoonfamilie voorstellen. Ik ben ontzettend blij dat ik deel mag uitmaken van jullie gezin. Lena mag maar in haar handjes knijpen met zo'n lieve opa en oma en oom en tante! Zet de klompjes en het kieltje maar vast klaar, want wij komen ongetwijfeld nog vaak uitwaaien op de boerderij.

Lieve papa en mama, bedankt voor jullie vanzelfsprekende liefde en betrokkenheid. Ik prijs mezelf enorm gelukkig om op te zijn gegroeid in zo'n fijn gezin als dat van ons. Jullie hebben me alle kansen gegeven en ik heb me altijd gesteund gevoeld in m'n keuzes (of het gebrek daaraan). Ondanks dat het antwoord vaak niet verder kwam dan "het gaat gewoon over wetenschappelijke dingen" bleven jullie geïnteresseerd in mijn proefschrift. Maar veel belangrijker zijn jullie adviezen en steun op alle andere gebieden – sorry, zelfs met een PhD op zak blijven jullie m'n raadgevers voor Grote Mensenzaken; belastingen, verzekeringen, kledingadvies, opvoedtips voor alles kan ik bij jullie terecht. En nóg veel belangrijker zijn natuurlijk de etentjes, weekendjes, en vakanties die we nog steeds met elkaar doorbrengen en jullie grenzeloze liefde en aandacht voor ons en Lena. Ik gun iedereen zo'n fijne jeugd en lieve ouders als die van mij. Sjo en Wout, opeens zijn we alledrie (semi-)ambtenaren, wie had dat gedacht! De 35 beloofde weekendjes Winterberg zullen er waarschijnlijk niet van komen, maar ik ben ontzettend blij dat we allemaal nog genieten van eindeloze potjes Boonanza, Kolonisten, en Party&Co met bitterballen en stinkkazen tijdens vakanties in de Ardennen en Limburg. De vanzelfsprekendheid waarop we met z'n allen kunnen niksen en lachen is me enorm veel waard. Janneke en Linde, ik ben heel blij dat jullie onze familie (en spelteam) zijn komen versterken. En eindeloze dank voor jullie bijdrage aan het animatieteam; ik kom zowaar tot rust de laatste vakanties.

Lieve Emiel, bedankt voor je onvoorwaardelijke steun, vertrouwen en liefde. Ik ben zo blij dat ik na een lange dag vol gecrashte R-scripts en vruchteloze schrijfsessies altijd weer met jou op de bank kan kruipen. Ik heb enorm veel zin in alles wat nog voor ons ligt en wat we samen gaan beleven. En sinds jij er bent, lieve Lena, is de wereld nóg een stukje mooier.

# List of Other Publications

van der Miesen, M. M., van der Lande, G. J. M., **Hoogeveen, S.**, Schjoedt, U., & van Elk, M. (2022). The effect of source credibility on the evaluation of statements in a spiritual and scientific context: A registered report study. *Comprehensive Results in Social Psychology.* https://doi.org/10.1080/23743603.2022.2041984

Pauw, L. S., **Hoogeveen, S.**, Breitenstein, C. J., Meier, F., Rauch-Anderegg, V., Neysari, M., Martin, M., Bodenmann, G., & Milek, A. (2021). Spillover effects when taking turns in dyadic coping: How lingering negative affect and perceived partner responsiveness shape subsequent support provision. *Frontiers in Psychology, 12.* https://www.frontiersin.org/article/10.3389/fpsyg.2021.637534

Tierney, W., Hardy, J. H., III, Ebersole, C. R., Viganola, D., Clemente, E., Gordon, E., **Hoogeveen, S.**, Haaf, J. M., Dreber, A. A., Johannesson, M., Pfeiffer, T., Chapman, H., Gantman, A., Vanaman, M., DeMarree, K., Igou, E., Wylie, J., Storbeck, J., Andreychik, M. R., … Uhlmann, E. L. (2021). A creative destruction approach to replication: Implicit work and sex morality across cultures. *Journal of Experimental Social Psychology, 93,* 104060.

van Elk, M., Groenendijk, E., & **Hoogeveen, S.** (2020). Placebo brain stimulation affects subjective but not neurocognitive measures of error processing. Journal of Cognitive Enhancement, 4(4), 389–400. https://doi.org/10.1007/s41465-020-00172-6

Sarafoglou, A., **Hoogeveen, S.**, Matzke, D., & Wagenmakers, E.-J. (2019). Teaching good research practices: Protocol of a research master course. *Psychology Learning & Teaching.* https://doi.org/10.1177/1475725719858807

**Hoogeveen, S.**, Schjoedt, U., & van Elk, M. (2018). Did I do that? Expectancy effects of brain stimulation on error-related negativity and sense of agency. *Journal of Cognitive Neuroscience*, 1–14. https://doi.org/10.1162/jocn_a_01297