Comparing Analysis Blinding With Preregistration In The Many-Analysts Religion Project

Alexandra Sarafoglou, Suzanne Hoogeveen, and Eric-Jan Wagenmakers

Department of Psychology, University of Amsterdam, The Netherlands

Abstract

In psychology, preregistration is the most widely used method to ensure the confirmatory status of analyses. However, the method has disadvantages: not only is it perceived as effortful and time consuming, but reasonable deviations from the analysis plan demote the status of the study to exploratory. An alternative to preregistration is analysis blinding, where researchers develop their analysis on an altered version of the data. In this experimental 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 = 71.40) 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 required approximately the same amount of time. 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.

Keywords: Open Science, Meta-Science, Replication Crisis, Many Analysts

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 one of the most vigorous. Preregistration protects the confirmatory status of the study by restricting the researchers’ degrees of freedom in conducting a study and analyzing the data (e.g., Chambers, 2017; Munafò et al., 2017; Wagenmakers et al., 2012). When preregistering studies, researchers specify in detail the study design, sampling plan, measures, and analysis plan before data collection. By specifying these aspects beforehand, researchers protect themselves against their (subconscious) tendencies to select favorable –that is, statistically significant– results.

Preregistration is fair in the sense that it restricts the researchers’ degrees of freedom. However, this implies that researchers must anticipate all possible peculiarities of the data and define analysis paths for each scenario, which can be perceived as effortful and time-consuming (Nosek & Lindsay, 2018; Sarafoglou 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. Such reasonable deviations include, for instance, removing outliers, transforming skewed data, or accounting for measurement invariance. From our personal experience, such deviations are usually small and do not affect the main conclusions of the study. However, adjusting the analysis to properties of the data, the analysis will be demoted from “confirmatory” to “exploratory” even when the adjustments were entirely appropriate and

Correspondence concerning this article should be addressed to: Alexandra Sarafgolou, Nieuwe Achtergracht 129B, 1001 NK Amsterdam, The Netherlands, E-mail: alexandra.sarafoglou@gmail.com.

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 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 (e.g., potential treatment effects or differences across conditions).

An overview on different blinding techniques for common study designs in experimental psychology is provided in Dutilh et al. (2019). One can create a blinded version of the data, for instance, by equalizing the group means across experimental conditions in factorial designs, by adding random noise to all values of the key outcome measure, or by shuffling the key outcome measures in regression designs. The latter technique was used in the present project. Shuffling the key outcome measures in regression designs implies reordering the dependent variable columns in the dataset while leaving all other columns untouched. The resulting blinded data are therefore complete, the column names are identical and the data have the same structure as the real data. Note that in contrast to the analysis of simulated data or data from a previously conducted (pilot) study, blinding of the analysis concerns the use of the actual data from a study.

As such, the analyst can examine the demographic characteristics of the sample, visualize the distribution of the variables, identify outliers, handle missing cases, or explore the factor structure of relevant measures. The analyst is thus able to create a reproducible analysis script including all steps in the analysis pipeline: from preprocessing the data to executing the appropriate statistical analysis. Most importantly, the analyst develops their analytic strategy without being able to determine how their analytic choices impact the significance level of the predictors. The blinding procedure has destroyed the relationship with the selected outcome variable, so that any analysis performed using this outcome variable will not be significant. 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.

Analysis blinding can be used either as stand-alone practice for data analysis or as a complement to preregistration. The latter was implemented, for example, in the study by Dutilh et al. (2017). The authors preregistered their analysis but anticipated deviations in the analysis plan due to the complexity of the statistical model and data structure. Analysis blinding allowed the authors to adjust the analysis plan to the specific peculiarities of the collected data while still maintaining its confirmatory status. In the current project, evaluate the differences between the two experimental conditions, we also deployed both strategies. That is, we preregistered our analysis plan on the OSF before data collection, but validated them on a blinded version the data.

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 how their change of beliefs on the two MARP research questions.

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 hours worked 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.

Table 1

Overview this Study's Materials Available on the Open Science Framework.

Resource	DOI	Citation
Project page	10.17605/osf.io/vy8z7	Hoogeveen, Sarafoglou, van Elk, et al. (2022c)
Preregistration	10.17605/osf.io/2cdht	Sarafoglou et al. (2022)
Data and analysis code	10.17605/osf.io/gkxqy	Hoogeveen, Sarafoglou, van Elk, et al. (2022b)
Stage 1 materials (preregistration)	10.17605/osf.io/a5ent	Hoogeveen, Sarafoglou, van Elk, et al. (2022e)
Stage 1 materials (blinding)	10.17605/osf.io/ktvqw	Hoogeveen, Sarafoglou, van Elk, et al. (2022d)
Surveys and ethics documents	10.17605/osf.io/kgqze	Hoogeveen, Sarafoglou, van Elk, et al. (2022f)
MARP data	10.31234/osf.io/dpex6	Hoogeveen, Sarafoglou, van Elk, et al. (2022a)

Disclosures

Preregistration and Analysis Blinding

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

Data and Materials

Table 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/.

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).

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.

Methods

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.

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.

Study Design

The current design was a between-subjects design (at the team level). Our dependent variables were (1) total hours worked, (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).

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. We assigned these teams to the preregistration condition since the blinded data were already available to them.

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 hours worked 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 (when indicating their hours worked) or give joint answers depending on the consensus within the team. The pre-survey, mid-survey, and post-survey were generated using Google Forms.

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).

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.

MARP Data and Data Documentation

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

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

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

Independent Variable: Assigned Analytic Strategy

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

Teams in the analysis blinding condition received a blinded version of the MARP data and a document which briefly explained what analysis blinding is, why analysis blinding can be beneficial, what analysts need to take into account when working with blinded data (e.g., analyses on blinded data may yield different results than when performed on the real data), and which blinding strategy was applied on the MARP data. Specifically, participants received the following information about the blinding strategy:

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.*

Dependent Variables: Hours Worked, Experienced Effort, Experienced Frustration, and Deviations From the Planned Analysis

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

One item asked to indicate how many hours it took the team to accomplish the tasks at the respective stage of the project. The hours of work required to complete a stage thus goes beyond simply writing the preregistration or developing the analysis script and also encompasses potential research that went into finding the appropriate analysis strategies, as well as discussions among team members. 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 (2006). The measures "Hours worked", "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. For researchers in the preregistration condition, deviations from the analysis plan concerned deviations from the analysis described in the preregistration document. For researchers in the analysis blinding condition, deviations from the analysis plan concerned adjustments of the analysis script they had developed for the blinded dataset. If researchers 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]

Reflection on Hours Worked

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

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. Other demographic information (e.g., age, gender, ethnicity) was not collected.

[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.

Respondents' Prior Beliefs

In the pre-survey, one item asked respondents about their subjective beliefs about the plausibility of the research questions *before* analyzing the data. This item was answered using a 7-point Likert-type scale from 1 (*Very unlikely*) to 7 (*Very likely*).

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. That is, researchers in the preregistration condition uploaded the filled out preregistration template, researchers in the analysis blinding condition uploaded their analysis script. 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.

Statistical Model

We used Bayesian inference for all statistical analyses. As preregistered, we aimed to collect at least strong evidence (i.e., a Bayes factor of at least 10) in favor for our hypotheses. Each hypothesis was tested against the null hypothesis that the respective outcomes are the same under both conditions. To test hypothesis 1 and 2, we conducted one-sided Bayesian independent samples *t*-tests. To test hypothesis 3, we conducted a one-sided Bayesian

Mann-Whitney U test. For hypothesis 1 and 2, we additionally conducted a robustness analysis to check how different prior specifications influence the results and a sequential analysis to check how the evidence changes as the data accumulates. For all three analyses, we assigned a one-sided Cauchy prior distribution with scale 0.707 to the effect size (i.e., $\delta \sim \text{Cauchy}^{-}(0, 0.707)$). These analyses were conducted in JASP (JASP Team, 2021).

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

We then estimated the log marginal likelihoods of these models using bridge sampling and computed the Bayes factor for these two models (Gronau et al., 2020a; Gronau 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, 2008; Klugkist et al., 2005; 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 Core Team, 2021).

Deviations from the Preregistration. The following deviations from the analysis plan were decided based on the blinded data. 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 hours worked" 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 hours worked" since this variable was heavily right-skewed.

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

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

Results

Sample Characteristics

The career stages and research backgrounds featured in each team are shown in Table 2. As apparent from Figure 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).

Prior beliefs for research question 1 were slightly higher in the preregistration group ($M = 4.95$, $SD = 1.12$) than in the blinding group ($M = 4.85$, $SD = 1.20$), yet the Bayes factor of the Mann-Whitney U test indicated moderate evidence against a difference: $\text{BF}_{01} = 4.60$, $\delta = -0.07$, $95\%[-0.41,\ 0.30]$. For research question 2, the same pattern emerged (preregistration: $M = 5.05$, $SD = 1.13$; blinding: $M = 4.88$, $SD = 1.12$), with again moderate evidence against a difference: $\text{BF}_{01} = 3.76$, $\delta = -0.14$, $95\%[-0.50,\ 0.21]$. As reported in Hoogeveen, Sarafoglou, Aczel, et al. (2022), there was no positive relation between prior beliefs about the plausibility of the two research questions and the reported effect sizes.

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.

Table 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.

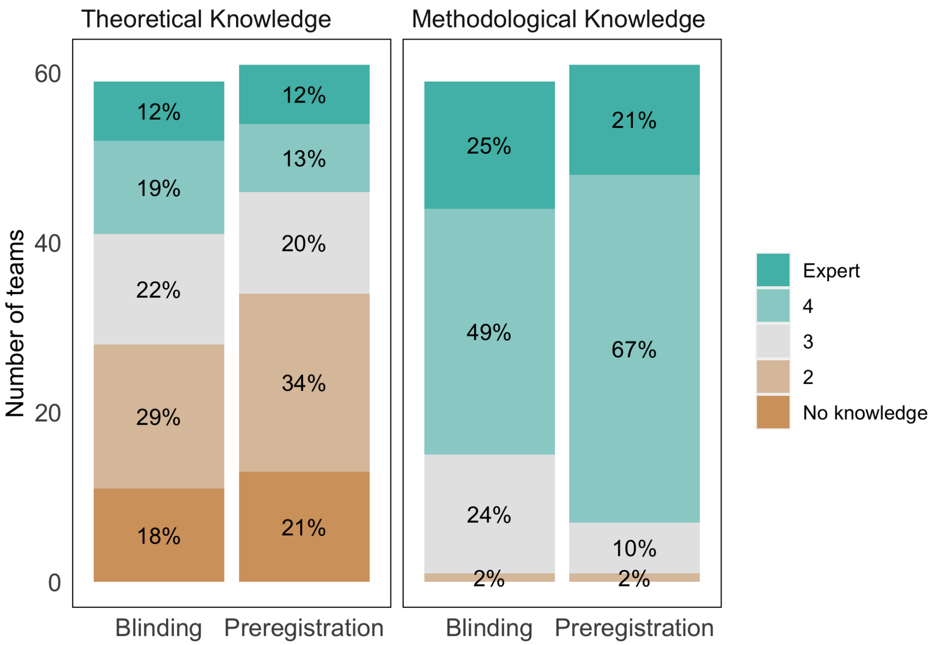


Figure 1

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 left bar represents responses from teams who did analysis blinding and the right bar represents responses from teams preregistered.

Table 3

For each condition, means and standard deviations for the hours worked (workload), perceived effort, perceived frustration, and reflection on hours worked. Statistics are shown for the total project duration and separately for each stage. For each stage, the number represents the mean with the standard deviation in round brackets. For correlations, the number represents the median estimate for the Bayesian Pearson correlation coefficient and the number in square brackets represents the 95% credible interval. The last column shows the median estimate for the Bayesian Pearson correlation coefficient ρ for values in stage 1 and 2, the number in square brackets the corresponding 95% credible interval.

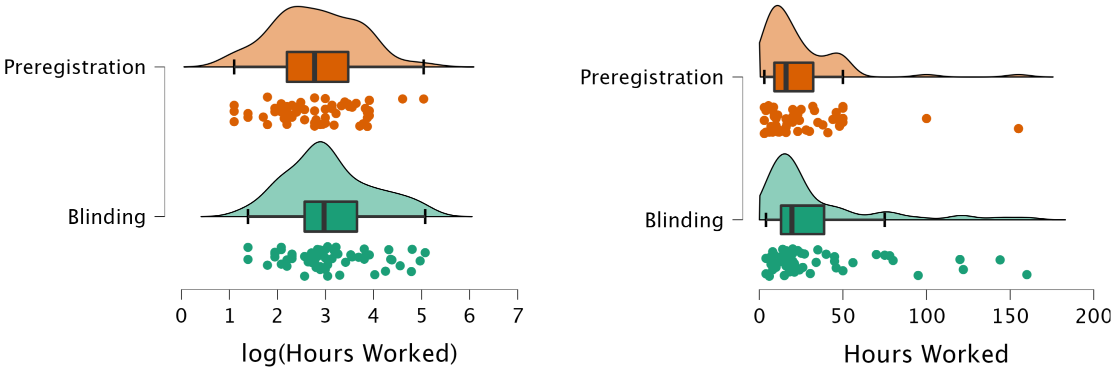
Measure	Condition	Total	Stage 1	Stage 2	$\rho(Stage1, Stage2)$
Effort	Blinding	8.44 (2.46)	4.42 (1.21)	3.95 (1.63)	0.47 [0.25, 0.64]
	Preregistration	8.78 (2.17)	4.37 (1.34)	4.46 (1.27)	0.38 [0.15, 0.57]
Frustration	Blinding	5.98 (2.66)	2.98 (1.41)	2.97 (1.82)	0.32 [0.08, 0.52]
	Preregistration	5.97 (2.22)	3.06 (1.55)	2.95 (1.36)	0.16 [-0.09, 0.38]
Reflection Hours Worked	Blinding	6.59 (1.39)	3.4 (0.86)	3.15 (0.96)	0.19 [-0.05, 0.41]
	Preregistration	6.32 (1.00)	3.12 (0.67)	3.23 (0.69)	0.12 [-0.12, 0.35]
Hours Worked	Blinding	33.12 (35.34)	19.11 (18.34)	13.78 (18.86)	0.76 [0.63, 0.85]
	Preregistration	23.94 (24.9)	8.43 (7.31)	15.75 (21.27)	0.32 [0.09, 0.52]
Log(Hours Worked)	Blinding	3.08 (0.89)	2.55 (0.90)	1.94 (1.18)	0.59 [0.40, 0.73]
	Preregistration	2.79 (0.88)	1.81 (0.82)	2.23 (1.01)	0.60 [0.42, 0.74]

Confirmatory Analyses

Table 3 shows the descriptive statistics of the dependent variables for each condition, for the entire project duration and separately for each stage.

The measures hours worked, perceived effort, and reflection on hours worked were positively correlated, yet not so strongly to suggest they measured the exact same concept. The Bayesian Kendall's tau correlations were as follows: For workload in hours and perceived effort is $\tau = .49$, $\text{BF}_{+0} = 2.6 \times 10^{12}$. For hours worked and reflection on hours worked $\tau = .32$, $\text{BF}_{+0} = 83476$. Finally, for perceived effort and reflection on hours worked, $\tau = .40$, $\text{BF}_{+0} = 2.3 \times 10^{8}$. Subsequently, for the t-tests, we report the median δ's with 95% credible intervals as effect size metrics.

Hours Worked. Hypothesis 1 stated that the total hours worked 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



(a) *Raincloud Plot for log hours worked* (b) *Raincloud Plot for hours worked*

Figure 2

Reported total hours worked 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.

is, that both teams take the same amount of time, with a Bayes factor of $\text{BF}_{0-} = 13.19$, $\delta = 0.29,\ 95\%[-0.05,\ 0.65]$. Figure 2 illustrates the responses of the reported hours worked. 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. An exploratory sequential analysis showed that as the data accumulate, the evidence in favor for the null hypothesis gradually increases.

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 $\text{BF}_{-0} = 0.41$, $\delta = -0.133,\ 95\%[-0.48,\ 0.21]$. These results are not robust against different prior settings. Depending on the prior choices, the evidence in favor of the null hypothesis fluctuates

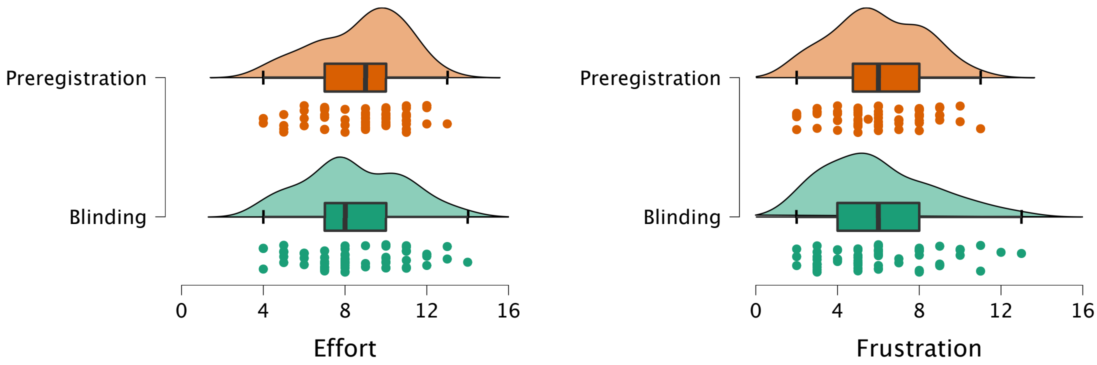


Figure 3

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.

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 3 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 $\text{BF}_{0-} = 5.00$, $\delta = -0.01$, $95\%[-0.35,\ 0.34]$. The right panel in Figure 3 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.

Table 4

Reported deviations from 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.

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 4. We collected strong evidence in favor for our hypothesis, that is, $\text{BF}_{r0} = 71.40$. The estimated probability that a team would deviate from their analysis plan was almost twice as high for for teams who preregistered (i.e., 38%) compared to team who did analysis blinding (i.e., 20%).

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

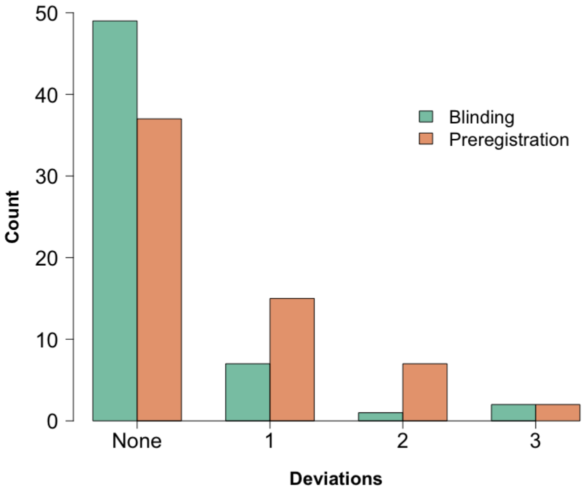


Figure 4

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.

Exploratory Analysis

Differences of the Many-Analysts' Conclusions Per Condition

Elaborate results of the many-analysts' conclusions about the substantive research questions are reported in Hoogeveen, Sarafoglou, Aczel, et al. (2022). Here, we briefly show the analysis teams' findings split per experimental condition. In Figure 5, the standardized effect sizes (β's) reported by the analysis teams are displayed per condition and research question.[2] For research question 1 ("Do religious people self-report higher well-being?"), all teams in the blinding condition reported positive effect sizes for which the 95% confidence/credible interval excludes zero. The median reported $\beta = 0.125$ and the median absolute deviation (MAD) = 0.030. Similarly, for the teams in the preregistration

[2] We were able to extract 99 β-coefficients for research question 1 and 101 for research question 2. The remaining teams provided effect size metrics that could not be converted to standardized regression coefficients. See the online Appendix to Hoogeveen, Sarafoglou, Aczel, et al. (2022) at https://osf.io/9kpfu/ for more information.

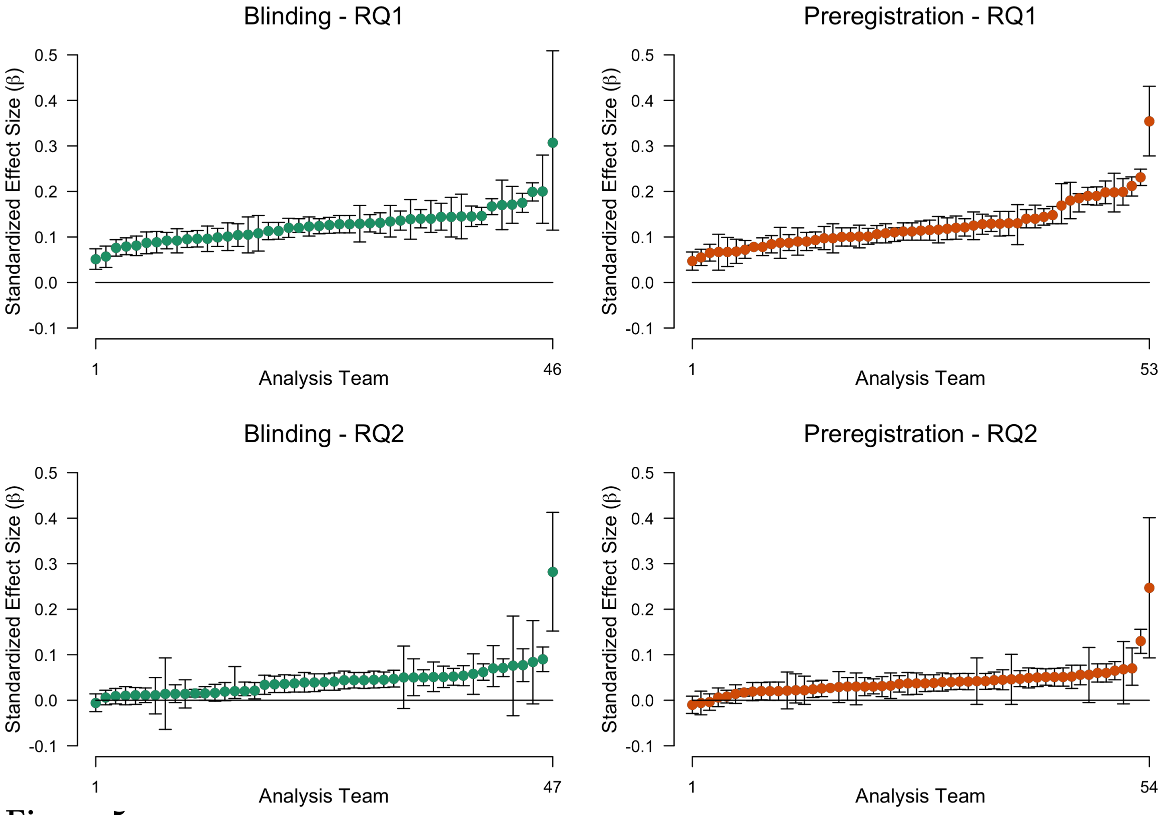


Figure 5

Effect sizes (β-coefficients) with 95% confidence or credible intervals for the two research questions reported by the analysis teams in the MARP. The top row shows the β's for the effect of religiosity on self-reported well-being (research question 1) and the bottom row for the effect of cultural norms of religion on the relation between religiosity and self-reported well-being (research question 2). The left panels show the β's for teams in the blinding condition (in green) and the right panels for teams in the preregistration condition (in orange). The β's are ordered from smallest to largest.

condition, all teams reported positive effect sizes with 95% confidence/credible intervals excluding zero. The median reported $\beta = 0.114$ and the MAD $= 0.039$. For research question 2 ("Does the relation between religiosity and self-reported well-being depend on perceived cultural norms of religion?"), the majority of teams again reported positive effect sizes with confidence/credible intervals excluding zero. That is, in the blinding condition, 97.9% of the β's were positive and 66.0% of the intervals excluded zero, median $\beta = .040$, MAD $= .030$. In the preregistration condition, 94.4% of the β's were positive and 64.8% of the intervals excluded zero, median $\beta = .037$, MAD $= .020$.

Total Hours Worked

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

In addition, we compared the reported hours worked between the two project stages. Figure 6 illustrates the responses of the reported work hours separately for stage 1 and stage 2. The difference in total hours worked 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 ($M = 19.11$, $SD = 18.33$) than teams in the preregistration condition ($M = 8.43$, $SD = 7.31$).

Reflection on Hours Worked

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 (i.e., $M = 13.78$ ($SD = 18.86$) for teams in the analysis blinding condition and $M = 15.75$ ($SD = 21.27$) for than teams in the preregistration condition). 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.

Independently Coded Deviations

In an additional exploratory analysis, we compared the deviations reported by the analysis teams with the deviations we identified. For this purpose, we (SH and AS) independently coded deviations from the analysis plan for each team. For the teams in the preregistration condition, the authors compared the analysis plan from the preregistration form with the responses from the post-survey. Only when information did not emerge from

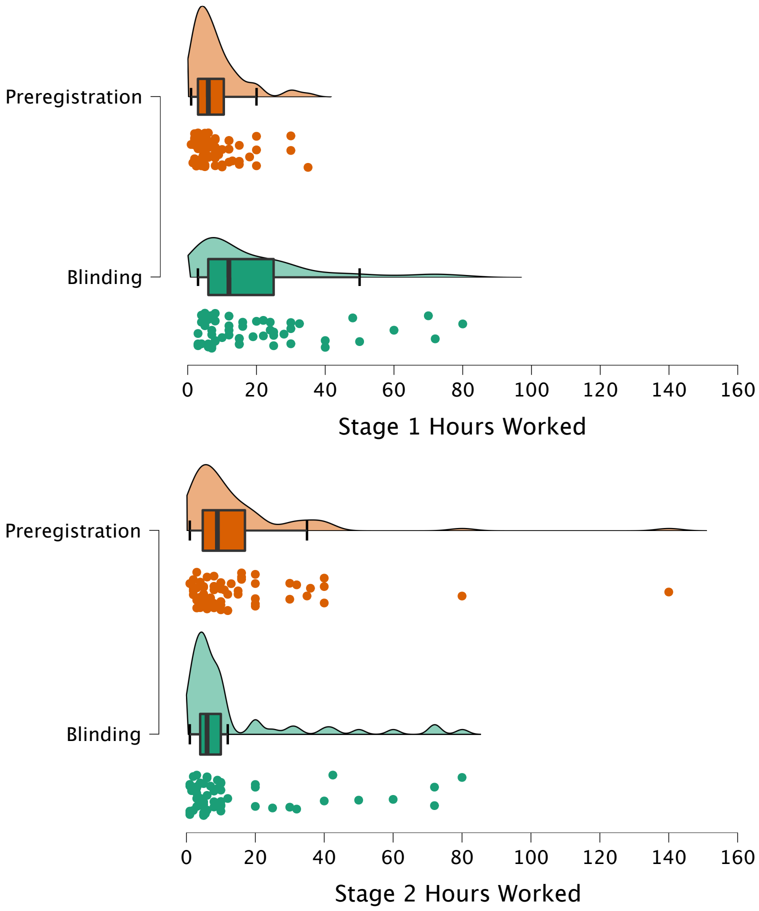


Figure 6

Reported total hours worked 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.

the post-survey did the authors review the authors' report or the final analysis scripts. For teams in the analysis blinding condition, the authors compared the analysis scripts from the blinded data with the analysis scripts on the real data. Initially, we evaluated three teams independently using the same checklist as we presented in the post-survey. Subsequently, we discussed their results and agreed upon the following adjustments. We decided not to consider it a deviation if teams had planned to conduct their statistical analyses with multiple dependent variables but reported only one of them in the post-survey. This was because we had explicitly instructed the teams to provide us with only one effect size. For aspects where we did not know the answer (e.g., because the analysis plan was too vague), we coded it NA. In addition, two teams were excluded since they did not submit a final analysis (though they completed the post-survey and self-reported deviations). The inter-class correlation between the two independently coded deviations was satisfactorily with $ICC = 0.71$. We then resolved any disagreement by discussion and then used the combined coding to test $\mathcal{H}_4$.

The result of this exploratory analysis is presented in Table 6. Based on the independent coding, we find extreme evidence for the hypothesis that teams in the analysis blinding condition deviated less from their planned analysis than teams who preregistered, with $BF_{r0} = 357.18$. The estimated probability that a team would deviate from their analysis plan was more than three times as high for for teams who preregistered (i.e., 37.7%) compared to team who did analysis blinding (i.e., 11.9%). Note that the independently coded deviations were fewer than those reported by the teams. When self-reported, a total of 50 deviations were disclosed, while the independent coders identified 44 deviations (based on 118 teams). Moreover, the inter-class correlation of $ICC = 0.43$ between the self-reported deviations and the independently coded deviations is not satisfying. We attribute these differences to the fact that teams have a better insight into their own analyses than independent coders, who might easily miss some deviations. We also judge the self-reported deviations as more accurate.

Table 5

Reported deviations from planned analysis per condition as coded by two independent raters.

	Preregistration	Analysis Blinding
Nr. of Teams Reporting Deviations	23/61 (37.7 %)	7/59 (11.86 %)
Domains		
Exclusion Criteria	15/61 (24.59 %)	0/59 (0 %)
Included Variables	11/61 (18.03 %)	4/59 (6.78 %)
Operationalization of IV	4/61 (6.56 %)	1/59 (1.69 %)
Statistical Model	3/61 (4.92 %)	5/59 (8.47 %)
Statistical Test	0/61 (0 %)	0/59 (0 %)
Operationalization of DV	1/61 (1.64 %)	0/59 (0 %)
Hypothesis	0/61 (0 %)	0/59 (0 %)
Direction of Effect	0/61 (0 %)	0/59 (0 %)

Note. Teams may report multiple deviations.

Robustness Checks

In this study, we deviated from our preregistration at several points. First, we have adapted our analyses to the properties of the data (e.g., transformations due to the skewness of the data). Second, we deviated from our sampling plan by assigning teams that merged to the analysis-blinding condition ($n = 4$). One team also switched from the analysis-blinded condition to the preregistration condition on its own. To confirm that our conclusions are not dependent on these deviations, we performed a series of robustness checks. The results of these analyses are shown in Table 6. This table contains for each hypothesis the Bayes factor of (1) the main analysis, (2) the original preregistered analysis, (3) the analysis in which merged or switched teams were excluded, and (4) the Bayes factor of hypothesis 4 with the independently coded deviations instead of the self-reported deviations.

Constraints on Generality

We believe that our results can be generalized to other research designs (i.e., experimental studies) and do not apply only to correlational studies. However, the outcomes of this study might be dependent on the complexity of the data and hypotheses researchers

Table 6

Robustness checks for the analysis of the four main hypotheses.

		BF_{r0}			
Robustness Set	n	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$	$\mathcal{H}_4$
Main Analysis	120	0.076	0.409	0.200	71.40
Exact Adherence to Preregistration	120	0.080	0.349	–	27.95
Excluding Merged and Switched Teams	115	0.072	0.496	0.180	45.19
Independently Coded Deviations	116	–	–	–	357.18

Note. For each hypothesis (columns) and robustness set (rows) the Bayes factor in favor of the restricted alternative hypothesis versus the null-hypothesis is given. See the main text for an explanation of the different robustness sets. Empty cells indicate that the adjustments were not relevant for the particular hypothesis.

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.

In addition, our results may not generalize to paradigms and topics that analysis teams are very familiar with. That is, researchers are better at anticipating analysis plans for paradigms they often work with than developing an analysis plan for a completely new dataset, measures, and theories. At the same time, most deviations in the present study occurred for the data exclusions, mostly related to unexpected peculiarities of the data that are unrelated to the topic or paradigm (e.g., some participants provided a nonsensical age).

Moreover, we cannot determine to which extend the results of the current study generalize beyond multi-team projects. It is possible that researchers conducting their own studies need to perform more preparatory steps than researchers in our study, especially when preregistering or blinding for their own projects. 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.

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

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

Descriptives on stage 2 showed that teams who preregistered and teams who did analysis blinding took about the same amount of time to execute the analysis. We speculate that this result may be due to an improper communication to the teams. To complete stage 2, the teams were instructed to execute their planned analyses on the real data and fill out the post-survey to indicate their conclusions and summarize their results. We also provided teams with the type of information required to fill in the post-survey and recommendations about how to organize their OSF folder. These recommendations included to add a "ReadMe" file that documents the uploaded files and a brief summary of the main conclusions. The time associated with creating these files might have distorted our measure on hours worked. 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 the teams reported their hours worked.

The current study has several limitations, the first one concerning the measurements. While our measures of workload, effort, and frustration have high face validity and were taken from a previous study (Hart, 2006), their validity in the present context is unknown. Especially the reported number of hours spent on the project should be interpreted with caution, as this was filled out in retrospect by one team member. Future projects could opt for a more objective measure and ask teams in advance to log their work hours (Parry et al., 2021).

The analysis teams, although co-authors of the manuscript, may have been less invested in this large-scale collaboration project than if it were their own research. On the one hand, less emotional commitment to the research hypotheses may be advantageous since it lowers the motivation to engage in questionable research practices such as *p*-hacking. On the other hand, it may also reduce the teams' effort and hence the quality of the analyses. This latter possibility was raised, for instance, by Ross et al. (2022), one of the commentaries of the MARP. The low number of deviations in the present study could be due to a possible lack of commitment of the teams: When teams have little emotional investment to a study, they might be less likely to deviate from their planned analyses, even if such adjustments would have been necessary. However, one could also turn the argument around and argue that the low number of deviations is a sign that the teams were indeed invested in the project and that the analyses presented were therefore of high quality and required few adjustments. Future research could thus assess whether the quality of proposed analysis plans is sufficiently high, or whether the quality of final analyses 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.

Although adherence to the analysis plan is desirable to ensure the confirmatory status of an analysis, we speculate that the teams’ deviations are not consequential. As the main results of the MARP show, almost all teams found a positive effect for the research question one. Thus, the fact that teams in the preregistration condition deviated from their analysis plans more often than teams in the analysis blinding condition most likely had no practical consequences. The extent to which this pattern of inconsequential deviations also holds for other data and research questions (e.g., an experiment in which the null hypothesis is true) needs to be investigated in future studies.

The current study focused on planning and executing an analysis whose confirmatory status could be guaranteed. As such, 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.

Regardless of our results, the decision whether to prefer preregistration or blinding of analyses is always a matter of circumstance and research design. In the MARP, analysis blinding has been particularly suitable since the data managers (i.e., the team with access to the real data) were completely independent of the analysis teams. From our subjective experience, we also found that researchers who had access to the blinded data asked us data managers fewer questions in stage 1 than researchers who had only access to the data documentation. Therefore, we can imagine that especially many-analysts projects can benefit greatly from analysis blinding. It would also be worth considering giving researchers access to blinded data first when they want to perform reanalyses or meta-analyses, rather than providing them directly with the real data.

In contrast, in very small research groups, there is often no guarantee that the analysis blinding has actually been done effectively. For instance, it cannot be ruled out that data managers and analysts discuss certain data patterns and thus develop new analyses that presumably lead to desirable results. Preregistrations allow for better control as they are time-stamped and it is possible to find out exactly in which time period data was collected.[3]

However, even in cases researchers in which researchers solely preregister their study, the analysis plan can be developed on the basis of simulated data or on data from previous work (which was recommended, for instance, in Nosek et al., 2019). The resulting syntax can then be added to the preregistration document. Refining an analysis plan on simulated data helps researchers anticipate an analytic strategy and removes ambiguities from the preregistration.

We would like to emphasize again, however, that researchers can also use preregistration and analysis blinding 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 in the social and behavioural sciences. Analysis blinding ties in with current methodological reforms for more transparency since it safeguards the confirmatory status

[3] Note that this excludes preregistrations of secondary data which again introduces the uncertainty of whether the analysts were truly naive to any potentially biasing information.

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.

Author Contributions

Contributorship was documented with CRediT taxonomy using tenzing Holcombe et al., 2020.

Alexandra Sarafoglou: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, and Writing - original draft.

Suzanne Hoogeveen: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, and Writing - original draft.

Eric-Jan Wagenmakers: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, and Writing - original draft.

Conflicts of Interest

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Acknowledgements

This research was supported by a talent grant from the Netherlands Organisation for Scientific Research (NWO) to AS (406-17-568), as well as an Advanced ERC grant from the NWO to EJW (743086 UNIFY). The analyses were conducted using the `R` packages .

The analyses were conducted in JASP (JASP Team, 2021) and in `R` version 4.0.3 (R Core Team, 2021) using the following packages: `BayesFactor` (Morey & Rouder, 2018), `bridgesampling` (Gronau et al., 2020b), `rstan` (Stan Development Team, 2022), `papaja` (Aust & Barth, 2020), `ggplot2` (Wickham, 2016), `purrr` (Henry & Wickham, 2020), `stringr` (Wickham, 2019), `dplyr` (Wickham et al., 2020), `tidyverse` (Wickham et al., 2019), `rlang` (Henry & Wickham, 2022), `RColorBrewer` (Neuwirth, 2014), `rethinking` (McElreath, 2020) and `bayesplot` (Gabry & Mahr, 2021).

References

Aust, F., & Barth, M. (2020). *papaja: Prepare reproducible APA journal articles with R Markdown* [R package version 0.1.0.9997]. https://github.com/crsh/papaja

Center for Open Science. (2021). Open Science Framework.

Chambers, C. D. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

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.

Dutilh, G., Sarafoglou, A., & Wagenmakers, E.-J. (2019). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, *198*, S5745–S5772.

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.

Gabry, J., & Mahr, T. (2021). Bayesplot: Plotting for Bayesian models [R package version 1.8.0]. https://mc-stan.org/bayesplot/

Gelfand, A. E., Smith, A. F., & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association*, *87*, 523–532.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020a). Bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software, Articles*, *92*(10), 1–29.

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.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020b). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, *92*(10), 1–29. https://doi.org/10.18637/jss.v092.i10

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*, 904–908.

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.* https://psyarxiv.com/nj4es

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools* [R package version 0.3.4]. https://CRAN.R-project.org/package=purrr

Henry, L., & Wickham, H. (2022). *Rlang: Functions for base types and core r and 'tidyverse' features* [R package version 1.0.2]. https://CRAN.R-project.org/package=rlang

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Chapman & Hall/CRC.

Holcombe, A. O., Kovacs, M., Aust, F., & Aczel, B. (2020). Documenting contributions to scholarly articles using CRediT and tenzing. *PLoS One*, *15*, e0244611.

Hoogeveen, S., Sarafoglou, A., van Elk, M., & Wagenmakers, E. (2022a). A many-analysts approach to the relation between religiosity and well-being: The dataset. *PsyArXiv*. https://psyarxiv.com/dpex6. https://doi.org/10.31234/osf.io/dpex6

Hoogeveen, S., Sarafoglou, A., van Elk, M., & Wagenmakers, E. (2022b). Many-analysts religion project: Data and analysis code. *OSF Project Page*. https://doi.org/10.17605/OSF.IO/GKXQY

Hoogeveen, S., Sarafoglou, A., van Elk, M., & Wagenmakers, E. (2022c). Many-analysts religion project: Main project page. *OSF Project Page*. https://doi.org/10.17605/OSF.IO/VY8Z7

Hoogeveen, S., Sarafoglou, A., van Elk, M., & Wagenmakers, E. (2022d). Many-analysts religion project: Stage 1 materials for the blinding condition. *OSF Project Page*. https://doi.org/10.17605/OSF.IO/KTVQW

Hoogeveen, S., Sarafoglou, A., van Elk, M., & Wagenmakers, E. (2022e). Many-analysts religion project: Stage 1 materials for the preregistration condition. *OSF Project Page*. https://doi.org/10.17605/OSF.IO/A5ENT

Hoogeveen, S., Sarafoglou, A., van Elk, M., & Wagenmakers, E. (2022f). The many-analysts religion project: Surveys. *OSF Project Page*. https://doi.org/10.17605/OSF.IO/KGQZE

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., Levy, N., Neely, A., Qiu, L., Rabelo, A., Ramsay, J. E., … van Elk, M. (2021). The Einstein effect: Global evidence for scientific source credibility effects and the influence of religiosity. *PsyArXiv*. https://doi.org/10.31234/osf.io/sf8ez

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., & van Elk, M. (2018). Advancing the Cognitive Science of Religion through Replication and Open Science. *Journal for the Cognitive Science of Religion*, *6*, 158–190. https://doi.org/10.1558/jcsr.39039

JASP Team. (2021). JASP (Version 0.16.2.0) [Computer software].

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.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*, 1–14.

MacCoun, R., & Perlmutter, S. (2015). Hide results to seek the truth: More fields should, like particle physics, adopt blind analysis to thwart bias. *Nature*, *526*, 187–190.

MacCoun, R., & Perlmutter, S. (2018). Blind analysis as a correction for confirmatory bias in physics and in psychology. In S. O. Lilienfeld & I. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 297–322). John Wiley; 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.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman & Hall/CRC Press.

McElreath, R. (2020). *Rethinking: Statistical rethinking book package* [R package version 2.13]. https://github.com/rmcelreath/rethinking

Morey, R. D., & Rouder, J. N. (2018). *Bayesfactor: Computation of bayes factors for common designs* [R package version 0.9.12-4.2]. https://CRAN.R-project.org/package=BayesFactor

Munafò, M., Nosek, B. A., Bishop, D., Button, K., Chambers, C., Du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J., & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.

Neuwirth, E. (2014). *Rcolorbrewer: Colorbrewer palettes* [R package version 1.1-2]. https://CRAN.R-project.org/package=RColorBrewer

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., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer*, *31*, 19–21.

Parry, D. A., Davidson, B. I., Sewall, C. J., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*, *5*, 1535–1547. https://doi.org/10.1038/s41562-021-01117-5

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.

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ross, R. M., Sulik, J., Buczny, J., & Schivinski, B. (2022). Many analysts and few incentives. *Religion, Brain, & Behaviour.*

Sarafoglou, A., Hoogeveen, S., van Elk, M., & Wagenmakers, E. (2022). Comparing analysis blinding with preregistration in the many-analysts religion project: Preregistration. *OSF Registries.* https://doi.org/10.17605/OSF.IO/2CDHT

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

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.

Silberzahn, R., & Uhlmann, E. L. (2015). Many hands make tight work. *Nature*, *526*, 189.

Stan Development Team. (2022). RStan: The R interface to Stan [R package version 2.26.9]. https://mc-stan.org/

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/https://doi.org/10.1016/j.jesp.2016.03.004

Veldkamp, C., Bakker, M., van Assen, M., Crompvoets, E., Ong, H., Soderberg, C., Mellor, D., Nosek, B. A., & 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).

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

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 01–17.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations* [R package version 1.4.0]. https://CRAN.R-project.org/package=stringr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H.

(2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation* [R package version 1.0.2]. https://CRAN.R-project.org/package=dplyr

Appendix

Preregistration Form: Analysis Plan

1. Title: Many Analysts Replication Project

2. Analysis Team

3. Research Question

   - Do religious people have higher well-being?
   - Does the relationship between religiosity and well-being depend on perceived cultural norms of religion?

4. Hypotheses

   List specific, concise, and testable hypotheses. Please state if the hypotheses are directional or non-directional. If directional, state the direction. A predicted effect is also appropriate here. If a specific interaction or moderation is important to your research, you can list that as a separate hypothesis.

   Example:

   If taste affects preference, then mean preference indices will be higher with higher concentrations of sugar.

5. Variables

   - Dependent variable(s)

     State which key dependent variable(s) you will use in your analysis. Name the specific column names of these variables as stated in the data documentation.

- Predictor variable(s)

  State which predictor variable (including moderators and covariates) you will use in your analysis. Name the specific column names of these variables as stated in the data documentation.

- Indices

  If applicable, please define how measures will be combined into an index (or even a mean) and what measures will be used. Include either a formula or a precise description of the method. If you are using a more complicated statistical method to combine measures (e.g. a factor analysis), please note that here but describe the exact method in the analysis plan section.

  Example:

  We will take the mean of the two questions above to create a single measure of “brownie enjoyment.”

6. Analysis Plan

- Statistical models

  What statistical model will you use to test each hypothesis? Please include the type of model (e.g. ANOVA, RMANOVA, MANOVA, multiple regression, SEM, etc) and the specification of the model. This includes each variable that will be included, all interactions, subgroup analyses, pairwise or complex contrasts, and any follow-up tests from omnibus tests. Provide enough detail so that another person could run the same analysis with the information provided. Remember that in your final article any test not included here must be noted as exploratory and that you must report the results of all tests.

  Note:

  This is perhaps the most important and most complicated question within the preregistration. Ask yourself: is enough detail provided to run the same analysis again with the information provided by the user? Be aware for instances where the statistical models appear specific, but actually leave openings for the precise test.

Example:

We will use a 2 X 3 repeated measures ANOVA (RMANOVA) with the mean preference indices as the outcome variable and the factors "sweetness" and "color" within subjects to analyze our results.

- Transformations

  If you plan on transforming, centering, recoding the data, or requiring a coding scheme for categorical variables, please describe that process.

  Example:

  The "Effect of sugar on brownie tastiness" does not require any additional transformations. However, if it were using a regression analysis and each level of sweet had been categorically described (e.g. not sweet, somewhat sweet, sweet, and very sweet), "sweet" could be dummy coded with 'not sweet' as the reference category. If any categorical predictors are included in a regression, indicate how those variables will be coded (e.g. dummy coding, summation coding, etc.) and what the reference category will be.

- Inference criteria

  What criteria will you use to make inferences? Please describe the information you'll use (e.g. specify the p-values, Bayes factors, specific model fit indices), as well as cut-off criterion, where appropriate. Will you be using one or two tailed tests for each of your analyses? If you are comparing multiple conditions or testing multiple hypotheses, will you account for this?

  Example:

  We will use the standard $p < .05$ criteria for determining if the ANOVA and the post hoc test suggest that the results are significantly different from those expected if the null hypothesis were correct. The post-hoc Tukey-Kramer test adjusts for multiple comparisons.

- Data exclusion

  How will you determine which data points or samples if any to exclude from your analyses, for instance, based on the attention check or missing data? How will outliers be handled?

  Example:

  We will verify that each subject answered each of the three tastiness indices. Outliers will be included in the analysis.

- Missing data

  How will you deal with incomplete or missing data?

  Note:

  For the well-being and religiosity measures there are no missing data. See the data documentation for a more detailed overview of missing data for each variable.

  Example:

  If a subject does not complete any of the three indices of tastiness, that subject will not be included in the analysis.

7. Other

   If there is any additional information that you feel needs to be included in your preregistration, please enter it here. Literature cited, disclosures of any related work such as replications or work that uses the same data, or other helpful context would be appropriate here.