Many-Analysts Religion Project: Reflection and Conclusion

Suzanne Hoogeveen*[1], Alexandra Sarafoglou*[1], Michiel van Elk[2], and Eric-Jan Wagenmakers[1]

[1]Department of Psychology, University of Amsterdam, The Netherlands

[2]Institute of Psychology, Leiden University, The Netherlands

Introduction

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

However, we believe the scope of the project and the effort of the analysis teams justifies highlighting some more in-depth observations. Here, we aim to address these supplementary findings, taking the points raised in the 17 commentaries written by various participating analysts as a guideline. We identified three overarching themes in the commentaries and our own experiences. First, there was a need for more focus on theoretical

---

*These authors contributed equally.
Corresponding author: Suzanne Hoogeveen. E-mail: suzanne.j.hoogeveen@gmail.com.

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.

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.; 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 re-

lated 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, 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).

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

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

religious majorities as a covariate or moderator in the analysis on the MARP data. (They found no effect, suggesting that well-being does not depend on the match between one's own religion and that of the majority in one's country.)[2] This approach is worth pursuing in future many-analysts projects on the topic of religion and well-being: since there are many large-scale surveys covering both constructs, this seems a feasible endeavor.

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

[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/.

[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.

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.

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 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).

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.

Acknowledgements

This research was supported by a talent grant from the Netherlands Organisation for Scientific Research (NWO) to AS (406-17-568) and a Templeton Foundation grant to MvE (60663).

References

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

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.*

Balkaya-Ince, M., & Schnitker, S. (2022). Advantages of using multilevel modeling approaches for the Many Analysts Religion Project. *Commentary in MARP special issue.*

Bulbulia, J. A. (2022). Causal models are needed to infer how religion affects mental health. *Commentary in MARP special issue.*

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.*

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.*

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.*

Islam, C.-G., & Lorenz, J. (2022). How to increase the robustness of survey studies. *Commentary in MARP special issue.*

Krypotos, A.-M., Klein, R., & Jong, J. (2022). Resolving religious debates through a multiverse approach. *Commentary in MARP special issue.*

Lodder, P. (2022). Why researchers should not ignore measurement error and skewness in questionnaire item scores. *Commentary in MARP special issue.*

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.*

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.*

Paloutzian, R. (2017). Invitation to the psychology of religion. New York.

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.*

Ross, R. M., Sulik, J., Buczny, J., & Schivinski, B. (2022). Many analysts and few incentives. *Commentary in MARP special issue.*

Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E.-J. (2022). Comparing analysis blinding with preregistration in the many-analysts religion project. *Manuscript submitted for publication*. https://doi.org/10.31234/osf.io/6dn8f

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*.

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*, 337–356. https://doi.org/10.1177/2515245917747646

Smith, E. (2022). Individual-level versus country-level moderation. *Commentary in MARP special issue*.

The MARP Team. (2022). A many-analysts approach to the relation between religiosity and well-being. *Manuscript submitted for publication*. https://doi.org/10.31234/osf.io/pbfye

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 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.

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*.

Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., van Dongen, N., Hoekstra, R., Moreau, D., van Ravenzwaaij, D., Sluga, A., Stanke, F., Tendeiro, J., & Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, *5*, 1473–1480.