Measurement Issues in the Many Analysts Religion Project

Marcel R. Schreiner[1], Brett Mercier[2], Susanne Frick[1], Dylan Wiwad[3], Marcel C. Schmitt[4], John Michael Kelly[5], Julian Quevedo Pütter[1]

[1] Department of Psychology, University of Mannheim, Mannheim, Germany

[2] Department of Psychology, University of Toronto, Toronto, ON, Canada

[3] Management and Organizations, Kellogg School of Management, Northwestern University, Evanston, IL, United States

[4] Faculty of Psychology, University of Koblenz-Landau, Landau, Germany

[5] Department of Psychological Science, University of California Irvine, Irvine, CA, United States

Author Note:

This material is based in part upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE-1839285).

This research was in part funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2277 “Statistical Modeling in Psychology”.

The Many Analysts Religion Project (MARP; The MARP Team, 2022) asked analysts to compare the relationship between religiosity and well-being across many different countries. Although this project makes a valuable effort towards improving our understanding of the psychology of religion, we highlight two measurement issues that should be considered when interpreting the results of the MARP.

First, when conducting cross-cultural research, researchers need to make sure they are measuring the same construct of interest in each cultural context (Steenkamp & Baumgartner, 1998) . This is particularly important in the psychology of religion, as religiosity may have different

components in different countries (Mercier et al., 2018). For example, attending religious services plays an important role in some religions (e.g. Christianity) and a relatively minor role in others (e.g. Buddhism; Pew Research Center, 2018). Thus, it is important to determine whether the MARP scale measuring "religiosity" is measuring the same psychological phenomenon in, for example, Lebanon as it is in the United States.

In statistical terms, this issue is often referred to as establishing measurement invariance. Measurement invariance occurs in several nested levels. Establishing invariance on one level allows for certain statistical comparisons and permits testing of subsequent levels (Steenkamp & Baumgartner, 1998). The first level, configural invariance, means that different groups have the same factor structure. In the context of the MARP, this means demonstrating that in each country, religiosity is best measured by a model with a single latent construct and that the items which make up this construct are the same across countries (for a description of the remaining levels of invariance, see Steenkamp & Baumgartner, 1998).

To test for measurement invariance, one of our analysis teams (Team 108; Mercier et al., 2021; analysis code available at https://osf.io/3sywn/) used the lavaan package (Rosseel, 2012) in R to fit a Multigroup Structural Equation Model to the MARP data (for an alternative test of measurement invariance, see the analysis code from Team 155, available at https://osf.io/kgt6b/; Schreiner et al., 2021). We fit a model where, in each country, each of the religiosity items loaded onto a common latent factor. We fixed the loadings for the first religiosity item at 1, while the remaining factor loadings and item intercepts were allowed to freely vary in each country. If configural invariance exists in the MARP religiosity items, this model should fit well. However, the model was a poor fit for the data, and most model fit statistics failed to reach commonly accepted benchmarks (CFI = .90, RMSEA = .16, SRMR = .05). This provides

evidence that responses to items measuring religiosity in the MARP are not measurement invariant, meaning that the factor structure of item responses is different across countries. Because a similar factor structure is required to appropriately compare the relationship between religiosity and well-being (and their interaction with cultural norms) across countries, it is unclear what the results of the MARP indicate. Future many analyst projects could avoid these issues by using only items that are measurement invariant across analysis groups.

The second measurement issue is the treatment of categorical items. Surveys, including those used in the MARP, generally consist of multi-item scales – often using a Likert scale format – from which (latent) trait estimates are generated. Trait estimates are often obtained by aggregating items into continuous indicators such as sum scores. This procedure assumes that all items are equally informative about the trait (McNeish & Wolf, 2020) and that distances between item categories are equal across items (Wakita et al., 2012). These assumptions may often be violated in empirical data. Consequently, trait scores can be biased, especially at the ends of trait distributions (Dumenci & Achenbach, 2008). This bias does not only occur for raw scores such as sum scores, but also for scores obtained from other models that assume continuous indicators, such as Principal Component Analysis (Dumenci & Achenbach, 2008). These problems are further aggravated by skewed data (Hutchinson & Olmos, 1998), as was the case for several items in the MARP data. On the applied level, biased trait scores may lead to wrong inferences, such as spurious correlations (Embretson, 1996). Compared to the original analyses of our teams, a post-hoc analysis using continuous indicators (scale means) yielded the same overall results regarding the main research questions, but the relation between religiosity and well-being was reduced (the incremental R^2 for the effect of religiosity changed from .03 to .01 and the

standardized regression coefficient of religiosity changed from 0.12, 95% CI [0.11, 0.14] to 0.10, 95% CI [0.08, 0.11], although it should be noted that the confidence intervals overlapped).

We argue that researchers, such as those analyzing the MARP data, should take the categorical nature of items into account by using appropriate models such as IRT (Lord, 1980; Lord & Novick, 1968) or item factor analysis models (Jöreskog & Moustaki, 2001). Alternatively, they should test the robustness of their results against violating the assumption of continuity explicitly. In addition, the MARP data provided by the organizers already contained precomputed means for the well-being items. Future many analyst projects should avoid providing any aggregated data, since this may guide modeling choices of the analyst teams.

We highlighted two measurement issues which should influence the interpretation of the MARP findings. These issues also illuminate a more general weakness of many analyst projects. Analysis approaches differ with respect to their quality, and inappropriate approaches may impact the overall results. In the worst case, a shared methodological shortcoming common in a research field may lead the majority of research teams to converge on an incorrect or biased result.

On the other hand, many analysts projects are a useful way to demonstrate how different research teams can reach different results with the same data and research questions (Silberzahn et al., 2018). A group of researchers with diverse methodological backgrounds generates heterogeneous ideas, extending the analysis multiverse beyond the capacities and resources of a single research team. Additionally, if inappropriate approaches are included next to appropriate ones and results nevertheless converge, this shows that results are robust against violations of specific model assumptions. For example, the relationship between religiosity and well-being

appears robust, regardless of the analysis assumptions. In contrast, the interaction with cultural norms appears to vary across research teams, and it might be illuminating to examine whether the diverging results can be attributed to different kinds of analysis approaches.

References

Dumenci, L., & Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment, 20*(1), 55–62. https://doi.org/10.1037/1040-3590.20.1.55

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20*(3), 201–212. https://doi.org/10.1177/014662169602000302

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36*(3), 347–387. https://doi.org/10.1207/S15327906347-387

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods, 52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Mercier, B., Kramer, S. R., & Shariff, A. F. (2018). Belief in god: Why people believe, and why they don't. *Current Directions in Psychological Science*, *27*(4), 263–268. https://doi.org/10.1177/0963721418754491

Mercier, B., Wiwad, D., Kelly, J. M. (2021). Team 108. https://osf.io/3sywn/

Pew Research Center. (2018). *The age gap in religion around the world*. https://www.pewforum.org/2018/06/13/the-age-gap-in-religion-around-the-world/

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. R package version 0.5-15. *Journal of Statistcial Software*, *48*(2), 1–36.

Schreiner, M. R., Frick, S., Schmitt, M. C., & Quevedo Pütter, J. (2021). Team 155.

https://osf.io/kgt6b/

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. https://doi.org/10.1086/209528

The MARP Team (2022). A many-analysts approach to the relation between religiosity and well-being. *Manuscript in preparation.*

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert scale: Comparing different numbers of options. *Educational and Psychological Measurement*, *72*(4), 533–546. https://doi.org/10.1177/0013164411431162