Accepted manuscript

Many analysts and few incentives

Robert M. Ross, Justin Sulik, Jacek Buczny, Bruno Schivinski

Macquarie University; Ludwig Maximilian University of Munich; Vrije Universiteit Amsterdam; SWPS University; RMIT University

To cite this article

Ross, R. M., Sulik, J., Buczny, J., & Schivinski, B. (2023). Many analysts and few incentives. *Religion, Brain & Behavior, 13*(3), 336–339. https://doi.org/10.1080/2153599X.2022.2070248

Accepted manuscript of a commentary published in Religion, Brain & Behavior, 13(3), 336–339 (2023), https://doi.org/10.1080/2153599X.2022.2070248. The authors' manuscript was posted on PsyArXiv (https://doi.org/10.31234/osf.io/jdtur) under the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/). Changes made: re-typeset and the corresponding author's email address left out.

This is the authors' accepted manuscript, as posted on PsyArXiv under CC BY 4.0; it differs from the published commentary in copy-editing, formatting and pagination, and the corresponding author's email address has been left out. Please cite the published version.

Re-typeset from the authors' accepted manuscript (CC BY 4.0); line spacing and layout changed. Downloaded from schivinski.com

Many analysts and few incentives

Robert M. Ross[1*], Justin Sulik[2], Jacek Buczny[3,4] and Bruno Schivinski[5]

[1]Department of Psychology, Macquarie University, NSW 2019, Australia; [2]Cognition, Values & Behavior, Ludwig Maximilian University of Munich, Germany; [3]Department of Experimental and Applied Psychology, Faculty of Behavioral and Movement Sciences, Vrije Universiteit Amsterdam, the Netherlands; [4]SWPS University of Social Sciences and Humanities, Sopot, Poland; [5]School of Media and Communication, RMIT University, VIC 3000, Australia

*Corresponding author: Robert M. Ross, Macquarie University, NSW 2019, Australia

ORCID: Robert M. Ross 0000-0001-8711-1675; Justin Sulik 0000-0002-0978-9496; Jacek Buczny 0000-0002-4932-196X; Bruno Schivinski 0000-0002-4095-1922.

Commentary on the Many Analysts Religion Project (Hoogeveen et al., Religion, Brain & Behavior, 13(3)).

The Many Analysts Religion Project (MARP) follows in a tradition of many analysts projects that have examined how variations in analytic choices can affect results (Silberzahn & Uhlmann, 2015). One of the core strengths of the many analysts approach is that there are no direct incentives for analysis teams to attempt to achieve statistical significance (Silberzahn et al., 2018). However, in this commentary we suggest that a lack of direct incentives to achieve particular outcomes can be a double-edged sword. We use our own submissions to the MARP project to illustrate our point.[1]

In our submissions we went to some trouble to develop multilevel models that were suitable for testing the focal hypotheses of the MARP project. Nonetheless, our submissions shared the same potentially consequential limitation: we ignored issues pertaining to construct validity.[2] In particular, consider how each team developed a religiosity predictor variable. Both of our teams operationalized this variable on the basis of intuitive judgements: Team 039 operationalised religiosity by sum scoring four out of the nine religiosity items, picking the two "best" behaviour-focused items and the two "best" belief-focused items in order to achieve a sense of "balance" between behaviour and belief in the variable; while Team 175 operationalized religiosity by sum scoring all nine items, feeling that a comprehensive coverage of all available religiosity items was most appropriate. The fact that we made different decisions for operationalising this variable is not of serious concern (after all, a focus of many analyst projects is to examine the extent to which different analytic decisions influence results). However, what is of concern is that neither team assessed the construct validity of the religiosity predictor variable. This is potentially problematic because using predictor variables with poor construct validity can invalidate analyses (Slaney, 2017). This lacuna in our submissions could be interpreted as manifesting a "measurement-schmesurement" attitude that is prevalent in the social sciences, whereby important measurement issues are down-played or entirely ignored (Flake & Fried, 2020; Flake, Pek, & Hehman, 2017; Hussey & Hughes, 2020). In conducting background reading for this commentary (i.e., after submitting our pre-registrations and analyses) we learnt that (1) unjustified sum scoring can have adverse consequences (McNeish & Wolf, 2020) and that (2) a failure to achieve measurement invariance for psychometric instruments across cultures can throw the validity of analyses into doubt (Boer, Hanke, & He, 2018; Fischer & Karl, 2019; van de Vijver & Leung, 2021).

Why did we fail to attend to construct validity in our submissions? We suggest that the incentive structure of the MARP project (which is typical of analysts projects) may have contributed.[3] While developing our respective pre-registrations we could not help but be aware that our submission would be merely one among many and, thus, would have little discernible impact on the overall success

of the project. Given that our teams have expertise in multilevel modelling, it was not overly time-consuming (and was rather enjoyable!) for us to put careful consideration in the pre-registrations of our multilevel models. Thus, it is not surprising that this aspect of our analyses is strong (or so we hope!). By contrast, our teams have little experience with construct validation in the context of cross-cultural research, meaning that it would have necessitated a considerable investment of time and effort to study the relevant literature in order to develop pre-registrations that thoroughly addressed key issues. Consequently, it is not surprising that construct validation failed to feature in our submissions. Had there been direct incentives for us to carefully attend to construct validation we may well have done so.

We strongly suspect that our MARP project submissions were not the only ones with limitations that could plausibly be attributed to a lack of direct incentives. While submissions will have a variety of strengths and weaknesses, we anticipate that insufficient attention being paid to construct validation will be a recurring limitation (given the ubiquity of the aforementioned "measurement-schmesurement" attitude in the social sciences). In fact, we would like to make a prediction: very few analysis teams have made any attempt to assess the construct validity of the variables they used in their analyses.[4] If our prediction proves to be correct then we would hasten to emphasise that this should not be taken as support for the traditional single team approach because it has a poor track-record in terms of assessing the construct validity of measures. Rather, such a finding would provide evidence that the status quo of poor measurement practices in the social sciences is not ipso facto resolved by adopting a many analysts approach. Moreover, such a finding would suggest that an important direction for future work is to explore whether analysis teams can be incentivised to carefully consider all key dimensions of their submissions. This might involve developing procedures by which the strongest submissions are identified and the authors of these submissions are rewarded (by placing their names in a more prominent position in the author list; by inviting them to write a commentary about their submission; by giving them awards or cash prizes; etc). Alternatively (or in addition), this might involve incentivising researchers to form analysis teams that bring together diverse analytic skills and domain specific knowledge.

Notes

[1] This commentary was written by members of two analysis teams: Ross and Sulik (along with Tappin) are Team 039, and Buczny and Schivinski are Team 175.

[2] The term "construct validity" encompasses a variety of subcomponents that have been classified in different ways (Slaney, 2017), and not all subcomponents could be assessed using the materials available to us in the MARP project. However, at least three important subcomponents (some of which are classified as "reliability" rather than "validity" according to some measurement classifications) could have been examined using available data: internal consistency, factor structure, and measurement invariance across cultures.

[3] We say "may have" contributed because we are not confident that we have good introspective access to what motivated our analysis decisions.

[4] We invite the leaders of the MARP project to test our prediction by coding each submission according to whether or not there is evidence of an attempt to assess construct validity.

References

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: a review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology, 49*(5), 713-734. doi:10.1177/0022022117749042

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology, 10*, 1507. doi:10.3389/fpsyg.2019.01507

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456-465. doi:10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research. *Social Psychological and Personality Science, 8*(4), 370-378. doi:10.1177/1948550617693063

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 166-184. doi:doi.org/10.1177/2515245919882903

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

Silberzahn, R., & Uhlmann, E. L. (2015). Many hands make tight work. *Nature, 526*(7572), 189-191. doi:10.1038/526189a

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., . . . Nosek, B. A. (2018). Many analysts, one data set: making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337-356. doi:10.1177/2515245917747646

Slaney, K. (2017). *Validating psychological constructs: historical, philosophical, and practical dimensions*: Palgrave Macmillan.

van de Vijver, F. J. R., & Leung, K. (2021). *Methods and data analysis for cross-cultural research, 2nd ed*: Cambridge University Press.