When do open science practices lead to higher quality data?

Danka Purić[12], Iris Žeželj[12], Ljiljana B. Lazarević[23], Goran Knežević[12]
[1] *Department of Psychology, University of Belgrade, Belgrade, Serbia*
[2] *Laboratory for Research of Individual Differences, University of Belgrade, Belgrade, Serbia*
[3] *Institute of Psychology, University of Belgrade, Belgrade, Serbia*
{dpuric, izezelj, ljiljana.lazarevic, gknezevi}@f.bg.ac.rs

Abstract: Open science initiatives and practices are gaining almost universal support. For example, the registered report manuscript format, designed with the aim to increase transparency and quality of science, is starting to make an impact across different fields. The second line of action focuses on creating standards for sharing research materials, analysis scripts, and databases. Third, the pressure on publishers to make published manuscripts openly accessible is increasing. Finally, scientific collaborations set the standards in data collection and enable the collection of high-quality data. In this paper, we discuss the mechanisms by which these practices may improve the quality of scientific data and offer a critical perspective on their outcomes and effectiveness.

Keywords: open science practices; registered reports; open data; collaborations.

I. Introduction

In the past two decades, there has been a number of initiatives supporting open science, e.g., [1], [2], [3], [4], and the idea of openness seems to be almost undisputed and widely promoted [5]. This resulted in a list of radical changes in recommended research practices for empirical sciences aimed to discourage the so called questionable ones (for a brief overview see [6]). In this paper, we argue that, as with all innovations, academic community should not take the positive impact of open science practices for granted, but should instead scrutinize it and carefully consider their benefits, as well as potential risks and drawbacks (this same sentiment was voiced in, for example, [7]). To this end, we will delve into the mechanisms by which open science practices could benefit science, and, more specifically, lead to higher quality data and we will discuss the yet unresolved issues related to open science practices.

II. Registered reports

By pre-registering their study, researchers commit to specific hypotheses and analysis plans before collecting research data (as seen in [8][9]). In a rather straightforward manner, this practice prevents some of the most common questionable research practices, such as post hoc hypothesizing (HARK-ing: Hypothesizing After the Results are Known), searching for statistically significant results (p-hacking), selective reporting (reporting only on significant findings and omitting the non-significant ones), etc. In this way, pre-registering a study ensures that the database which will be the result of a study is complete and known in advance. Additionally, if the pre-registered plan is peer-reviewed before the data is collected it should also make sure that study samples are large enough to enable a valid test of the research hypothesis, i.e., that the studies have enough statistical power to detect the effect of certain size.

The design of the study is not the only output that can be peer-reviewed - instead of the fully completed manuscript, researchers can now submit the whole manuscript omitting only the exact results to solicit feedback from the experts in the field. In general, the practice of results-blind reviews should guarantee that a manuscript is evaluated on the basis of the relevance of the research question it poses, its methodological stringency, the soundness of the data analytic approach, and not the significant and “attractive” results it is expected to obtain. This practice makes it more probable for both significant (hypothesis confirming) and non-significant results to be published, thus helping future studies determine the focus of their research. If a finding is repeatedly proven to be non-replicable, this is an important signal for other researchers that their resources can be better used to explore different, more promising research lines and to collect databases with higher usability.

III. Sharing materials, databases, and scripts

Let us go through the most important arguments for the claim that sharing is the most obvious way to improve the quality of scientific data. Firstly, making the data public means that the results obtained on those data are verifiable - i.e., other members of the community can engage in its quality control and point out to eventual omissions or errors that can be corrected. This mere fact, in turn, increases the researcher’s responsibility to double-check both the data and all the results, consequently reducing the probability of both unintentional errors and questionable research practices. Sharing the data also requires the researcher to make the database readable to others (e.g., labeling raw variables, describing the scale and calculated scores) thus increasing the likelihood of it being (re)used (for example, see [10]). Moreover, open databases can be aggregated into large secondary databases. This allows for analyses that would not be possible on single-study data alone (one such attempt which is known to a wider public is the Gapminder Foundation, https://www.gapminder.org/).

Closely related to sharing data is sharing analysis scripts, as these two types of open resources complement each other. Sharing scripts further increases both the

verifiability and the communicability of research data, so it improves scientific practices in a way similar to that of sharing data. But more than that, by sharing their scripts, researchers make it easier for others to build upon their work, extend their analyses and generate secondary data (e.g., by proposing new indices calculated from the raw data). It also prevents loss of resources, as the researchers do not need to start from scratch but can build upon what has already been done - unfortunately, it is very common in science to have several independent teams of researchers working on the same tasks simultaneously or having to redo what has already been done but not shared. Additionally, sharing scripts encourages the use of free/open software, since it makes the script more accessible to a larger audience.

Finally, by sharing their research materials, researchers provide necessary information for understanding the scope and generalizability of the data they have collected and the results they have obtained. It is also a prerequisite for testing the replicability of findings, which is an important step in the development of any empirical science.

IV. Opening access to published manuscripts

Empirical analyses show a growing trend of open access publishing in the past four decades [11]. But how exactly is this contributing to the quality of science? The paywall is notorious for fueling a disparity between researchers depending on the number of resources available to them through the institutions they work in. Opening access to academic journals makes academic research widely available thus reducing these inequalities. While this is a valuable goal per se, having more researchers keeping up to date with the newest scientific findings also increases probabilities for scientific break-throughs, enhances the collaborative potential of scientists in deprived countries and supports cumulative science in general.

Researchers can now also share the pre-print versions of their manuscripts on one of the many designated websites (e.g., ArXiv, OSF, bioRxiv, PrePubMed) and thus recruit the help of the scientific community. After receiving comments from a number of peers, the researcher is bound to make better use of the collected data.

V. Scientific collaborations

Even though the sheer quantity of data does not guarantee its quality, collaborative efforts (whether they are widely known mega collaborations such as CERN [12] or less formalized ones such as Psychological Science Accelerator [13]), make it more likely that the data is being collected in an optimal way for testing a research hypothesis. Large databases that are created as a result of collaborative studies are high-powered for drawing conclusions and, in the case of human participants, most often include people from diverse cultural contexts or of different racial/ethnic backgrounds, which may be important for understanding the generalizability of findings (especially in, for example, medicine and psychology). Mega collaborations are also more likely to have an academic impact (for a review of how the number of citations relates to team size see [14]) and be visible in the mainstream media. They also enable knowledge sharing, especially through empowering the “weaker” partners to conduct future research, thus improving the quality of future data. In some instances, such as the Psychological Science Accelerator [13], collaborations also encourage a greater diversity of research topics, by employing a bottom-up approach to selecting projects that will be funded.

VI. Open questions on open science practices

Even though the advantages of open science practices are clear, we should be cautious about their potential to backfire, and for the practical issues they bring up. For example, relying on study materials shared by primary researchers may lead to an over-standardization, i.e., a lack of diversity of research data, which may put the generalizability of results in question. Some critiques voiced their concerns that focusing on only robust and replicable findings could discourage creative explorations that are supposed to be the “engine of science” [15][16], thus the recommended practice is now to clearly label confirmatory and exploratory analyses.

Regarding open data - special care must be taken to ensure that the data is properly anonymized so that no sensitive information on participants is disclosed and so that no participants can be identified based on their responses (see [10]). Full openness of data also allows non-experts to look into them and analyze them, which in some cases might lead to inaccurate conclusions or data interpretations.

Another issue is data ownership - as soon as the data is shared with the public, it is free to be used by anyone. Sometimes, substantial resources, e.g., government resources of a specific country taxpayers, such as GESIS panel in Germany [17] have gone into data collection, and yet the collected data is available to researchers from all over the world. However, this is a notable exception since, usually, the more time and effort had gone into data collection, the more the original authors are reluctant to share their data, at least immediately after collection/publication.

Finally, open science practices are cumbersome and can sometimes seem like an unnecessary bureaucratic burden that slows down the research and publishing process, thus making the open science research proponents less competitive. Therefore, the researchers should be made aware of all the benefits these practices bring but also incentivized to actually follow them. To this end, there have been a number of initiatives coming from academic publishers. For example, awarding badges for pre-registration, open data, and open materials has proven a good method for increasing transparency [3]. More importantly, following open science practices

should be viewed as an advantage, if not a prerequisite when evaluating projects or job applicants.

VII. Conclusion

This is most certainly not an exhaustive list of potential issues that open science faces. Many of the practices we have mentioned are rather new and we are still learning how to be open, so running into some issues seems inevitable. It is, therefore, crucial to be transparent about the decisions we make and to acknowledge the limitations of our practices.

References

[1] B. D. Borges, “Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities”, 2018

[2] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, pp. Aac4716, 2015

[3] Platforma za otvorenu nauku, http://www.mpn.gov.rs/wp-content/uploads/2018/07/Platforma-za-otvorenu-nauku.pdf [Accessed July. 15, 2019]

[4] M.C. Kidwell, L. B. Lazarević, E. Baranski, T.E. Hardwicke, S. Piechowski, L. S. Falkenberg, ... and T. M. Errington, “Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency," *PLoS biology*, vol 14(5), pp. e1002456, 2016

[5] B.A. Nosek, G. Alter, G.C. Banks, D. Borsboom, S. Bowman, S.J. Breckler, ... & M. Contestabile, "Promoting an open research culture," *Science*, vol 348, pp. 1422-1425, 2015.

[6] Lj. Lazarević and I. Žeželj, “How open science norms improve scientific practices”. *Proceedings of PSSOH conference*, 2018, pp. 13-15, doi: 10.5281/zenodo/1411157

[7] E. J. Finkel, P. W. Eastwick, and H. T. Reis, ”Replicability and other features of a high-quality science: Toward a balanced and empirical approach”. *Journal of Personality and Social Psychology*, vol 113(2), pp. 244, 2017

[8] H. IJzerman, S. M. Lindenberg, I. Dalğar, S. Weissgerber, R. C. Vergara, A. H. Cairo, … and J. H. Zickfeld, “The Human Penguin Project: Climate, Social Integration, and Core Body Temperature," *Collabra: Psychology*, vol 4(1), pp. 37. 2018, doi: http://doi.org/10.1525/collabra.165

[9] R. A. Klein, et al., "Many Labs 2: investigating variation in replicability across sample and setting," *Advances in Methods and Practices Psychological Science*, vol 1(4), pp. 443-490, 2018. https://doi.org/10.1177/2515245918810225

[10] C. P. Hu, J. Yin, S. Lindenberg, I. Dalğar, S. S. Weissgerber, R. C. Vergara, … and H. IJzerman, “Data from a cross-national project testing principles from social thermoregulation theory," *Scientific Data*, vol 6:32, 2019,. https://doi.org/10.1038/s41597-019-0029-2

[11] M. Laakso, P. Welling, H. Bukvova, L. Nyman, B. C. Björk, and T. Hedlund, “The development of open access journal publishing from 1993 to 2009," *PloS one*, vol 6(6), pp. e20961, 2011

[12] https://home.cern/ [Accessed Aug. 20, 2018]

[13] H. Moshontz, L. Campbell, C. R. Ebersole, H,. IJzerman, H. L. Urry, P. S. Forscher,..., and C. R. Chartier, "The Psychological Science Accelerator: Advancing Psychology through a Distributed Collaborative Network," *Advances in Methods and Practices in Psychological Science*, vol 1(4), 501-515, 2018

[14] D. Hsiehchen, M. Espinozaand A. Hsieh, “Multinational Teams and Diseconomies of Scale in Collaborative Research," *Science Advances*. vol 1(8), pp. e1500211, 2015

[15] R. F. Baumeister, and K. D. Vohs, “Misguided effort with elusive implications,” *Perspectives on Psychological Science*, vol 11(4), pp. 574-575, 2016

[16] D. T. Gilbert, G. King, S. Pettigrew, and T. D: Wilson, “Comment on “Estimating the reproducibility of psychological science,"" *Science*, vol. 351(6277), pp. 1037-1037, 2016

[17] M. Bosnjak, T. Dannwolf, T. Enderle, I. Schaurer, B. Struminskaya, and A. Tanner, “Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The GESIS Panel,” Social Science Computer Review, vol 23(1), pp. 103-115. 2015, http://journals.sagepub.com/doi/pdf/10.1177/0894439317697949