Industrial and Organizational Psychology (2021), **14**, 527–530
doi:10.1017/iop.2021.97

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

Raw data + analysis code > descriptive statistics

Cort W. Rudolph[1]* and Hannes Zacher[2]

[1]Saint Louis University and [2]Leipzig University
*Corresponding author. Email: cort.rudolph@health.slu.edu

The focal article by Murphy (2021) raises a number of excellent points about legacy statistical reporting practices that would be well-addressed by increasing transparency in the research process. Here, we argue that a complementary solution to many of the ills that were proposed by Murphy is to consider raw data[1] and analysis code as a "product" of the research process on par with the manuscript that describes one's research efforts. That is to say, many of the problems that are outlined in the focal article could be (at least partially) solved if journals required researchers to submit their raw data and analysis code as a supplement to their manuscript, which would be made publicly available upon publication. Murphy specifically notes three issues that are associated with increasing complexity and diversity of data-analytic methods in organizational research: (a) incorrect applications and interpretations of analyses, (b) increasing reliance on significance testing, and (c) increasing difficulty in interpretation that widens the gaps between science and practice. We argue here that requiring open sharing of raw data and analysis code is at least a partial remedy for each of these issues and would also have a much broader benefit to the industrial-organizational (I-O) psychology literature in terms of increasing the transparency (and thus the credibility) of our science.

To the first point, regarding the potential for incorrect applications and interpretations of analyses, we argue that one antidote to such errors is to provide readers with the raw data and the analysis code that gave rise to the analyses and corresponding results in the first place. Indeed, to the extent that readers can triangulate the results that are reported in manuscripts to the raw data and analysis code that produced them, more confidence in the conclusions of the research can be instilled. Also, the potential to catch errors and/or fraud in analyses during the review process would be increased to the extent that editors and reviewers have access to this information when making evaluations of such work (Simonsohn, 2013). Ultimately, with raw data and analysis code in hand, anyone can reproduce Table 1 (and Table 2, Table 3 … Table *k*).

To the second point, regarding the limitations of (and overreliance on) null hypothesis significance testing (NHST), sharing raw data and analysis code helps to overcome two related issues. First, it facilitates meta-analytic techniques, which generally focus on effect size estimation and to some extent eschew NHST logic (e.g., Hunter & Schmidt, 2004; Schmidt & Hunter, 2002). Regarding meta-analysis, it is true that such analyses are typically undertaken with summary data. However, we should point out that (a) it is often difficult to extract the necessary information from summary data alone to compute appropriate effect sizes (e.g., Rudolph & Jundt, 2017) and (b) there are more advanced meta-analytic procedures that would certainly benefit from access to both raw and summary data (e.g., meta-analytically testing nonlinear associations; see Katz

Cort W. Rudolph, Department of Psychology, Saint Louis University, St. Louis, MO (USA). Hannes Zacher, Institute of Psychology–Wilhelm Wundt, Leipzig University, Leipzig, Germany.

[1]We use the term "raw data" rather broadly here to refer to data that have *not* been aggregated into summary statistics, such as means and standard deviations (i.e., "summary" data).

et al., 2019). Second, sharing raw data facilitates "mega-analysis" (also called "integrative data analysis," Curran & Hussong, 2009; Eisenhauer, 2021) of pooled raw data sets of the same phenomena across multiple samples; a "high-*n*" and thus "high-powered" technique that can help avoid certain pitfalls of NHST logic.

Finally, to the third point about interpretation fueling science–practice gaps, open sharing of raw data and analysis code empowers consumers of science on both sides of the divide to take action to answer their own questions about the results that are being presented; it facilitates the broad translation of increasingly complex analyses to consumers of research across different roles (i.e., students, researchers, and practitioners—but also editors and reviewers) by providing a reproducible pipeline from data to code to manuscript.

Implementing these ideas will require a few things, not least of which is changing norms about what are the "products" of the research process. To do this, first we need to normalize the expectation that papers will include online supplemental appendices with raw data and analysis code. Indeed, various journals in our field already (at least tacitly) encourage this. For example, in its instructions for authors, the *Journal of Applied Psychology* suggests, "We recommend sharing data and materials via trusted repositories." Similarly, the *European Journal of Work and Organizational Psychology* suggests, "Authors are encouraged to share or make open the data supporting the results or analyses presented in their paper."

Second, the move toward open sharing of raw data and analysis code will require a willingness of editors and reviewers to do "extra work" by looking at and evaluating these materials as part of the peer review process. However, it could be argued that this will also save time in the long run, as papers with shoddy efforts toward raw data and analysis code sharing would be quickly dispensed with in favor of those whose authors have made the effort to do so. Moreover, these suggestions open up the potential for new editorial roles focusing on raw data and code review (i.e., the "methods editor" as a stand-alone editorial role).

Finally, arguably, moving toward open access to raw data and analysis code requirements would necessitate a broader set of knowledge and skills that goes beyond just understanding complex analyses, particularly an understanding of complex raw data and analysis code structures. To this end, we would argue that these are contemporary competencies that are required of top researchers, reviewers, and editors, as well as consumers of research (including scientists and practitioners).

We can also anticipate push back on these ideas. For example, the argument could be offered that one cannot share their raw data and/or analysis code for various reasons (e.g., the raw data contains identifying information; the code contains proprietary algorithms). A couple of rejoinders here bear consideration. First, anonymized raw data can readily be shared and, depending on the models considered, it may be the case that summary statistics (e.g., correlation/covariance matrices) may suffice in place of raw data (e.g., to reproduce simple regression or path analyses). Second, there are advanced tools available to researchers to create anonymous versions of data sets that maintain the original structure of one's data (e.g., the "synthpop" package for R; Nowok et al., 2016). Finally, with the advent and use of more advanced statistical analysis techniques, researchers should question the reproducibility of studies based on proprietary analysis methods and demand open-source methodologies whenever possible.

Moreover, it could also be rightly argued that requiring open access to raw data and analysis code does not solve all of the issues that are associated with research misconduct; for example, it would not necessarily curtail people from creatively maneuvering their data before it is shared. We contend, however, that it is arguably easier to spot manipulated raw data than manipulated summary data (e.g., for an exceptionally notable example, see Levelt et al., 2012). Moreover, it is almost certainly easier to spot errors in analysis code than in consequent results derived therefrom (e.g., Poldrack & Poline, 2015).

Another possible critique of our suggestions is the possibility that others may use or "scoop" one's data (i.e., use without credit or attribution). Although this issue has been debated variously

https://doi.org/10.1017/iop.2021.97 Published online by Cambridge University Press

in the open-science community (e.g., Bishop, 2015; Laine, 2017), we would argue that in the rare case in which this occurs, this is more of an issue for the "scoop-er" than the "scoop-ee." Somewhat related, there is also the potential fear that opening up raw data and analysis code would invite reanalysis that could debunk one's original claims (e.g., if substantial errors are found in the way in which data has been analyzed). However, we would argue that this is just "good science" and as such is a feature (not a "bug") of our proposition (e.g., Silberzahn et al., 2014).

Finally, and related to several of the points we raise here, it could be argued that requiring open sharing of raw data and analysis code would add "hurdles" or levels of bureaucracy to the research process. To this point, we would suggest that the value of psychological research has to some extent suffered from issues of credibility, stemming in part from a lack of transparency at various levels (see Rudolph, 2021). Arguably, the effort required to remove poorly conducted research from the literature is much higher than preventing such work from ever making it into our journals. As such, the suggestions we have offered represent a much easier means of curtailing "bad science" (while encouraging "good science") in the first place.

In closing, we ask the question, "How can we encourage the open sharing of raw data and analysis code?" Beyond the requirements that are laid out by journals to this end (which, so far, have come in the form of "encouragements"), there are creative ways to incentivize this practice. Indeed, it is often a requirement of funding agencies that data be eventually deposited in public repositories (e.g., ICPSR; https://www.icpsr.umich.edu). However, more immediately, programs such as the Center for Open Science's "badges initiative" (https://www.cos.io/initiatives/badges) give credit to authors who adopt open science practices, including the sharing of raw data and analysis code, by awarding "badges" that are displayed graphically at the top of published research articles. Despite widespread adoption in other fields of psychology, to our knowledge, no journals in I-O psychology currently participate in this initiative. By no means are the suggestions offered here a perfect answer to each of the challenges that are raised by Murphy (2021). However, we would argue that moving toward more open and transparent research practices in I-O psychology, especially open sharing of raw data and analysis code, would go a long way toward addressing these challenges. We hope that researchers and journal editors will take us up on these suggestions and look forward to a future in which I-O psychology research (i.e., manuscripts, but also supporting raw data and analysis code) is more accessible to all.

References

Bishop, D. (2015, December 15). Who's afraid of open data: Scientists' objections to data sharing don't stand up to scrutiny. *LSE Impact Blog*. https://blogs.lse.ac.uk/impactofsocialsciences/2015/12/16/whos-afraid-of-open-data-dorothy-bishop/

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, **14**(2), 81–100. https://doi.org/10.1037/a0015914

Eisenhauer, J. G. (2021). Meta-analysis and mega-analysis: A simple introduction. *Teaching Statistics*, **43**(1), 21–27. https://doi.org/10.1111/test.12242

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. SAGE.

Katz, I. M., Rudolph, C. W., & Zacher, H. (2019). Age and career commitment: Meta-analytic tests of competing linear versus curvilinear relationships. *Journal of Vocational Behavior*, **112**, 396–416. https://doi.org/10.1016/j.jvb.2019.03.001

Laine, H. (2017). Afraid of scooping–case study on researcher strategies against fear of scooping in the context of open science. *Data Science Journal*, **16**, 29. http://doi.org/10.5334/dsj-2017-029

Levelt, P., Noort, E., & Drenth, P. (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Tilburg University. https://www.rug.nl/about-ug/latest-news/news/archief2012/nieuwsberichten/stapel-eindrapport-eng.pdf

Murphy, K. (2021). In praise of Table 1: The importance of making better use of descriptive statistics. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **14**(4), 461–477.

Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. Journal of Statistical Software, 74(11), 1–26. https://doi.org/10.18637/jss.v074.i11

Poldrack, R. A., & Poline, J. B. (2015). The publication and reproducibility challenges of shared data. *Trends in Cognitive Sciences*, **19**(2), 59–61. https://doi.org/10.1016/j.tics.2014.11.008

Rudolph, C., & Jundt, D. (2017, July 18). Why betas should not rule metas. *PsyArXiv*. https://doi.org/10.31234/osf.io/jacdy

https://doi.org/10.1017/iop.2021.97 Published online by Cambridge University Press

Rudolph, C. W. (2021). Improving careers science: Ten recommendations to enhance the credibility of vocational behavior research. *Journal of Vocational Behavior*, 126, Article 103560. https://doi.org/10.1016/j.jvb.2021.103560

Schmidt, F., & Hunter, J. (2002). Are there benefits from NHST? *American Psychologist*, **57**(1), 65–66. https://doi.org/10.1037/0003-066X.57.1.65

Silberzahn, R., Simonsohn, U., & Uhlmann, E. L. (2014). Matched-names analysis reveals no evidence of name-meaning effects: A collaborative commentary on Silberzahn and Uhlmann (2013). *Psychological Science*, **25**(7), 1504–1505. https://doi.org/10.1177/0956797614533802

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, **24**(10), 1875–1888. https://doi.org/10.1177/0956797613480366

Cite this article: Rudolph, CW. and Zacher, H. (2021). Raw data + analysis code > descriptive statistics. *Industrial and Organizational Psychology* **14**, 527–530. https://doi.org/10.1017/iop.2021.97