



The Growing Importance of Reproducibility and Responsible Workflow in the Data Science and Statistics Curriculum

Nicholas J. Horton, Rohan Alexander, Micaela Parker, Aneta Piekut & Colin Rundel

To cite this article: Nicholas J. Horton, Rohan Alexander, Micaela Parker, Aneta Piekut & Colin Rundel (2022) The Growing Importance of Reproducibility and Responsible Workflow in the Data Science and Statistics Curriculum, Journal of Statistics and Data Science Education, 30:3, 207-208, DOI: [10.1080/26939169.2022.2141001](https://doi.org/10.1080/26939169.2022.2141001)

To link to this article: <https://doi.org/10.1080/26939169.2022.2141001>



© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 18 Nov 2022.



Submit your article to this journal [↗](#)



Article views: 688



View related articles [↗](#)

The Growing Importance of Reproducibility and Responsible Workflow in the Data Science and Statistics Curriculum

Modern statistics and data science uses an iterative data analysis process to solve problems and extract meaning from data in a reproducible manner. Models such as the PPDAC (Problem, Plan, Data, Analysis, Conclusion) Cycle (n.d) have been widely adopted in many secondary and post-secondary classrooms (see the review by Lee et al. 2022). The importance of the data analysis cycle has also been described and reinforced in guidelines for statistics majors (ASA Curriculum Guidelines 2014), undergraduate data science curricula (ACM 2021), and in data science courses and teaching materials (e.g., Wickham and Golemund 2022).

In 2018, the National Academies of Science, Engineering, and Medicine’s “Data Science for Undergraduates” consensus study (NASEM 2018) broadened the definition of the data analysis cycle by identifying the importance of workflow and reproducibility as a component of data acumen needed in our graduates. The report noted that “documenting, incrementally improving, sharing, and generalizing such workflows are an important part of data science practice owing to the team nature of data science and broader significance of scientific reproducibility and replicability.” The report also tied issues of reproducibility and workflow to the ethical conduct of science.

The importance of others being able to have confidence in our findings is built into the foundations of statistics and data science (Parashar, Heroux, and Stodden 2022). For instance, in theoretical research, theorems are introduced along with their proof. As statistics has changed to rely more on computational methods, innovation is needed to ensure that the same level of rigor characterizes claims based on data and code. Efforts to foster reproducibility in science (NASEM 2019; Parashar, Heroux, and Stodden 2022) and to accelerate scientific discoveries (NASEM 2021) have highlighted the importance of reproducibility and workflow within the broader scientific process.

Robust workflows matter. For instance, COVID-19 counts in the United Kingdom were underestimated because the way that Excel was used resulted in dropped data (Kelion 2020). The economists Carmen Reinhart and Kenneth Rogoff made Excel errors that resulted in miscalculated GDP growth rates (Herndon, Ash, and Pollin 2014). Cut and paste errors are all too common in many workflows (Perkel 2022). The reproducibility crisis that was first identified in psychology is now known to afflict much of the physical and social sciences. Steps taken to address this crisis, including improved reporting of methods, code and data sharing, version control, are increasingly com-

mon (Munafò et al. 2017), but workflow and reproducibility issues can be subtle.

Where does this leave statistics and data science education? Many instructors incorporate their research in teaching and research-led teaching (Schapper and Mayson 2010). It is often neglected, however, that most time in research involving data science and statistics is spent on tidying up, documenting data provenance, improving group collaboration and sharing, anonymizing data, and creating analytic datasets as well as repository and replication files. While it is widely advocated that open science practice should be embedded in everyday research practice (Sandve et al. 2013), it is less clear why on the pedagogy side we should give most of the classroom attention to the middle part of the project—data analysis—and fail to incorporate the entire data analysis cycle (De Veaux et al. 2017; Wickham and Golemund 2022). Future scientists need to have multiple opportunities to undertake the entire data analysis cycle with real data and appropriate workflows. Teaching about reproducibility and workflow aligns well with the research-led approach.

However, there are many barriers that make it challenging to incorporate sound workflow and reproducible analysis into our courses and programs. These include the rapid and constant evolution of technology and tools, the minimal training that most instructors received in the use of reproducible methods, the lack of well-established best practices, the paucity of vetted and inclusive curricular materials, and minimal comprehension of important aspects of student understandings (or misunderstandings) when we teach about the data analysis cycle, workflows, and reproducibility.

To highlight work in this important and developing area, the *Journal of Statistics and Data Science Education* invited papers related to “Teaching reproducibility and responsible workflow.” The November, 2022 issue of the journal is devoted to this important topic. We are excited at what the community brought forward in these 11 papers. It’s our hope that the collected papers in this issue and related resources provide motivation, guidance, and examples that complement prior published work (see, e.g., Çetinkaya-Rundel and Ellison 2021; Smith, Yu, and Schmid 2021).

Integrating reproducibility into our practice and our teaching can seem intimidating initially (Monajemi et al. 2019). One way forward is to start small. Make one small change to add an element of exposing students to reproducibility in one class, then

make another the next semester. Our students can get much of the benefit of reproducible and responsible workflows even if we just make a few small changes in our teaching. These efforts will help them to make more trustworthy insights from data. If it leads, by way of some virtuous cycle, to us improving our own practice, then even better! Improving our teaching through providing curricular guidance about reproducible science will take time and effort that should pay off in the long term.

We look forward to seeing your innovations, and welcome future submissions on these important issues.

Beyond these papers, the interactions between the reviewers, associate editors, and authors over the past year and a half have led us to review policies and expectations for data and code sharing for the journal. As a result of these discussions and deliberations, the *Journal of Statistics and Data Science Education* has joined other American Statistical Association journals (e.g., JASA 2022) that require submission of code, data, and the workflow to reproduce the paper. New submissions to the journal now require a data availability statement from authors that indicate how the data and code that are associated with a paper have been made accessible. These resources have the potential to be helpful for the review process and for readers of that paper—that is, will allow reusing or adapting them to other courses. We believe that these policies, which are consistent with broader guidance about data and code sharing (e.g., Nosek et al. 2016), will be of benefit for the entire community.

Acknowledgments

We thank Micaela Parker for her work as one of the guest editors for the issue as well as the many reviewers who provided helpful feedback as part of the editorial review.




ORCID

Nicholas Horton  <https://orcid.org/0000-0003-3332-4311>
 Rohan Alexander  <https://orcid.org/0000-0003-1279-0700>
 Aneta Piekut  <https://orcid.org/0000-0002-3478-0354>
 Colin Rundel  <https://orcid.org/0000-0002-6058-8251>

References

- ACM. (2021), “Computing Competencies for Undergraduate Data Science Curricula”, Available at https://www.acm.org/binaries/content/assets/education/curricula-recommendations/dstf_ccdsc2021.pdf.
- American Statistical Association. (2014), “Curriculum Guidelines for Undergraduate Programs in Statistics”, Available at <https://www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx>.
- Çetinkaya-Rundel, M., and Ellison, V. (2021), “A Fresh Look at Introductory Data Science”, *Journal of Statistics and Data Science Education*, 29, S16–S26. <https://doi.org/10.1080/10691898.2020.1804497>.
- De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., Tiruvilumala, N., Uhlig, P. X., Washington, T. M., Wesley, C. L., White, D., and Ye, P. (2017), “Curriculum Guidelines for Undergraduate Programs in Data Science”, *Annual Review of Statistics and Its Application*, 4, 15–30. <https://doi.org/10.1146/annurev-statistics-060116-053930>.

- Herndon, T., Ash, M., and Pollin, R. (2014), “Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff”, *Cambridge Journal of Economics*, 38, 257–279. <https://doi.org/10.1093/cje/bet075>.
- JASA. (2022), “The Journal of the American Statistical Association (JASA) Reproducibility Guide”, Available at <https://jasa-acsgithub.io/repro-guide>.
- Kelion, L. (2020), “Excel: Why Using Microsoft’s Tool Caused Covid-19 Results to be Lost”, BBC News, 5 October, Available at <https://www.bbc.com/news/technology-54423988>.
- Lee, H., Mojica, M., Thrasher, E., and Baumgartner, P. (2022), “Investigating Data Like a Data Scientist: Key Practices and Processes”, *Statistics Education Research Journal*, 21, 1–23. <https://doi.org/10.52041/serj.v21i2.41>.
- Monajemi, H., Murri, R., Jonas, E., Liang, P., Stodden, V., and Donoho, D. (2019), “Ambitious Data Science Can Be Painless”, *Harvard Data Science Review*, 1. <https://doi.org/10.1162/99608f92.02ffc552>.
- Munafò, M., Nosek, B., Bishop, D., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. A. (2017), “A Manifesto for Reproducible Science”, *Nature Human Behavior*, 1, 0021. <https://doi.org/10.1038/s41562-016-0021>.
- National Academies. (2018), “Data Science for Undergraduates: Opportunities and Options”, Available at <https://nap.nationalacademies.org/read/25104>.
- National Academies. (2019), “Reproducibility and Replicability in Science”, Available at <https://nap.nationalacademies.org/read/25303>.
- National Academies. (2021), “Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop”, Available at <https://nap.nationalacademies.org/read/26532>.
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., Yarkoni, T., Stodden, V., DeHaven, A. C. (2016), “Transparency and Openness Promotion (TOP) Guidelines”, Available at <https://osf.io/vj54c>.
- Parashar, M., Heroux, M. A., and Stodden, V. (2022), “Research Reproducibility”, *Computer Magazine*, 55, 16–18. <https://doi.org/10.1109/MC.2022.3176988>.
- Perkel, J. M. (2022), “An End to Copy-and-Paste Errors”, *Nature*, 603, 191–192. <https://media.nature.com/original/magazine-assets/d41586-022-00563-z/d41586-022-00563-z.pdf>.
- PPDAC: The Problem-Plan-Data-Analysis-Conclusion Cycle. (n.d), Available at <https://dataschools.education/about-data-literacy/ppdac-the-data-problem-solving-cycle>.
- Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013), “Ten Simple Rules for Reproducible Computational Research”, *PLoS Computational Biology*, 9, e1003285. <https://doi.org/10.1371/journal.pcbi.1003285>.
- Schapper, J., and Mayson, S. E. (2010), “Research-led Teaching: Moving from a Fractured Engagement to a Marriage of Convenience”, *Higher Education Research & Development*, 29, 641–651. <https://doi.org/10.1080/07294360.2010.489236>.
- Smith, L. M., Yu, F., and Schmid, K. K. (2021), “Role of Replication Research in Biostatistics Graduate Education”, *Journal of Statistics and Data Science Education*, 29, 95–104. <https://doi.org/10.1080/10691898.2020.1844105>.
- Wickham, H., and Grolemund, G. (2022), *R for Data Science: Visualize, Model, Transform, Tidy, and Import Data*, O’Reilly, Available at <https://r4ds.had.co.nz>.

Nicholas J. Horton , Rohan Alexander , Micaela Parker,
 Aneta Piekut , and Colin Rundel 
 Guest Editors