Open Science Practices: Reproducibility

What it is and why to practice it

Daniela Palleschi

Leibniz-Zentrum Allgemeine Sprachwissenschaft

Wed Sep 25, 2024

What is Open Science?

> "Open science" is an umbrella term used to refer to the concepts of openness, transparency, rigor, reproducibility, replicability, and accumulation of knowledge, which are considered fundamental features of science"

— Crüwell et al. (2019), p.3

- a movement developed to respond to crisis in scientific research
  - lack of accessibility, transparency, reproducibility, and replicability of previous research
- transparency is key to all facets of Open Science
  - it allows for full evaluation of all stages of science
- Open Access, software, data, code, materials…

Systemic problem in science

- the combination of
  - publication bias
    - journals favour novel, significant findings
  - publish or perish
    - researchers' careers depend on publications
- can/does/did lead to:
  - HARKing
    - Hypothesising After Results are Known
  - p-hacking
    - (re-)running analyses until a significant effect is found
  - replication crisis
    - pervasive failure to replicate previous research

Why do Open Science?

- open science is good science

- it encourages organisation and planning

  - helpful for future you

- increases *transparency*

- makes our work more robust

  - so future work stands on solid ground

- not all-or-nothing

- there are things I consider the bare minimum

  - detailed experiment plan, ideally public

  - openly available materials (e.g., stimuli)

  - share code and data

- the important thing is to do what you can

Figure 1. Open Science research practices across the research cycle
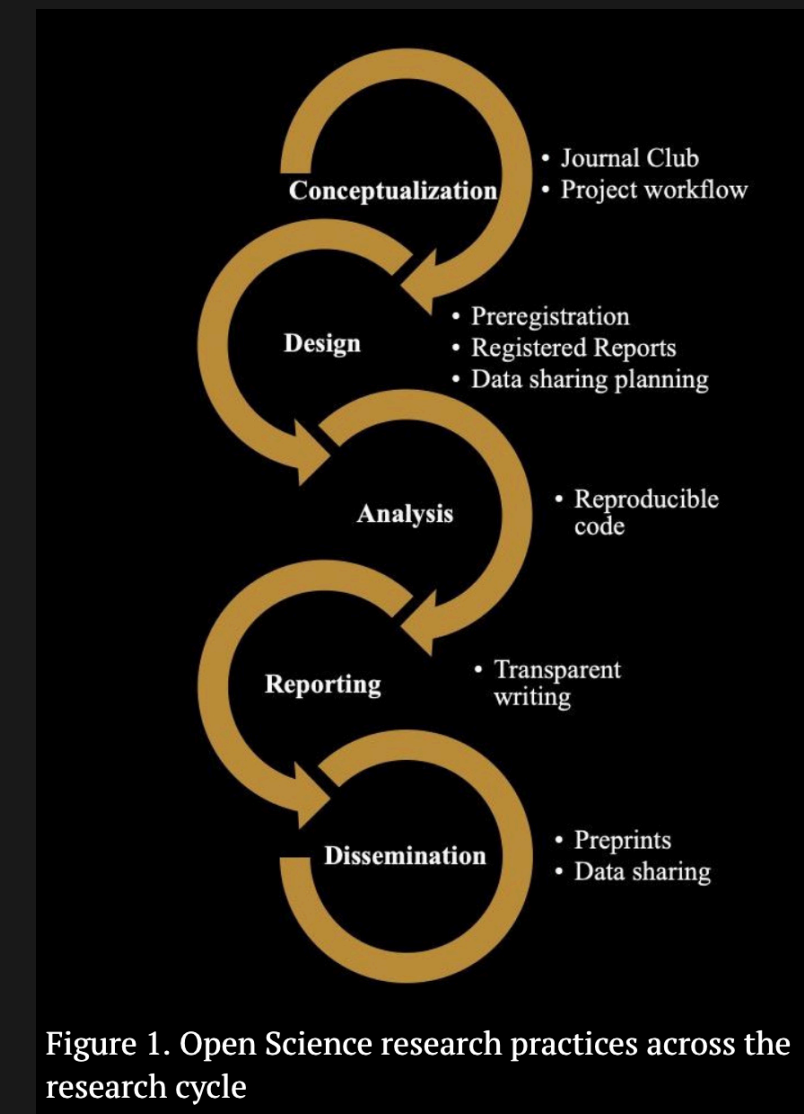
What is reproducibility?

- one piece of the Open Science pie

- generating the same results with the same data and analysis scripts

- seems obvious, but requires organisation and forethought before and during data collection/analysis

- bare minimum: share the code and the data (Laurinavichyute et al., 2022)

Reproducibility vs. replication

- the two terms have been used interchangably in the past (e.g., in the title of Open Science Collaboration, 2015)
  - we'll define them as follows (and this is becoming the standard distinction, imo)

Reproducibility

- re-analysing the same data using (ideally) the same scripts, software, etc
- aim: produce the same results (means, model estimates, etc.)
- why: tests for errors, coding mistakes, biases, etc.

- in short:
  - reproducibility = re-analysis of the *same data*
 - replication = collection of *new data*

Replication

- re-running a previous experiment, ideally with the same materials, set-up, etc.
- ideally the same analysis workflow as the original study (i.e., like reproducing the analyses but with new data)
- aim: test whether results are replicated with new data in terms of direction and magnitude

Why implement reproducibility in my workflow?

- firstly: the help future you (or collaborators/other researchers)!
    - you may return to your analyses tomorrow, next month, or next year
- to ensure robustness and to document your steps
    - 'researcher degrees of freedom' and the 'garden of forking paths': there's more than one way to analyse a certain dataset
    - we can try to plan ahead in detail (e.g., pre-reigster your analysis plan), but there will always be decisions made that were not foreseen
- lastly: it makes your life *much* easier and streamlines your workflow

The reproducibility spectrum

- reproducibility is on a continuum, referred to as the *reproducibility spectrum* in Peng (2011) (Figure 2)

  - *linked* means "*all data, metadata, and code [is] stored and linked with each other and with corresponding publications*" (Peng, 2011, p. 1227)

  - *executable* is not explained, and is more difficult to guarantee long-term as it depends on software versions

  - but at minimum we can assume it refers to code running on someone else's machine

Fig. 1. The spectrum of reproducibility.

Figure 2: Source: Peng (2011)
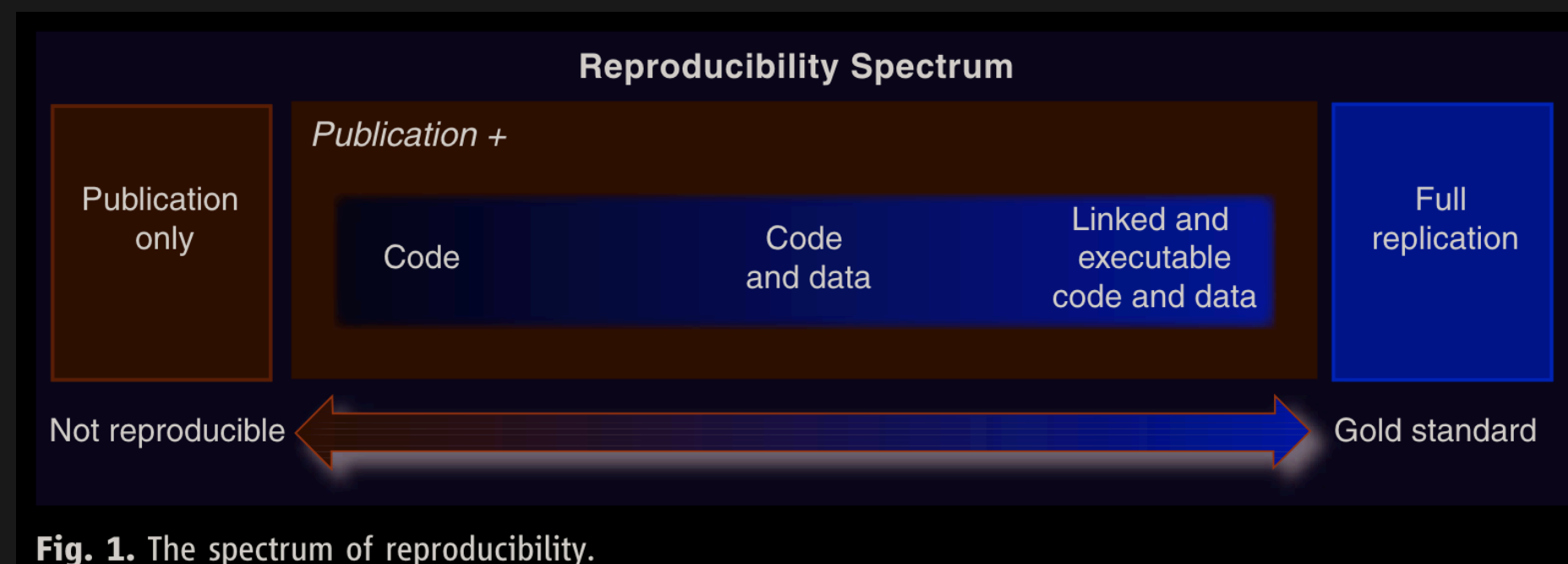
Steps we'll take

1. Open source software:

- R, an open source statistical programming language

- in RStudio, an IDE (integrated developer environment)

- with R Projects

2. Project-oriented workflow:

- establish folder structure

- and file/variable naming conventions

- use project-relative filepaths with the `here` package

- establish and maintain project-relative package library with `renv` (time permitting)

3. Practice literate programming:

- writing clean, commented, linear code

- in dynamic reports (e.g., Quarto, R markdown)

- practice modularity, i.e., 1 script = 1 purpose

4. Sharing and checking our code

- uploading our code and data to an OSF repository

- conducting a code review

References

Crüwell, S., Van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., Orben, A., Parsons, S., & Schulte-Mecklenbeck, M. (2019). Seven Easy Steps to Open Science: An Annotated Reading List. *Zeitschrift Für Psychologie*, *227*(4), 237–248. https://doi.org/10.1027/2151-2604/a000387

Kathawalla, U.-K., Silverstein, P., & Syed, M. (2021). Easing Into Open Science: A Guide for Graduate Students and Their Advisors. *Collabra: Psychology*, *7*(1), 18684. https://doi.org/10.1525/collabra.18684

Laurinavichyute, A., Yadav, H., & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language*, *125*, 12.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847