Reproducibility

Principles and Practice

Daniela Palleschi

2024-04-30

Table of contents

Learning Objectives

Today we will learn about…

Resources

- this lecture covers

1 The replication crisis

2 Reproducibility

- generating the same results with the same data and analysis scripts
  - seems obvious, but requires organisation and forethought
- bare minimum: share the code and the data (Laurinavichyute et al., 2022)
- rates of reproducibility across fields (Bochynska et al., 2023)
  - open access: 25-65%
  - data and analyses sharing: 11-33%
  - pre-registrations: 0-3%
- Journal of Memory and Language (JML) (Laurinavichyute et al., 2022)
  -

2.1 What should (ideally) be shared?

- materials
  - stimuli
  - experiment set-up
- documentation
  - README
  - metadata

- data
  - raw
    - e.g., text files, audio, video, or images
  - processed
- analysis code
  - pre-processing
  - analyses
- materials are helpful for replication
  - but also for inspection of e.g., design
- necessary for reproducibility
  - along with proper documentation of software used

2.2 Linguistic research

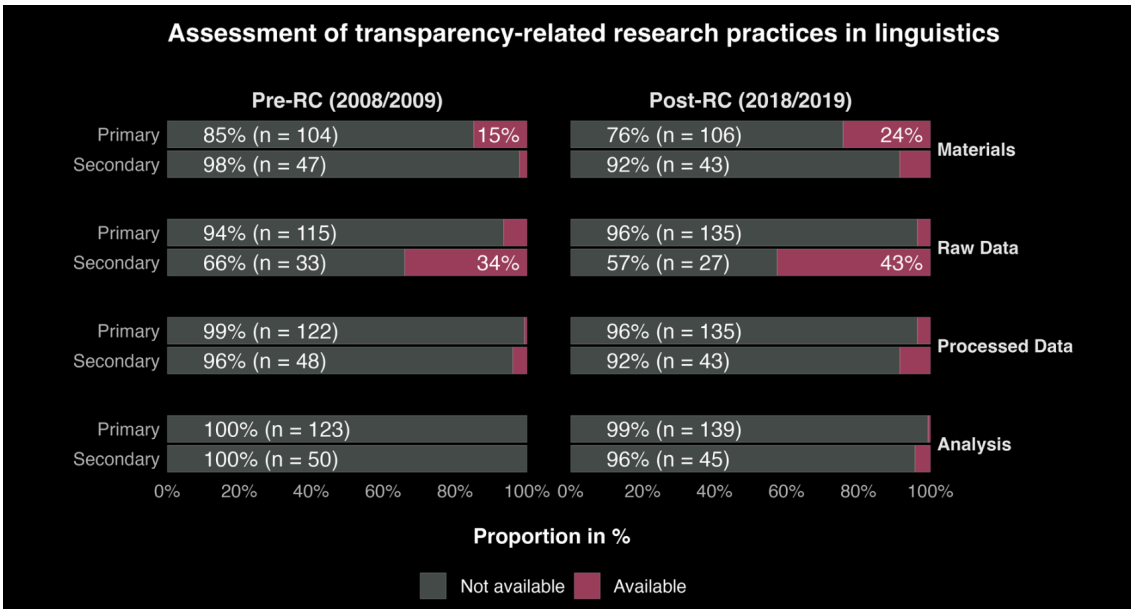

Figure 2: Percentages of the available and not available materials, raw data, processed data, and analysis scripts for the pre-RC (left) and post-RC (right) time windows, displayed separately for primary data (Primary) and secondary data (Secondary), for the empirical study articles in the sample. The Other category was excluded.

Figure 1: Source: Bochynska et al. (2023), p. 11 (all rights reserved)

- meta-analysis of 519 randomly sampled articles from various linguistic journales
  - pre- and post-reproducibility crisis (2008/9, 2018/19) (Bochynska et al., 2023)
  - differentiated between primary (collected for study) and secondary (pre-existing) data
- found a slight increase in shared materials, data, and analyses
  - but still low rates of each
- higher rates of secondary data sharing, presumably due to publicly available corpora

2.3 Journal of Memory and Language

- meta-analysis of articles from JML (Laurinavichyute et al., 2022)
  - before and after an Open Science Policy was introduced in 2019

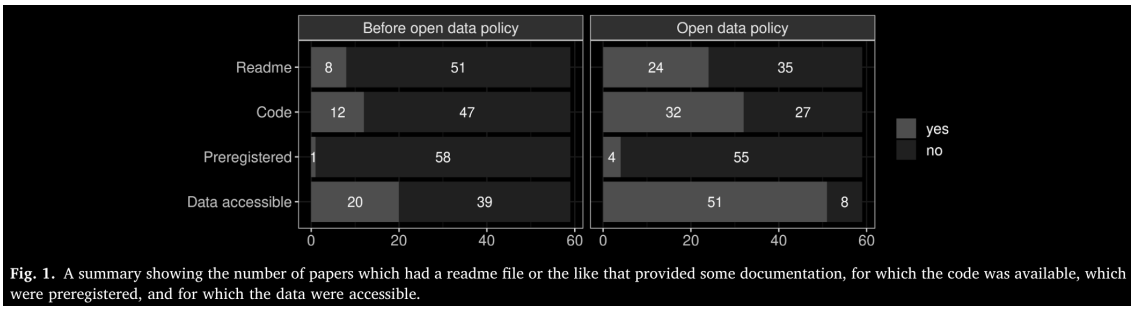

Fig. 1. A summary showing the number of papers which had a readme file or the like that provided some documentation, for which the code was available, which were preregistered, and for which the data were accessible.

Figure 2: Source: Laurinavichyute et al. (2022), p. 5 (all rights reserved)

- code and data availability improved
- but reproducibility rate ranged from 34-56%, depending on criteria
- higher rates compared to field-wide meta-analysis (Bochynska et al., 2023)

3 FAIR data

- refers broadly to data, but we'll consider it in terms of analyses
- findable and accesssible refer to where materials are stored
  - in *findable* repositories
  - that are *accessible*, i.e., do not require an account
- interoperable and reusable emphasis the format of data (and code)

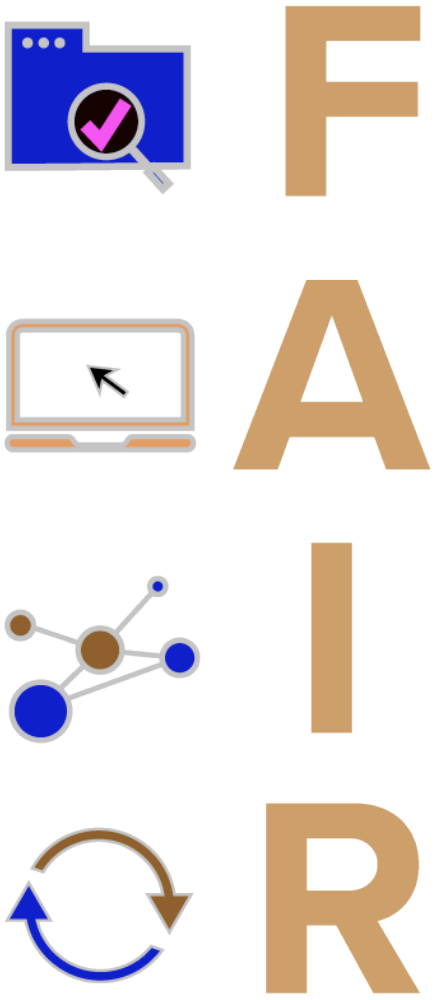

Figure 3: Source: National Library of Medicine (all rights reserved)

  - the importance of future use
  - and use beyond your precise computational environment

- a great way to test the FAIR principles
  - code review!
  - i.e., have a colleague try to access your data/run your code
    * either via an online repository
    * or send them your project folder

- Findable
  - refers to data and supplementary materials
  - persistant and unique identifer
  - relevant metadata

- Accessible
  - human-readable
  - available on a trusted repository, e.g., the OSF

- Interoperable
  - not dependent on an operating system
  - nor entirely on software/package versions

- Reusable
  - data should be reusable for future research
  - we can swap with 'reproducible' in the context of analyses

3.1 Findable

- materials should have a "persistant identifier"
  - e.g., Digital Object Identifier (DOI) for scholarly articles

- a digital, long-term storage of data
  - *not* on a personal or professional website
  - GitHub files don't typically have sufficient metadata
  - ideally: OSF, Zenodo or some other repository

- in recent papers, an OSF link is typically provided
- also: *discoverable*
  - e.g., in data-specific search engines (Google's Dataset search)

3.2 Accessible

3.3 Interoperable

3.4 Reusable

3.5 Task: finding data

Go to datasetsearch.research.google.com/

- do a search for data related to a topic of interest to you
- what type of information does the search provide?
- what type of links?
- do you find analysis code, or just data?
- do the same search at osf.io
- and at zenodo.org/
  - are there the same amount of hits?

4 Data and code availability

- "data available upon (reasonable) request"
  - generally not true
- data was not available in 68% of the most cited psychology studies (2006-2016) (Hardwicke & Ioannidis, 2018)
  - a further 18% were available with restrictions
  - only 11% available without restriction

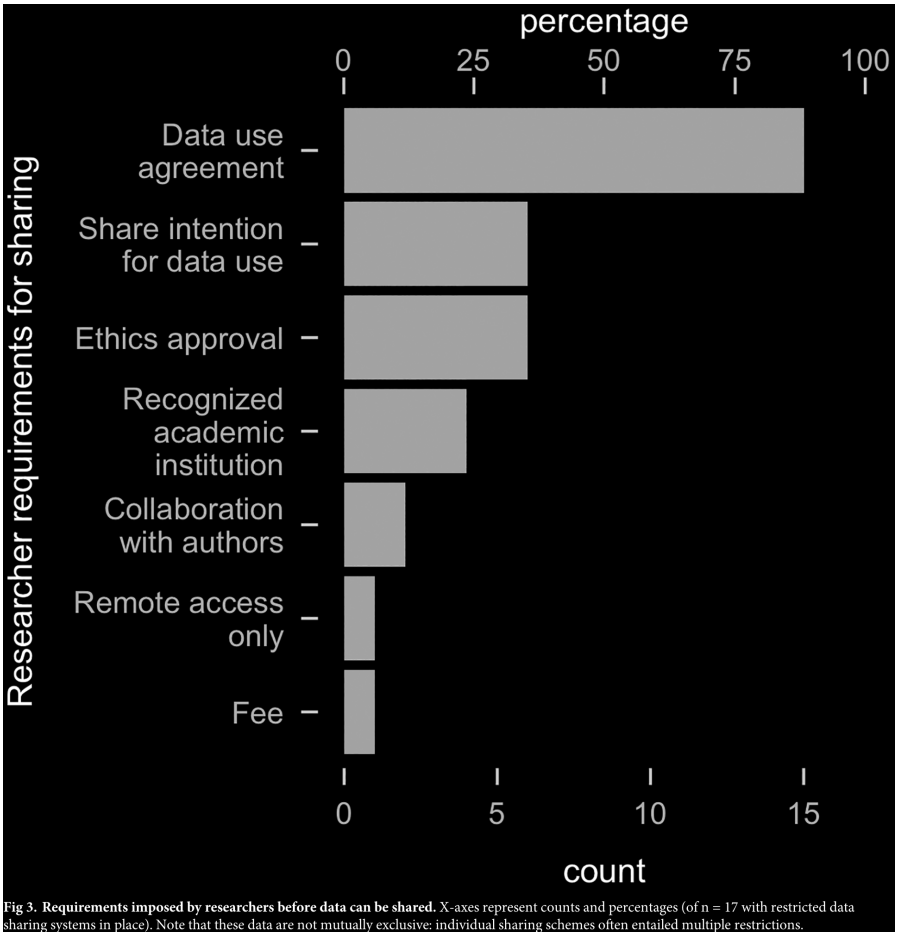

Fig 3. Requirements imposed by researchers before data can be shared. X-axes represent counts and percentages (of n = 17 with restricted data sharing systems in place). Note that these data are not mutually exclusive: individual sharing schemes often entailed multiple restrictions.

Figure 4: Source: Hardwicke & Ioannidis (2018), p. 6 (all rights reserved)

4.1 Data and code ≠ Reproducibility

- access to data and code do not mean analyses are reproducible
- what can go wrong?

1. Data problems
    - inaccessible data
    - incomplete data (e.g., 2/3 experiments)
2. Code problems
    - incomplete code
    - error messages
    - code rot: outdated syntax or environment
    - proprietary software
3. Documentation problems
    - data difficult to interpret
    - no README file/data dictionary
    - unclear folder/file/variable naming convention
    - manuscript contradicts code
4. Unclear terms of use
    - no licence specification

4.2 Share the code, not just the data

- suggestions for researchers from Laurinavichyute et al. (2022)

1. Share data in usable form
    - with pre-processing code
2. Use publicly accessible repositories
3. Use non-proprietary data formats
    - e.g., not `.mat` files (MATLAB)
4. Provide documentation
5. Share code *and* data
    - they estimate a 38% increase in reproducibility
    - different results can b
6. Teach data management and computing skills

5 Reproducible workflow

- project-oriented
- project-specific
- contained in a single project folder
- we will be using RProjects

5.1 Project management

- folder structure
- project-relative file paths
- appropriate documentation
  - e.g., README

5.2 Literate programming

- code is linear
- concise commenting
- one script per goal
- file paths are preferably interoperable
- facilitates maintainence

5.3 Documentation

- metadata
  - project README
  - codebook/data dictionary

5.4 Version control

- git: local tracking
- useful for the analysis and writing phases
  - but can be tricky for collaboration
- GitHub/GitLab: remote tracking
  - store your changes to your local git repository

  - then push them to your remote repository
- safe guards against local hardware/software issues
  - lost or damaged computer or local files
- and allows for collaboration or sharing

5.5 Persistant (pubic) storage

- GitHub/Lab are sub-optimal
  - developer-focused
  - typically lack thorough documentation/metadata
  - not very user-friendly for non-users
- OSF, Zenodo
  - Open Science-focused
  - can be linked to a GitHub/Lab repository
  - facilitate thorough documentation
  - user-friendly

5.6 Writing

- dynamic reports with Markdown syntax
  - e.g., Rmarkdown, Quarto
  - integration of data, code, and prose
    - facilitates cross-referencing within document
    - integration of citation management tools
    - supports LaTeX syntax for example sentences and tables
- `papaja` package for APA-formatted Rmarkdown documents
- challenge: collaboration
  - not all collaborators know these tools
  - track changes not currently possible
  - some tools to facilitate version control with dynamic reports
    - `trackdown` package: link an `.Rmd` file with a Google Doc
    - GitHub/Lab: push and pull changes to source code

6 Setting up a project

- next week: hands-on
- required installations/recent versions of:
  - R
    - version `4.4.0`, "Puppy Cup"
    - check current version with `R.version`
    - download/update: https://cran.r-project.org/bin/macosx/
  - RStudio
    - version `2023.12.1.402`, "Ocean Storm"
    - Help > Check for updates
    - new install: https://posit.co/download/rstudio-desktop/

Learning objectives

Today we learned...

-

Important terms

References

Bochynska, A., Keeble, L., Halfacre, C., Casillas, J. V., Champagne, I.-A., Chen, K., Röthlisberger, M., Buchanan, E. M., & Roettger, T. B. (2023). Reproducible research practices and transparency across linguistics. *Glossa Psycholinguistics*, *2*(1). https://doi.org/10.5070/G6011239

Hardwicke, T. E., & Ioannidis, J. P. A. (2018). Populating the Data Ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLOS ONE*, *13*(8), e0201856. https://doi.org/10.1371/journal.pone.0201856

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.