gesis Leibniz Institute for the Social Sciences

Workflows for Reproducible Research with R & Git

Introduction

Johannes Breuer, Bernd Weiss, & Arnim Bleier

2023-11-16

Leibniz Association

About us

Johannes Breuer

- Senior researcher in the team *Digital Society Observatory*, department *Computational Social Science* at *GESIS*
    - digital trace data for social science research
    - data linking (surveys + digital trace data)
- Leader of the team *Research Data & Methods* at the *Center for Advanced Internet Studies* (CAIS)
- Ph.D. in Psychology, University of Cologne
- Other research interests
    - Use and effects of digital media
    - Computational methods
    - Data management
    - Open science

johannes.breuer@gesis.org, personal website

About us

Bernd Weiß

- Head of team *GESIS Panel* and deputy head of the department *Survey Design and Methodology* at *GESIS*
- Obtained a doctorate (Sociology) from the University of Cologne in 2008
- Research interests: methods of empirical research in the social sciences, survey methodology, family sociology and juvenile delinquency
- Approach to this workshop (and a disclaimer): Open Science and reproducible research are tools, but not part of my research agenda and I do not claim to be an expert in any of the things I will be talking about...

bernd.weiss@gesis.org, ORCID: 0000-0002-1176-8408

About us

Arnim Bleier

- Senior researcher in the team *Designed Digital Data* in the department *Computational Social Science* at *GESIS*
  - natural language processing
  - probabilistic graphical models
- Ph.D. in statistics / machine learning, Leipzig University
- Other research interests
  - Structural equation modeling
  - Distributed databases
  - Right to replicate

arnim.bleier@gesis.org, Google Scholar

About you

- What's your name?
- Where do you work & what is your field?
- What are your experiences with reproducible research practices (and the tools we cover in this course)?
- What do you hope to get out of this course?

Please try to keep it brief.

Goals of this course

After this course you should be...

- familiar with key concepts of reproducible research workflows
- able to work with frameworks and tools that can be used for maximizing reproducibility, such as `Git`, `R` packages for dependency management, or *Binder*
- able to publish reproducible computational analysis pipelines with `R`

Prerequisites

For this course (esp. the exercises) you should have the following things installed on your computer:

- A version of `R` that is >= 4.0.0
  - the following `R` packages: `usethis`, `gitcreds`, `groundhog`

```r
# check if packages are installed and install missing ones
packages = c("usethis", "gitcreds", "groundhog")

install.packages(setdiff(packages, rownames(installed.packages())))
```

- A recent version of *RStudio*
- `Git`

In addition, you should also have/create a *GitHub* account.

Prerequisites

Did you have any trouble with the setup for this workshop?

Installing/setting up...

- `git`
- `R`
- *RStudio*
- the required `R` packages `usethis`, `gitcreds`, & `groundhog`
- a *GitHub* account

?

Workshop Structure & Materials

- The workshop consists of a combination of lectures and hands-on exercises
- Slides and other materials are available at

https://github.com/jobreu/reproducible-research-gesis-2023

- The workshop repository on the *GESIS ILIAS* contains some literature on tools and workflows for reproducibility as well as a timetable for this workshop

Online format

- If possible, we invite you to turn on your camera
- Feel free to ask questions anytime
  - If you have an immediate question during the lecture parts, please send it via text chat, publicly or privately (ideally to a person who is currently not presenting)
  - If you have a question that is not urgent and might be interesting for everybody, you can also use audio (& video) to ask it at the end of a lecture part or during the exercises (please use the use the "raise hand" function in *Zoom* for this)
- We would kindly ask you to mute your microphones when you are not asking (or answering) a question

Course schedule - Day 1

Time	Topic
09:30 - 10:45	Introduction
10:45 - 11:00	Coffee Break
11:00 - 12:00	Computer literacy
12:00 - 13:00	Lunch Break
13:00 - 15:00	Git & GitHub - Part 1
15:00 - 15:15	Coffee Break
15:15 - 16:30	Git & GitHub - Part 2
16:30 - 17:00	Q & A

Course schedule - Day 2

Time	Topic
09:00 - 09:30	Recap Day 1
09:30 - 11:00	Dependency management
11:00 - 11:15	Coffee Break
10:15 - 12:00	Build your own Binder
12:00 - 13:00	Lunch Break
13:00 - 14:30	Binder & Notebooks
14:30 - 14:45	Coffee Break
14:45 - 16:00	Saving computational environments
16:00 - 17:00	Recap & Outlook

Disclaimer

We will cover several different tools that can be used for reproducible research in the quantitative social sciences. We will only be able to cover the basics of those tools, so if you want to continue to use them and use them in more advanced ways, you will probably need to "dig deeper" eventually and consult further resources (documentation, further tutorials, blog posts, or other publications).

Our goal is that people with no or only very limited experience with the tools we will cover in this workshop are able to follow and keep up. For those who already have quite some experience with one or more of the tools we will cover, feel free to try out some additional things or options (you can, e.g., check some of the additional topics and tools from our Outlook slides), or have a closer look at documentation of the tools. This also applies if the exercises are too easy for you and you are done earlier with them).

Disclaimer

As you probably already know, there are a lot of different tools and workflows that can be employed for increasing the reproducibility of research. We will introduce you to some of those, but there is more, and, in the end, it depends on your personal preferences and needs what tools and workflows you employ.[1]

In this course, we will focus on free and open-source software (FOSS).[2] We will also focus on `R`, but there are solutions for reproducible research with other programming languages, such as `Python` or `Julia`, as well as statistical software packages, such as *SPSS* and *Stata*.

[1] As you will see, we instructors also all have different preferences in our workflows and tool use.

[2] The only exceptions are *Docker* and *GitHub* which belongs to *Microsoft*. We chose *GitHub* over *GitLab*, however, as it is generally more widely used and available to everybody, regardless of whether your institution maintains its own *GitLab* server or not.

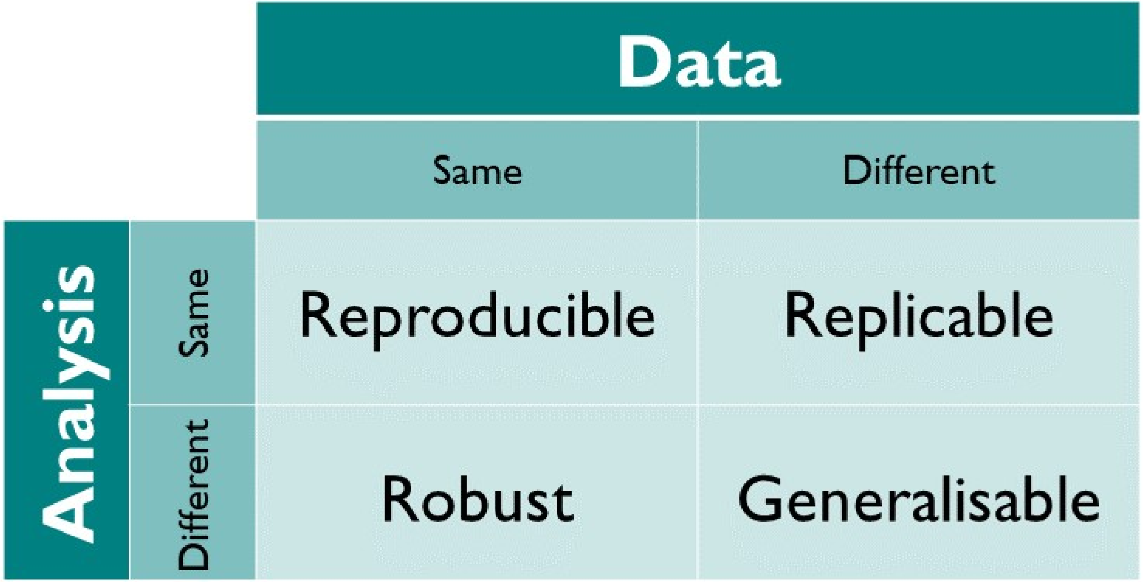
Any questions so far?

Next up: What is reproducibility and why does it matter?

Defining reproducibility

> A minimum standard on a spectrum of activities (“reproducibility spectrum”) for assessing the value or accuracy of scientific claims based on the original methods, data, and code […] In some fields, this meaning is, instead, associated with the term “replicability” or ‘repeatability’ (FORRT Glossary - Reproducibility)

The Turing Way definition

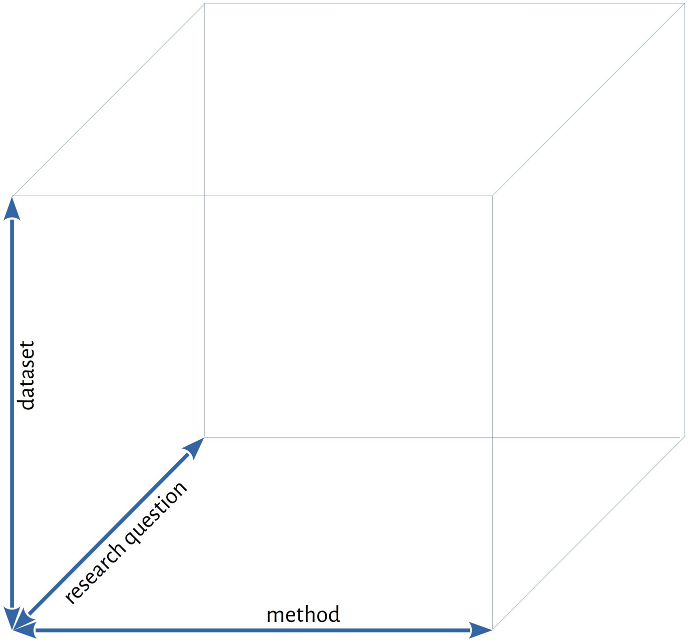


		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Source: https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html

Defining dimensions

3-dimensional concept space



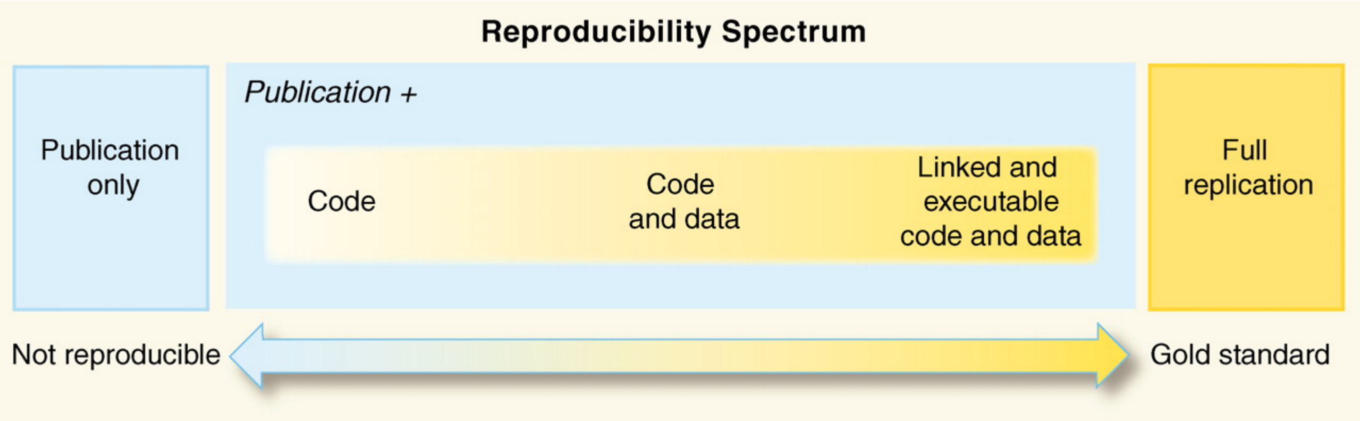
By Christof Schöch. Source: https://dh-trier.github.io/trr/#/2/1

Computational reproducibility

There also are distinctions between different types (or components) of reproducibility. One type/component that is especially relevant for this workshop is "Computational reproducibility":

> Ability to recreate the same results as the original study (including tables, figures, and quantitative findings), using the same input data, computational methods, and conditions of analysis. The availability of code and data facilitates computational reproducibility, as does preparation of these materials (annotating data, delineating software versions used, sharing computational environments, etc). Ideally, computational reproducibility should be achievable by another second researcher (or the original researcher, at a future time), using only a set of files and written instructions (FORRT Glossary - Computational reproducibility)

Reproducibility as a continuum



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

Why reproducibility matters?

Reproducibility ensures that research is transparent, verifiable, and trustworthy.

Studies have repeatedly shown suboptimal reproducibility of research in the social sciences (see, e.g., Artner et al., 2021, Hardwicke et al., 2020, Krähmer et al., 2023, Trisovic et al., 2022).

Motivations for reproducibility

- Increasing the robustness and trustworthiness of your own research
- Facilitating collaboration (through the use of common tools and standards)
- Being kind to future you:
    - Editing and reusing your own code (e.g., for a paper revision or a follow-up study)
    - Easily being able to pick up a project after a break or finding and understanding things again after a longer time

Tools & workflows 🛠 📋

In our case, **tools** are programming languages, programs, and other pieces of software that we can use to make our research (more easily) reproducible.

Workflows are the ways in which we combine these tools to achieve our goal.

Tools & workflows 🛠️ 📋

Choosing tools and establishing workflows are somewhat idiosyncratic processes that depend on...

- the requirements of your project (methods, data types...)
- the availability of tools
- your skills and knowledge
- the preferences of collaborators

...

Reproducible research workflows

> being an open scientist means adopting a few straightforward research management practices, which lead to less error-prone, reproducible research workflows (Klein et al., 2018, p. 11)

Research management practices

There are quite a few practices that researchers can adopt to increase the reproducibility of their work.

- Using free and open source software (FOSS)
- Project-oriented workflow
- Clear folder structures
- Naming things
- ...

Notably, all of those practices as well as the use of the tools we cover in this workshop require a certain degree of **Computer literacy**.