gesis Leibniz Institute for the Social Sciences

Tools and Workflows for Reproducible Research in the Quantitative Social Sciences

Introduction

Johannes Breuer, Bernd Weiss, & Arnim Bleier

2022-11-17

Leibniz Association

About us

Johannes Breuer

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

johannes.breuer@gesis.org, @MattEagle09, 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, @berndweiss

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, @arnimb, 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)?

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 (start) work(ing) with tools for reproducible research, such as `Git`, *GitHub*, `R Markdown`, `Jupyter Notebooks`, and `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: `rmarkdown`, `usethis`, `gitcreds`

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

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 `rmarkdown`, `usethis`, & `gitcreds`
- 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-2022

- 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
10:00 - 11:00	Introduction
11:00 - 12:00	Computer literacy for reproducible research
12:00 - 13:00	Lunch Break
13:00 - 15:00	Introduction to R Markdown
15:00 - 15:15	Break
15:15 - 17:00	Git & GitHub - Part 1

Course schedule - Day 2

Time	Topic
09:30 - 10:30	Git & RStudio
10:30 - 10:45	Break
10:45 - 11:45	Jupyter Notebooks & Binder
11:45 - 12:30	Build your own Binder
12:30 - 13:30	Lunch Break
13:30 - 14:00	Sharing & Publishing
14:00 - 14:15	Break
14:15 - 15:45	Git & GitHub - Part 2
15:45 - 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).

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]

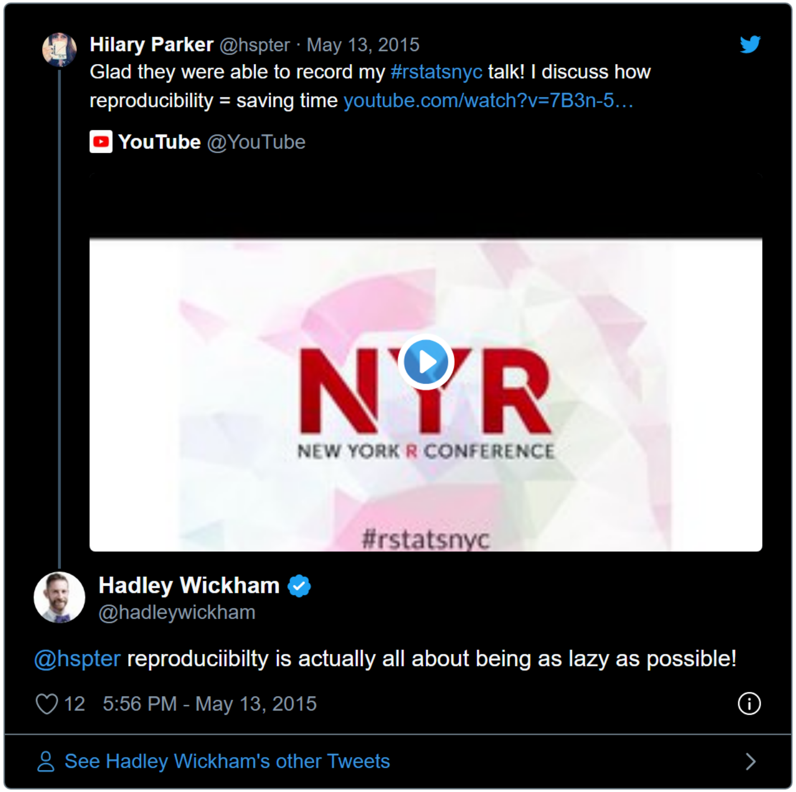
In this course, we will focus on reproducible research using `R`. However, there also are solutions for reproducible research with other programming languages, such as `Python` or `Julia`, as well as statistical software packages, such as *SPSS* and *Stata*.

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

Any questions so far?

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

Why should we aim for reproducibility?



Hilary Parker @hspter · May 13, 2015

Glad they were able to record my #rstatsnyc talk! I discuss how reproducibility = saving time youtube.com/watch?v=7B3n-5…

YouTube @YouTube

Hadley Wickham
@hadleywickham

@hspter reproduciibilty is actually all about being as lazy as possible!

♡ 12 5:56 PM - May 13, 2015

See Hadley Wickham's other Tweets

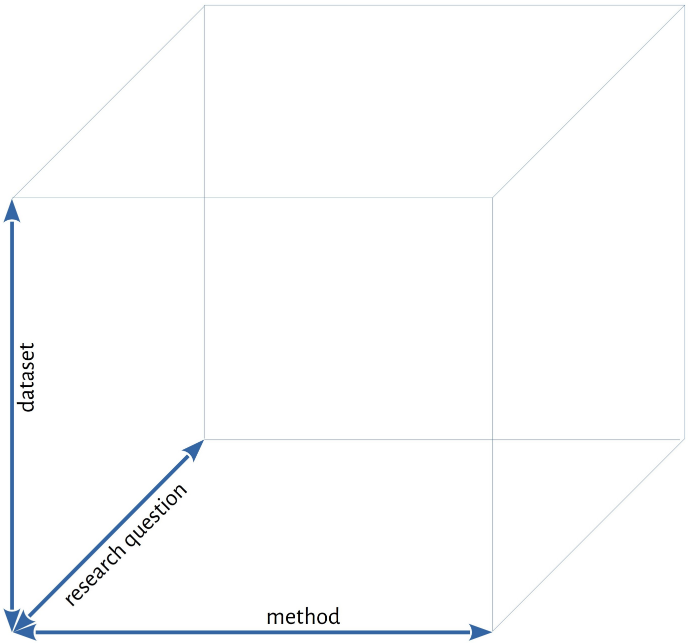
https://twitter.com/hadleywickham/status/598532170160873472?ref_src=twsrc%5Etfw

What is reproducibility?

As with (almost) everything in science, there are different definitions of reproducibility. We will discuss some of them in the following.

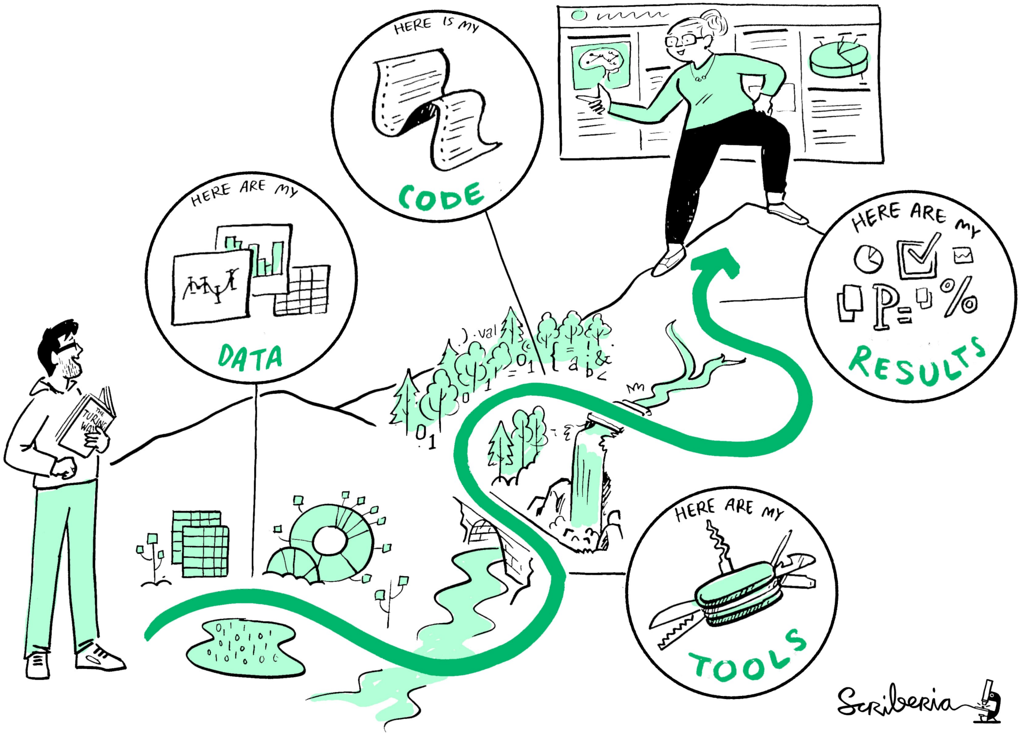
Defining dimensions

3-dimensional concept space



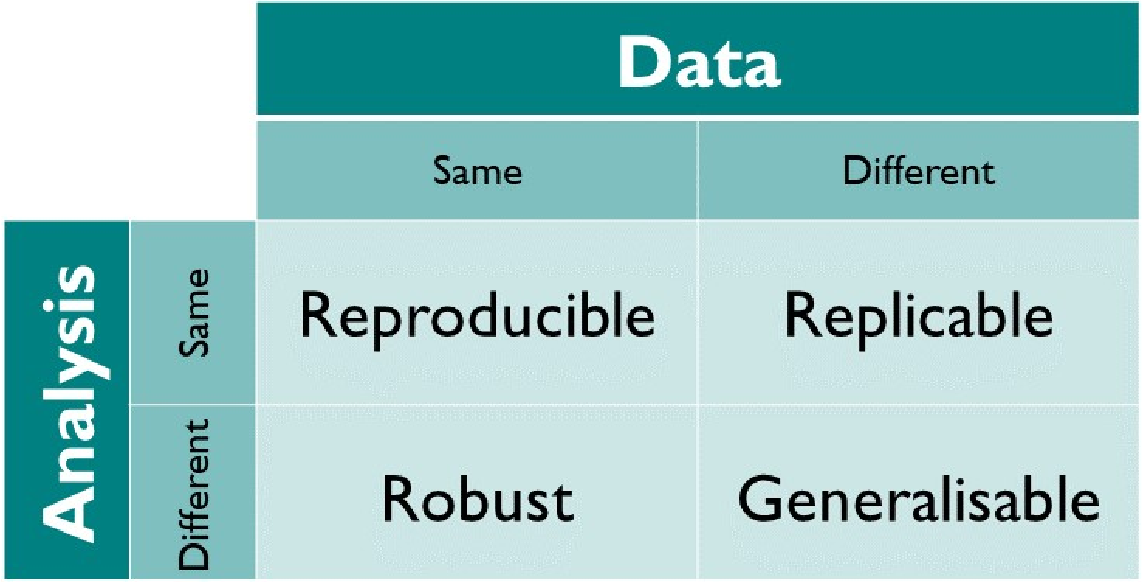
By Christof Schöch. Source: https://dh-trier.github.io/trr/#/2/1

Defining dimensions



The Turing Way Project illustration by Scriberia. DOI: 10.5281/zenodo.3332807

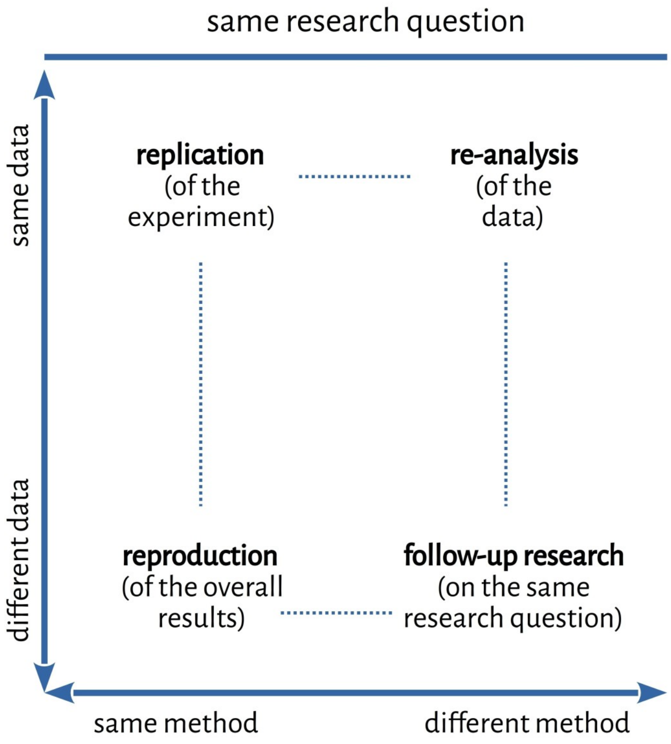
The Turing Way definition



		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

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

Replication or reproduction?



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

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)

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)

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 worklflows 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.

- Project-oriented workflow
- Clear folder structures
- Naming things
- ...

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