Reproducible Analyses

What is it and why should I care?

Daniela Palleschi

2023-04-12

Table of contents

Replication **2**
An example from language research 2

Reproducibility **3**
Replication vs. Reproducibility 3

Open Science: Why should I care? **3**
What can I do? 4
How to do better science 4
What will we learn here? 4

R is for Reproducibility **5**

Exercises **6**
RStudio 6
. 6
Quarto 8
YAML 8
Headings and text 8
Code snippets 9
In-line code 9
Altogether 9
Quarto Exercises 10
Quarto cont'd 10

```
knitr::opts_chunk$set(eval = T, # change this to 'eval = T' to reproduce the analyses; mak
                      echo = T, # 'print code chunk?'
```

```
                    message = F, # 'print messages (e.g., warnings)?'
                    error = F,
                    warning = F)
```

Replication

> "There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims"
>
> – Ioannidis (2005)

- replication refers to re-running a previous experiment with as few differences as possible
  - aim: determine whether the original results were *robust* and are *replicable*
  - if yes, great! the original findings are reliable
  - if no, hmm, maybe the original findings were false positives? or due to some other factor?
- in recent years, researchers have tried to *replicate* classic studies in their field
  - but in many cases, they did not get the same effects the original study reported (and were famous for)
- this began the ***replication crisis***

An example from language research

- Nieuwland et al. (2018): a *direct* EEG[1] replication (versus *conceptual* replication)
- a multi-lab replication of DeLong et al. (2005)'s impactful paper
  - DeLong et al. (2005): reported N400 effects elicted at unexpected nouns, but also on preceding determiners (English *a/an*) when it signalled an unexpected word,
    - e.g., *The day was breezy so the boy went outside to fly…a kite/*an airplane*
    - taken as evidence of pre-activation of phonological form, graded by cloze probability
  - Nieuwland et al. (2018): replicated N400 at noun, but not at adjective
    - i.e., *failure to replicate* a famous finding

[1] electroencephalography

Reproducibility

- reproducibility refers to the ability to *reproduce* somebody's analyses with their
  - data
  - *and* code
- it is not something we do once, nor is it something that will get us published
  - but it's important for open science and encourages transparency

Replication vs. Reproducibility

- **replication** of a study
  - repeating an **experiment**
  - getting *similar* results
- **reproducibility** of analyses
  - repeating **analyses** of the *same data*
  - getting the *same* results
- e.g., when you submit a paper to a journal, they make ask for your data and code so reviewers can *reproduce* your analyses
  - requires data and code
- if you have interesting findings, other researchers (or future you) may want to *replicate* your study to see if they can *replicate* your findings
  - (may require) stimuli, set-up and presentation information, participant demographics

Open Science: Why should I care?

1. Science is cumulative
   - We should ensure we're building on reliable, robust findings
   - i.e., it's *good* scientific practice
2. Because the field cares
   - replication/reproducibility are beginning to be foregrounded by e.g., journals/job advertisements
3. Helps future you

- pre-registration, reproducible analyses, clean and shareable data: all help *future you*

What can I do?

- there's a variety of open science practices that we can choose to implement
- some suggestions from Kathawalla et al. (2021):

Level: Easy

1. Journal Club
2. Project workflow
3. Pre-prints

Level: Medium

4. Reproducible code
5. Sharing data
6. Transparent manuscripts
7. Pre-registration

Level: Difficult

8. Registered reports

How to do better science

- don't be afraid of making mistakes
  - (most) researchers aren't statisticians or programmers
  - do the best you can, and ***be transparent***
- doing *some* of the steps is better than doing *none*

What will we learn here?

Design and Reporting

- Preregistration/Registered Reports
- Transparent writing

Analysis

- Reproducible code

  - with open source software (R, RStudio, packages)
  - dynamic reports with Quarto/Rmarkdown

- Project workflow
  - folder structure
    - how to sensibly set up your folders
  - contained environments
    - using RProjects and the `here` package

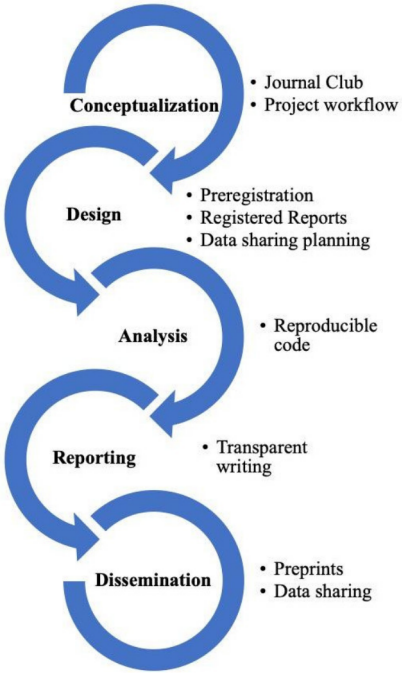

Figure 1. Open Science research practices across the research cycle

Figure 1: Image source: Kathawalla et al. (2021) (all rights reserved)

R is for Reproducibility

- we will be working with R, RStudio, Quarto, and RProjects
  - R: a programming language for statistical computing and graphics
  - RStudio: an integrated development environment (IDE)
    - RStudio Desktop

    - RStudio Server
  - Quarto (similar to Rmarkdown): dynamic reports
    - combining text, code, and printed tables and figures
  - RProjects: a workflow tool
    - contains all files necessary for a project
    - works with *relative* file paths

Exercises

RStudio

1. Open RStudio

   - locate the Environment, Files, and Console panes
   - File > New File > R script
   - write `[your birth-month number]*[the your birth day]` and hit Enter
   - write `print("Hello World!")`
   - write `number <- 3*32`; this will create an object/variable 'number'
   - write `string <- "Hello World!"`; this will create an object/variable 'string'
   - write `number`
   - write `string`
   - add comments describing each step using `#`
   - File > Save As

```
# multiply 5 by 7
5*7
```

```
[1] 35
```

```
# print some text
print("Hello World!")
```

```
[1] "Hello World!"
```

```r
# save an object 'number' with 5*7
number <- 5*7
```

```r
# save an object 'string' with text
string <- "Hello World!"
```

```r
# print number
number
```

```
[1] 35
```

```r
# print string
string
```

```
[1] "Hello World!"
```

```r
# do math with objects
number+number
```

```
[1] 70
```

```r
number*number
```

```
[1] 1225
```

```r
number*2
```

```
[1] 70
```

```r
month <- 5
```

```r
day <- 7
```

```r
month*day
```

```
[1] 35
```

Quarto[2]

- R scripts are a great way to keep track of what you did
  - however, the output is not saved, and adding comments with `#` gets kind of chunky
  - enter: dynamic reports!
- dynamic reports are those that combine text, code, and output
  - they are a great tool for communicating, collaborating, and documenting
  - they are also fantastic for note-taking
- Rmarkdown vs. Quarto
  - both can combine text with code, outputting PDFs, Word Documents, html, or slides
  - main difference: Quarto has native support of a wider range of programming languages (e.g., Python and Julia)
- Want to know more? Check out Hadley Wickham's intro (Wickham et al., n.d.)

YAML

```
---
title: "My title"
author: "My name"
format: html
---
```

- YAML is a human-readable programming language used to configure documents
- formatting is important: but be sandwiched between `---` and `---`
- in Quarto the output type must at least be given (with R: pdf, html, revealjs)

Headings and text

```
# This is a heading

This is text.

## This is a sub-heading

This is more text.
```

[2]https://r4ds.hadley.nz/quarto.html#workflow

- headings are indicated by `#`
  - the number of `#`'s indicates the heading level

Code snippets

```
# do some math
year <- 1989
dog <- "Lola"
```

- sandwiched between `markdown```{r}` and `markdown`
  - shortcut: Ctrl/Cmd+Alt+I

In-line code

```
I was born on `r month`/`r day`/`r year`. My dog's name is `r dog`.
```

I was born on 5/7/1989. My dog's name is Lola.

- code output that was run *above* text can be called in-line using `` `r ` ``

Altogether

````
---
title: "My title"
author: "My name"
format: html
---

# This is a heading

This is text.

## This is a sub-heading

This is more text.

Add some code chunks.

```{r}
````
````

```
# do some math
year <- 1989
dog <- "Lola"
```

And use call objects for in-line code: I was born on `r month`/`r day`/`r year`. My dog's

Quarto Exercises

3. Create a new Quarto document
    - File > New File > Quarto Document
    - Read the instructions
    - Practice running the chunks individually
    - render the document
    - verify that you can modify the code, re-run it, and see modified output
4. Create one new Quarto document for each of the three built-in formats: HTML, PDF and Word.
    - Render each of the three documents
    - How do the outputs differ?
    - How do the inputs differ?[3]

Quarto cont'd

- Choose a Quarto document:
    - give it a title, your name (author), and unclick 'Use visual markdown editor'
- Render
- YAML:

```
title: "Eye-tracking during reading"
subtitle: "Lecture 2 notes"
author: "[YOUR NAME HERE]"
lang: en
date: `r Sys.Date()`
```

- Render

[3]You may need to install LaTeX in order to build the PDF output — RStudio will prompt you if this is necessary.

- you can now try writing your class notes in this document (if you're brave)

References

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. https://doi.org/10.1038/nn1504

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), 2–8. https://doi.org/10.1371/journal.pmed.0020124

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

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (n.d.). *R for Data Science* (2nd ed.). https://r4ds.hadley.nz/