

Reproducible Manuscripts in R Markdown

Juli Tkotz

27-05-2020

ZI Mannheim, RG Psychology and Neurobiology of Sleep and Memory

Nice to meet(up) you!



- Juli Tkotz
- PhD student ZI Mannheim
- RG Psychology and Neurobiology of Sleep and Memory

What is a reproducible manuscript?

- A manuscript that directly embeds your **research data** and **analysis code**.
- Any person with the raw data can run the code and **reproduce** your manuscript.
- Interactive stand-alone versions are possible.

Why do we need it?

Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology

Pepijn Obels¹, Daniel Lakens¹, Nicholas A. Coles², Jaroslav Gottfried³, & Seth Ariel Green⁴

¹ Eindhoven University of Technology, The Netherlands

² University of Tennessee, Knoxville, USA

³ Masaryk University, Brno, Czech Republic

⁴ Code Ocean, New York, USA

Ongoing technological developments have made it easier than ever before for scientists to share their data, materials, and analysis code. Sharing data and analysis code makes it easier for other researchers to re-use or check published research. These benefits will only emerge if researchers can reproduce the analysis reported in published articles, and if data is annotated well enough so that it is clear what all variables mean. Because most researchers have not been trained in computational reproducibility, it is important to evaluate current practices to identify practices that can be improved. We examined data and code sharing, as well as computational reproducibility of the main results, without contacting the original authors, for Registered Reports published in the psychological literature between 2014 and 2018. Of the 62 articles that met our inclusion criteria, data was available for 40 articles, and analysis scripts for 37 articles. For the 35 articles that shared both data and code and performed analyses in SPSS, R, Python, MATLAB, or JASP, we could run the scripts for 31 articles, and reproduce the main results for 20 articles. Although the articles that shared both data and code (35 out of 62, or 56%) and articles that could be computationally reproduced (20 out of 35, or 57%) was relatively high compared to other studies, there is clear room for improvement. We provide practical recommendations based on our observations, and link to examples of good research practices in the papers we reproduced.

263 3.1 Epistemic Trustworthiness

264 Participants placed more epistemic trust in the debaters when reading a neutral debate: Student
265 teachers in the neutral condition ($M = 5.06, SD = 1.00$) perceived the debaters to have more expertise
266 than those in the uncivil condition ($M = 5.06, SD = 1.00$), $t(218.49) = 1.99, p = .047, d = 0.27$.
267 Furthermore, participants reading a neutral debate ($M = 4.76, SD = 1.02$) reported higher ratings of
268 debaters' integrity than those reading an uncivil debate ($M = 4.05, SD = 1.15$), $t(219.41) = 4.87, p <$
269 $.001, d = 0.65$. Additionally, ratings of benevolence were higher in the neutral condition ($M = 4.77,$
270 $SD = 0.98$) than in the uncivil condition ($M = 4.05, SD = 0.89$), $t(214.11) = 5.67, p < .001, d = 0.76$
271 (see Figure 2).

272 We further explored the correlation between the conflict explanation items and the METI subscales,
273 that is, if the perception of various aspects of a conflict was associated with different degrees of
274 epistemic trust. Those who agreed that the debaters in the scenario referred to different research
275 results also thought them to have more expertise, $r(220) = .14, p = .039$. There was no relation with
276 integrity, $r(220) = .07, p = .321$, or benevolence, $r(220) = .03, p = .679$. Assuming personal reasons
277 for the conflict had the strongest relationship with epistemic trust. The more participants perceived
278 the conflict to be personal, the less expertise they assigned to the debaters $r(220) = -.25, p < .001$. In
279 a similar manner, perception of a personal conflict lead to decreased ratings of integrity, $r(220) =$
280 $-.36, p < .001$, and benevolence, $r(220) = -.41, p < .001$. How much participants agreed that the
281 debaters referred to different goals of PAVLOV did not correlate with any of the METI subscales,
282 neither with expertise, $r(220) = .10, p = .122$, nor with integrity, $r(220) = -.00, p = .946$, nor with
283 benevolence $r(220) = -.00, p = .994$. Embracement of the statement that debaters referred to different
284 effects of PAVLOV was not associated with epistemic trust either, neither with expertise, $r(220) =$
285 $.01, p = .863$, nor with integrity, $r(220) = -.06, p = .348$, nor with benevolence $r(220) = -.05, p =$
286 $.475$. Internal consistency of the METI subscales was somewhat lower than initially found by
287 Hendriks et al. (2015), with a Cronbach's α of .87 for expertise, .83 for integrity and .76 for
288 benevolence.

R Markdown to the rescue

```
## R Markdown to the rescue
```

```
```{r intext_stats, echo = TRUE}  
nerd <- read.csv("./data/nerd.csv", sep = "\t")
```
```

```
```{r copy_paste_hell}  
include_graphics("./pics/slide_inception.png")
```
```

This example dataset consists of $N = \text{r nrow(nerd)}$ participants with an age range between $\text{r min(nerd[["age"]])}$ and $\text{r max(nerd[["age"]])}$ years. Overall, $\text{r sum(nerd$age > 100)}$ participants reported to be older than 100, so we probably can't trust this data set a lot.

This example dataset consists of $N = 14955$ participants with an age range between 13 and 38822 years. Overall, 8 participants reported to be older than 100, so we probably can't trust this data set a lot.

Data retrieved from <https://openpsychometrics.org/>

R Markdown to the rescue

```
## R Markdown to the rescue
```

```
```{r intext_stats, echo = TRUE}  
nerd <- read.csv("./data/nerd.csv", sep = "\t")
```
```

```
```{r copy_paste_hell}  
include_graphics("./pics/slide_inception.png")
```
```

This example dataset consists of $N =$ ``r nrow(nerd)`` participants with an age range between ``r min(nerd[["age"]])`` and ``r max(nerd[["age"]])`` years. Overall, ``r sum(nerd$age > 100)`` participants reported to be older than 100, so we probably can't trust this data set a lot.

This example dataset consists of $N =$ **14955** participants with an age range between **13** and **38822** years. Overall, **8** participants reported to be older than 100, so we probably can't trust this data set a lot.

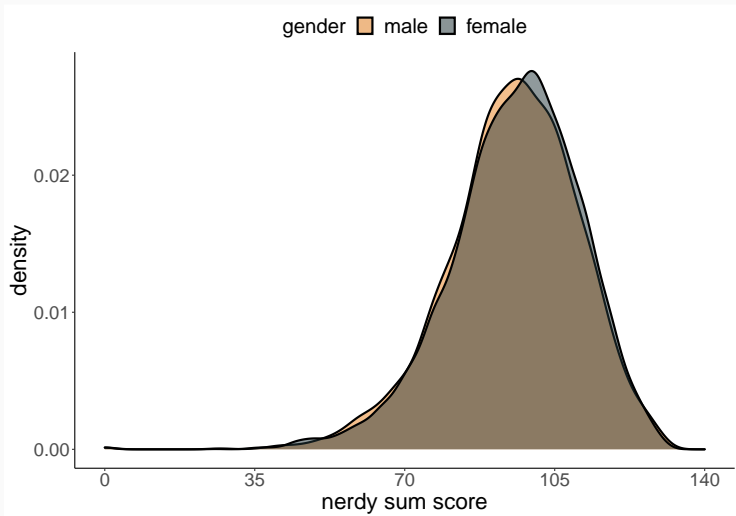
How about some stats?

```
nerd_ttest <- t.test(sum_score ~ gender, data = nerd)
nerd_effsize <- cohen.d(sum_score ~ gender, data = nerd)
```

```
In this dataset, men `r print_mean_sd(nerd[["sum_score"]][nerd[["gender"]] == "male")` have a
significantly lower nerd score than women `r print_mean_sd(nerd[["sum_score"]][nerd[["gender"]]
== "female")`, `r print_ttest(nerd_ttest, nerd_effsize)`.
```

In this dataset, men ($M = 95.18$, $SD = 15.27$) have a significantly lower nerd score than women ($M = 95.82$, $SD = 15.16$), $t(9800.06) = -2.39$, $p = .017$, $d = -0.04$.

Yeah, plots!



Yeah, references!

```
## Yeah, references!
```

```
``{r citation}  
include_graphics("./pics/citation.png")  
``
```

```
If I want to cite a paper, I can do this [san_martin_1968].  
This also works if I cite @san_martin_1968 as an in-text citation.
```

If I want to cite a paper, I can do this (San-Martin et al. 1968).
This also works if I cite San-Martin et al. (1968) as an in-text
citation.

How to get bibtex references

The screenshot shows a journal article page for "ASPECTS OF REPRODUCTION IN THE ALPACA". A modal window titled "Preview Citation" is open, displaying the citation in APA format and options to download or export the citation. The export options are .ris, .bib, and .enw. A red arrow points to the .bib option, which is associated with BibTeX and Zotero. The background page shows the article title, authors (M. SAN-MARTIN, M. COPAIRA, J. ZUNIGA, R. RODRIGUEZ), DOI, page numbers (395-399), and volume/issue information (Volume 16, Issue 3). The journal title "J. Reprod. Fert." is also visible.

Reproduction

Home Browse

ASPECTS OF REPRODUCTION IN THE ALPACA

Free access

in Reproduction

Authors: M. SAN-MARTIN, M. COPAIRA, J. ZUNIGA, R. RODRIGUEZ

DOI: <https://doi.org/10.1530/jrf.0.0160395> Article Type: Research Article

Page(s): 395-399 Online Publication Date: 2020

Volume/Issue: Volume 16, Issue 3

Abstract/Excerpt PDF

J. Reprod. Fert. (1968) 16, 395-399

ASPECTS OF REPR

Preview Citation

Format:

APA Download

SAN-MARTIN, M., COPAIRA, M., ZUNIGA, J., RODRIGUEZ, R., BUSTINZA, G., & ACOSTA, L. (1968). ASPECTS OF REPRODUCTION IN THE ALPACA, *Reproduction*, 16(3), 395-399. Retrieved May 20, 2020, from [https://rep.bioscientifica.com/view/journals/rep/16\(3\)/jrf_16_3_009.xml](https://rep.bioscientifica.com/view/journals/rep/16(3)/jrf_16_3_009.xml)

Export Citation

.ris .bib .enw

ProCite RefWorks Reference Manager BibTeX Zotero EndNote

Fully formatted articles

The R-packaga papaja offers you documents that are formatted according to APA (6) style.

<https://github.com/crsh/papaja>

1How to use papaja: An Example Manuscript Including Basic Instructions

2Frederik Aust¹

3¹ University of Cologne

HOW TO USE PAPAJA

6

Table 1

Descriptive statistics of correct recall by dosage.

| | Mean | Median | SD | Min | Max |
|--|------|--------|------|-------|-------|
| | 4.19 | 14.00 | 4.45 | 5.00 | 25.00 |
| | 3.50 | 14.00 | 5.15 | 4.00 | 22.00 |
| | 9.19 | 19.00 | 3.52 | 13.00 | 25.00 |

table was created with apa_table()

References

145

146Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., ... Hyndman, R. (2016).

147*Rmarkdown: Dynamic documents for r*. Retrieved from

148<https://CRAN.R-project.org/package=rmarkdown>

149

150Aust, F., & Barth, M. (2015). *Papaja: Create apa manuscripts with rmarkdown*.

151

152Bates, D., & Maechler, M. (2016). *Matrix: Sparse and dense matrix classes and methods*.

153Retrieved from <https://CRAN.R-project.org/package=Matrix>

Or a whole book?

R for Data Science

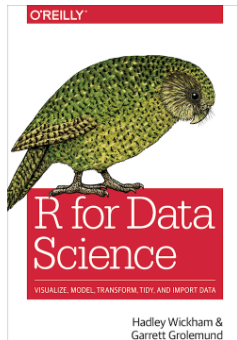
build passing

This repository contains the source of [R for Data Science](#) book. The book is built using [bookdown](#).

The R packages used in this book can be installed via

```
devtools::install_github("hadley/r4ds")
```

This is the website for “**R for Data Science**”. This book will teach you how to do data science with R: You’ll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you’ll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You’ll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You’ll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualising, and exploring data.



Talking about reproducibility ...

```
> anticlusters <- anticlust::anticlustering(  
+   iris[, -5],  
+   K = 3,  
+   objective = "variance",  
+   method = "exchange"  
+ )  
Error in loadNamespace(name) : there is no package called 'anticlust'
```

Code capsules

The screenshot displays the Code Ocean web interface for a capsule named 'MOSM_talk'. The interface is divided into several sections:

- Top Bar:** Includes the Code Ocean logo, a 'Private' status indicator, the capsule name 'MOSM_talk', and navigation links for 'Capsule', 'File', 'Edit', 'View', 'Tabs', 'Settings', and 'Help'. On the right, there are buttons for 'Launch Cloud Workstation' and 'Collaborate', along with a user profile icon.
- Left Panel:** Contains a 'Files' sidebar with a tree view of the capsule's structure. The tree shows folders like 'metadata', 'environment', 'code', 'data', 'pics', and 'results', each with its size and a green status icon. Below the tree is a 'Results' section with a link to 'View latest results'.
- Center Panel:** Displays the main content of the capsule, which is a document titled 'Reproducible Manuscripts in R Markdown'. The document content includes the author 'Juli Tkotz', affiliation 'Z1 Mannheim, RG Psychology and Neurobiology of Sleep and Memory', and the date '27-05-2020'.
- Right Panel:** Features a 'Reproducible Run' section with buttons for 'lab', 'Jupyter', 'RStudio', 'Python', and 'Shell'. Below this is a 'Timeline' section showing a list of runs. The most recent run is by 'Juli Tkotz' on 'May 25, 2020' at '00:08:50', with a status of 'Run 406321'. The run details show files like '00reproducible_man...', 'buildLog', and 'output' with their respective sizes. A 'Submit for publication...' button is also visible.

<https://codeocean.com/capsule/3923848/tree/v1>

There will be pain



But it's worth it

- Saves time for others and for future you.
- Mistakes are easier to spot and easier to correct.
- Your data and your manuscript will survive longer.
- Others can learn from your analyses.

Thank you!

Find this presentation on GitHub.

Or on CodeOcean.

Or on the OSF.

References

Obels, Pepijn, Daniel Lakens, Nicholas A Coles, Jaroslav Gottfried, and Seth A Green. 2019. "Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology." PsyArXiv.

<https://doi.org/10.31234/osf.io/fk8vh>.

San-Martin, M., M. Copaira, J. Zuniga, R. Rodreguez, G. Bustinza, and L. Acosta. 1968. "Aspects of Reproduction in the Alpaca." *Reproduction* 16 (3): 395–99. https://rep.bioscientifica.com/view/journals/rep/16/3/jrf_16_3_009.xml.