The 3 R’s of open science

Rick O. Gilmore

2018-10-14 18:56:03

# Preliminaries

## Themes

1. **R**eproducibility
2. **R**eproducible research with R and R Markdown
3. **R**egistered reports and pRe-registration

# Reproducibility

<https://www.youtube.com/embed/66oNv_DJuPc>

## Is there a reproducibility crisis?

* Yes, a significant crisis
* Yes, a slight crisis
* No crisis
* Don’t know

[Baker 2016](http://doi.org/10.1038/533452a)

## Not just psychology

## Psychological science is harder than physics

## Much harder

“*…psychologists tend to treat other peoples’ theories like toothbrushes; no self-respecting individual wants to use anyone else’s.*”

[(Mischel, 2009)](https://www.psychologicalscience.org/observer/becoming-a-cumulative-science)

“*The toothbrush culture undermines the building of a genuinely cumulative science, encouraging more parallel play and solo game playing, rather than building on each other’s directly relevant best work.*”

[(Mischel, 2009)](https://www.psychologicalscience.org/observer/becoming-a-cumulative-science)

## Why is reproducibility hard?

[Baker 2016](http://doi.org/10.1038/533452a)

## Risks from idea to product

[(Munafo et al. 2017)](http://doi.org/10.1038/s41562-016-0021)

## What do we mean by ‘reproducibility’?

* *Methods* reproducibility
  + Enough details about materials & methods recorded (& reported)
  + Same results with same materials & methods
* *Results* reproducibility
  + Same results from independent study
* *Inferential* reproducibility
  + Same inferences from one or more studies or reanalyses

[Goodman et al., 2016](http://doi.org/10.1126/scitranslmed.aaf5027)

## Achieving *methods* reproducibility

* Data collection
* Cleaning
* Visualization
* Analysis
* Reporting
* Manuscript, talk, poster preparation

No one is irreplaceable, but we can all strive to be indispensible.

What’s your project’s ‘bus number’?

# R-eproducible research with R, RStudio, and R Markdown

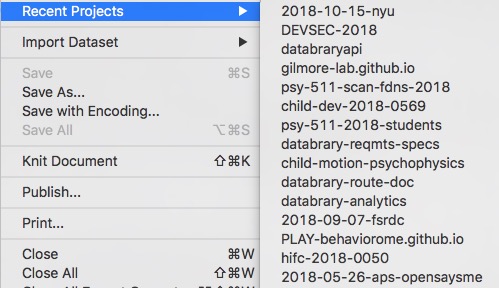
## Why R?

* It’s fun
* It’s free
* Amaze your friends; dazzle your rivals
* Powerful data manipulation, plotting, analysis

[RStudio](http://rstudio.com)

## Why [RStudio](http://rstudio.com)?

* Integrated development environment (IDE) for R
* It’s free, open source
* Suitable for beginners and experts
* Integration with git version control and [GitHub](http://github.com) web repository
* Projects management: Keep projects separate



## Reproducible workflows

* Scripted, automated = minimize human-dependent steps.
* Well-documented
* Kind to your future (forgetful) self
* Transparent to me & colleagues == transparent to others

# Import/gather data  
  
# Clean data  
  
# Visualize data  
  
# Analyze data  
  
# Report findings

# Import data  
my\_data <- read.csv("path/2/data\_file.csv")  
  
# Clean data  
my\_data$gender <- tolower(my\_data$gender) # make lower case  
...

# Import data  
source("R/Import\_data.R") # source() runs scripts, loads functions  
  
# Clean data  
source("R/Clean\_data.R")  
  
# Visualize data  
source("R/Visualize\_data.R")  
...

## Objections

* But my SPSS syntax file already does this
  + Great! How are you sharing them?
* But I prefer {Python, Julia, Ruby, Matlab, …}
  + Great! Let’s talk about [R Markdown](https://rmarkdown.rstudio.com/)

## Reproducible research with [R Markdown](https://rmarkdown.rstudio.com/)

* Add-on package to R, developed by the RStudio team
* Combine text, code, images, video, equations into one document
* Render into PDF, MS Word, HTML (web page or site, slides, a blog, or even a book)
* [R Markdown documentation](https://rmarkdown.rstudio.com/); online [tutorial](http://r4ds.had.co.nz/r-markdown.html); Mike Frank and Chris Hartgerink’s [tutorial](https://libscie.github.io/rmarkdown-workshop/handout.html)
* Similar to Mathematica notebooks, [Jupyter notebooks](http://jupyter.org)
* R Markdown extends Markdown

## R Markdown syntax

- List item 1  
- List item 2  
  
\*\*Bold text\*\* and \*italicized text\* and ~~struck-through text~~  
  
`Monospace font for code or computer commands`  
  
> Box quote

* List item 1
* List item 2

**Bold text** and *italicized text* and ~~struck-through text~~

Monospace font for code or computer commands

Box quote

One tool to rule them all and in the R-ness, bind them.

## *Why* write reproducible papers/reports in R Markdown?

[**Mike Frank & Chris Hartgerink’s tutorial**](https://github.com/mcfrank/openscience_tutorial/blob/master/rmarkdown_handout.Rmd) **on GitHub**.

There are three reasons to write reproducible papers. To be right, to be reproducible, and to be efficient. There are more, but these are convincing to us. In more depth:

**To avoid errors**. Using an automated method for scraping APA-formatted stats out of PDFs, [Nuijten et al. 2015](http://doi.org/10.3758/s13428-015-0664-2) found that over 10% of p-values in published papers were inconsistent with the reported details of the statistical test, and 1.6% were what they called “grossly” inconsistent, e.g. difference between the p-value and the test statistic meant that one implied statistical significance and the other did not. Nearly half of all papers had errors in them.

**To promote computational reproducibility**. Computational reproducibility means that other people can take your data and get the same numbers that are in your paper. Even if you don’t have errors, it can still be very hard to recover the numbers from published papers because of ambiguities in analysis. Creating a document that literally specifies where all the numbers come from in terms of code that operates over the data removes all this ambiguity.

**To create spiffy documents that can be revised easily**. This is actually a really big neglected one for us. At least one of us used to tweak tables and figures by hand constantly, leading to a major incentive never to rerun analyses because it would mean re-pasting and re-illustratoring all the numbers and figures in a paper.

That’s a bad thing! It means you have an incentive to be lazy and to avoid redoing your stuff. And you waste tons of time when you do. In contrast, with a reproducible document, you can just rerun with a tweak to the code. You can even specify what you want the figures and tables to look like before you’re done with all the data collection (e.g., for purposes of preregistraion or a registered report).

## Ways to use R Markdown

* Web [pages](http://gilmore-lab.github.io/) | [Rmd](https://raw.githubusercontent.com/gilmore-lab/gilmore-lab.github.io/master/index.Rmd)
* [Talks](https://raw.githubusercontent.com/gilmore-lab/DEVSEC-2018/master/promise-of-open-dev-sci/index.Rmd) | [Rmd](https://raw.githubusercontent.com/gilmore-lab/DEVSEC-2018/master/promise-of-open-dev-sci/index.Rmd)
* [Project notes](https://github.com/gilmore-lab/infant-moco-eeg/blob/master/plan-of-work.Rmd)
* Documentation: [HTML](https://gilmore-lab.github.io/databrary-reqmts-specs/databrary_1.0.html) | [Rmd](https://raw.githubusercontent.com/gilmore-lab/databrary-reqmts-specs/master/databrary_1.0.Rmd)

## Ways to use R Markdown

* Scriptable figures: [PDF](https://github.com/PLAY-behaviorome/site-demographics/blob/master/img/ed-attain-bars-plot-1.pdf) | [Rmd](https://raw.githubusercontent.com/PLAY-behaviorome/site-demographics/master/educational_attainment.Rmd)
* Scriptable timelines from Datavyu files: [Example](https://github.com/PLAY-behaviorome/video-coding/blob/master/NYU020-timeline.md) | [Rmd](https://github.com/PLAY-behaviorome/video-coding/blob/master/video-code-file-export-cleaning.Rmd)
* Papers: [PDF](https://github.com/psu-psychology/r-bootcamp-2018/blob/master/talks/gilmore-hallquist-bootcamp-2018-papaja.pdf) | [Rmd](https://raw.githubusercontent.com/psu-psychology/r-bootcamp-2018/master/talks/gilmore-hallquist-bootcamp-2018-papaja.Rmd)

## The future of R-reproducible workflows

# Login to Databrary  
  
databraryapi::login\_db("rogilmore@psu.edu")  
  
# Download data  
  
# Process data

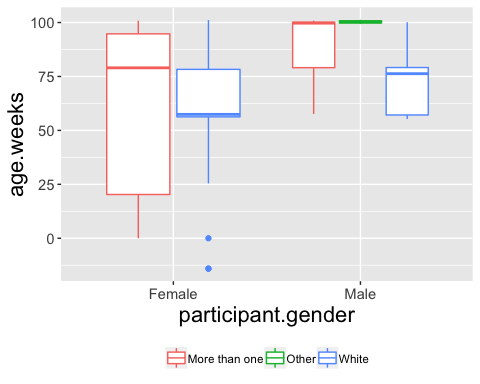
<https://github.com/PLAY-behaviorome/databraryapi>

lego\_vol <- 563  
databraryapi::list\_volume\_metadata(lego\_vol) %>%   
 kable()

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | name | owners | permission | doi |
| 563 | LEGO | Adolph, Karen; Tamis-LeMonda, Catherine | 4 | NA |

databraryapi::summarize\_demog(lego\_vol)

## Warning in `[<-.factor`(`\*tmp\*`, unreported, value = "Unknown or not  
## reported"): invalid factor level, NA generated



play\_pilot\_vol <- 444  
session\_id <- 18825  
(vol\_444\_sess\_18825\_assets <- databraryapi::list\_assets\_in\_session(session.id = session\_id,  
 vol.id = play\_pilot\_vol))

## asset.id asset.type.id asset.name  
## 1 88121 2 NYU-019-survey-demog  
## 2 86063 2 S#19\_DecibelData  
## 3 85853 -800 S#19\_House Map.mov  
## 4 85851 -800 S#19\_Dyadic Play.mov  
## 5 85863 -800 S#19\_Solitary Play.mov  
## 6 85849 -800 S#19\_1-Hour.mov  
## 7 85867 -800 S#19\_Questionnaires.mp4  
## 8 85865 -800 S#19\_Solitary&Dyadic Play (Experimenter).mov  
## 9 85855 -800 S#19\_1-Hour (Experimenter).mov  
## 10 85857 -800 S#19\_House Map (Experimenter).mov  
## 11 85861 -800 S#19\_Phone Questionnaire.mov  
## permission size duration segment vol.id session.id  
## 1 5 18310 NA NULL 444 18825  
## 2 5 880044 NA NULL 444 18825  
## 3 5 128407155 523136 69908000, 70431136 444 18825  
## 4 5 170298675 208267 70027000, 70235267 444 18825  
## 5 5 42970367 129259 70041000, 70170259 444 18825  
## 6 5 2958618399 3668800 78777000, 82445800 444 18825  
## 7 5 1758919170 3012864 82179000, 85191864 444 18825  
## 8 5 337009013 461590 85767000, 86228590 444 18825  
## 9 5 2106349637 3715840 86009000, 89724840 444 18825  
## 10 5 352815199 506646 86069000, 86575646 444 18825  
## 11 5 111188798 432064 89725840, 90157904 444 18825  
## mimetype extension asset.type transcodable  
## 1 text/csv csv Comma-separated values NA  
## 2 text/csv csv Comma-separated values NA  
## 3 video/mp4 mp4 MPEG-4 video -800  
## 4 video/mp4 mp4 MPEG-4 video -800  
## 5 video/mp4 mp4 MPEG-4 video -800  
## 6 video/mp4 mp4 MPEG-4 video -800  
## 7 video/mp4 mp4 MPEG-4 video -800  
## 8 video/mp4 mp4 MPEG-4 video -800  
## 9 video/mp4 mp4 MPEG-4 video -800  
## 10 video/mp4 mp4 MPEG-4 video -800  
## 11 video/mp4 mp4 MPEG-4 video -800

## Report on PLAY pilots and LEGO data

rmarkdown::render('PLAY-LEGO-report.Rmd', params = list(db\_account = "rogilmore@psu.edu"))

[Rendered Report in HTML](https://gilmore-lab.github.io/2018-10-15-nyu/PLAY-LEGO-report.html)

## But my advisor prefers MS Word…

rmarkdown::render('PLAY-LEGO-report.Rmd',   
 params = list(db\_account = "rogilmore@psu.edu"),  
 output\_format = "word\_document")

[MS Word version](PLAY-LEGO-report.docx)

## Key points

* Databrary API returns video/audio *segments*
* Video/audio segments can be extracted automatically from a Datavyu file
* (Soon) export updated Datavyu file with Databrary URIs
* R Markdown document can render code in [other languages](https://bookdown.org/yihui/rmarkdown/language-engines.html), including Python, bash, Octave, Julia, Haskell, JavaScript, …)
* Same input -> multiple output formats

## Learning more about R Markdown

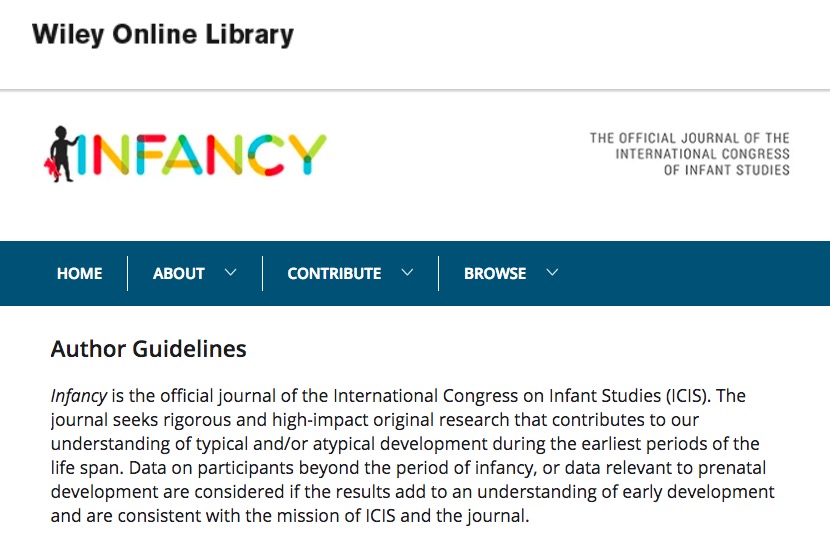
* [Online book](https://bookdown.org/yihui/rmarkdown/)

# Registered reports and pre-registration

## What



<https://cos.io/rr/>



<https://onlinelibrary.wiley.com/page/journal/15327078/homepage/ForAuthors.html>

## Why

* Nosek: “Don’t fool yourself”
* Separate *confirmatory* from *exploratory* analyses
* Confirmatory (hypothesis-driven): *p*-hacking matters
* Exploratory: *p*-values hard(er) to interpret

## How/where

* OSF: <http://help.osf.io/m/registrations/l/524205-register-your-project>
* [AsPredicted.org](https://aspredicted.org/)

# Psychological science (esp. developmental) can be a leader in the development of **r**obust and **r**eproducible **r**esearch practices

“*The advancement of detailed and diverse knowledge about the development of the world’s children is essential for improving the health and well-being of humanity. We regard scientific integrity, transparency, and openness as essential for the conduct of research and its application to practice and policy.*”

SRCD Task Force on Scientific Integrity and Openness

“*the principles of human subject research require an analysis of both risks and benefits…such an analysis suggests that researchers may have a positive duty to share data in order to maximize the contribution that individual participants have made.*”

([Brakewood & Poldack, 2013](http://dx.doi.org/10.1016/j.neuroimage.2013.02.040))

[rogilmore@psu.edu](mailto:rogilmore@psu.edu)

<https://gilmore-lab.github.io>

<https://gilmore-lab.github.io/2018-10-15-nyu/>

This talk was produced on 2018-10-14 18:56:05 in [RStudio 1.1.453](http://rstudio.com) using R Markdown and the reveal.JS framework. The code and materials used to generate the slides may be found at <https://github.com/gilmore-lab/DEVSEC-2018/promise-of-open-dev-sci/>. Information about the R Session that produced the slides is as follows:

## R version 3.5.1 (2018-07-02)  
## Platform: x86\_64-apple-darwin15.6.0 (64-bit)  
## Running under: macOS Sierra 10.12.6  
##   
## Matrix products: default  
## BLAS: /System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versions/A/libBLAS.dylib  
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib  
##   
## locale:  
## [1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8  
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] bindrcpp\_0.2.2 knitr\_1.20   
## [3] forcats\_0.3.0 stringr\_1.3.1   
## [5] dplyr\_0.7.6 purrr\_0.2.5   
## [7] readr\_1.1.1 tidyr\_0.8.1   
## [9] tibble\_1.4.2 ggplot2\_3.0.0   
## [11] tidyverse\_1.2.1 databraryapi\_0.1.3.9006  
## [13] revealjs\_0.9   
##   
## loaded via a namespace (and not attached):  
## [1] tidyselect\_0.2.4 reshape2\_1.4.3 haven\_1.1.2 lattice\_0.20-35   
## [5] colorspace\_1.3-2 htmltools\_0.3.6 yaml\_2.2.0 utf8\_1.1.4   
## [9] rlang\_0.2.2 pillar\_1.3.0 withr\_2.1.2 glue\_1.3.0   
## [13] modelr\_0.1.2 readxl\_1.1.0 jpeg\_0.1-8 bindr\_0.1.1   
## [17] plyr\_1.8.4 munsell\_0.5.0 gtable\_0.2.0 cellranger\_1.1.0  
## [21] rvest\_0.3.2 evaluate\_0.11 labeling\_0.3 curl\_3.2   
## [25] fansi\_0.3.0 highr\_0.7 broom\_0.5.0 Rcpp\_0.12.18   
## [29] scales\_1.0.0 backports\_1.1.2 jsonlite\_1.5 hms\_0.4.2   
## [33] digest\_0.6.16 stringi\_1.2.4 keyring\_1.1.0 grid\_3.5.1   
## [37] rprojroot\_1.3-2 cli\_1.0.0 tools\_3.5.1 magrittr\_1.5   
## [41] lazyeval\_0.2.1 crayon\_1.3.4 pkgconfig\_2.0.2 xml2\_1.2.0   
## [45] lubridate\_1.7.4 rstudioapi\_0.7 assertthat\_0.2.0 rmarkdown\_1.10   
## [49] httr\_1.3.1 R6\_2.2.2 nlme\_3.1-137 compiler\_3.5.1