

# Using R for Corpus Linguistics - an Introduction and Discussion Note on Sustainability and Replicability in Corpus Linguistics

Martin Schweinberger  
The University of Queensland  
m.schweinberger@uq.edu.au

CoEDL: Corpus Workshop



ARC CENTRE OF EXCELLENCE FOR  
**THE DYNAMICS OF LANGUAGE**



# Aims of this talk

- ▶ One of my core concerns: “Best Practices”  
(with respect to research technology and data analysis in linguistics and language studies)
- ▶ Raise awareness for best practices
- ▶ Start a discussion about issues related to best practices
- ▶ Introduce R as a remedy to some issues related to best practices. . .

# Replication crisis (RC)

... ongoing methodological crisis primarily affecting parts of the social and life sciences beginning in the early 2010s.

- ▶ growing awareness of the problem that results of many scientific studies are difficult or impossible to replicate/reproduce.
- ▶ reproducibility is an essential part of the scientific method,
- ▶ inability to replicate the studies of others has potentially grave consequences for many fields of science in which significant theories are grounded on unreproducible work.

# Replication crisis (RC)



# Replication crisis (RC)

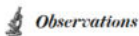
SCIENTIFIC  
AMERICAN.

HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCASTS

## More social science studies just failed to replicate. Here's why this is good.

What scientists learn from failed replications: how to do better science.

By Brian Resnick | @B\_Resnick | brian@vox.com | Aug 27, 2018, 11:00am EDT



## (Dis)trust in Science

Can we cure the scourge of misinformation?

By Gleb Tsipursky on July 5, 2018

# Replication crisis (RC)

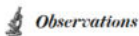
SCIENTIFIC  
AMERICAN.

HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCAS

## More social science studies just failed to replicate. Here's why this is good.

What scientists learn from failed replications: how to do better science.

By Brian Resnick | @B\_Resnick | brian@vox.com | Aug 27, 2018, 11:00am EDT



## (Dis)trust in Science

Can we cure the scourge of misinformation?

By Gleb Tsipursky on July 5, 2018

NOBA

NOBA Online - The Replication Crisis in Psychology

## The Replication Crisis in Psychology

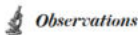
By Edward Olfend and Robert Brown-Olfend  
University of Utah, University of Virginia, Portland State University

# Replication crisis (RC)

SCIENTIFIC  
AMERICAN.

HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCASTS

## More social science studies just failed to replicate. Here's why this is good.



What scientists learn from failed replications: how to do better science.

By Brian Resnick | @B\_Resnick | bresnick.com | Aug 27, 2015, 11:00am EDT

## (Dis)trust in Science



## The Replication Crisis in Psychology

By Edward O'Keefe and Robert Brown O'Keefe  
University of Utah, University of Virginia, Portland State University

AMERICAN PSYCHOLOGICAL ASSOCIATION

MEMBERS TOPICS PUBLICATIONS & DATABASES PSYCHOLOGY HELP CENTER NEWS & EVENTS

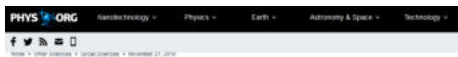
Home / Monitor on Psychology / 2015 / 10 / A reproducibility crisis?

## A reproducibility crisis?

The headlines were hard to miss: Psychology, they proclaimed, is in crisis.

October 2015, Vol. 46, No. 9  
Print version: page 39

# Replication crisis (RC)



SCIENTIFIC  
AMERICAN.

TAINABILITY EDUCATION VIDEO PODCASTS

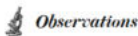
**Researcher discusses the the science replication crisis**

November 21, 2018 by Emily Raposo, California Institute of Technology

**MORE SOCIAL SCIENCE STUDIES JUST FAILED TO REPLICATE. Here's why this is good.**

What scientists learn from failed replications: how to do better science.

By Brian Resnick | @B\_Resnick | bresnick@uconn.edu | Aug 27, 2018, 11:00am EDT



## (Dis)trust in Science

Can we cure the scourge of misinformation?

By Gleb Tsipursky on July 5, 2018



## The Replication Crisis in Psychology

By Edward Dener and Robert Brown-Dener  
University of Utah, University of Virginia, Portland State University



# Replication crisis (RC)

PHYS ORG  
Nanotechnology - Physics - Earth - Astronomy & Space - Technology

Researcher discusses the the science replication crisis  
November 21, 2018 in Emily Weiss, California Institute of Technology

**more social science studies just failed to replicate. Here's why this is good.**

What scientists learn from failed replications: how to do better science.  
By Brian Resnick | @B\_Resnick | brian@vox.com | Aug 27, 2018, 11:00am EDT

SCIENTIFIC AMERICAN.

TAINABILITY EDUCATION VIDEO PODCAS

*Observations*

## (Dis)trust in Science

**More social science studies just failed to replicate. Here's why this is good.**

What scientists learn from failed replications: how to do better science.  
By Brian Resnick | @B\_Resnick | brian@vox.com | Aug 27, 2018, 11:00am EDT

**Technology**

By Edward Omer and Robert Weiss-Omer  
University of Utah, University of Virginia, Portland State University

urge of misinformation?

---

rsky on July 5, 2018

# Replication crisis (RC)

PHYS ORG Nanotechnology Physics Earth Astronomy & Space Technology

November 21, 2018 to Emily Weiss, California Institute of Technology

**Researcher discusses the the science replication crisis**

**more social science studies just failed to replicate. Here's why this is good.**

What scientists learn from failed replications: how to do better science.

By Brian Resnick | @B\_Resnick | brian@hox.com | Aug 27, 2018, 11:00am EDT

**(DIS)U**

**More social science studies just fa  
replicate. Here's why this is good.**

What scientists learn from failed replications: how to do better science.

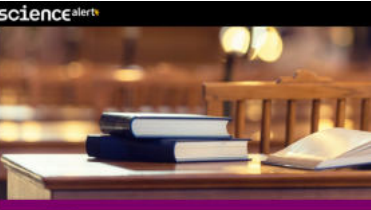
By Brian Resnick | @B\_Resnick | brian@hox.com | Aug 27, 2018, 11:00am EDT

By Edward Omeier and Robert Brown-Omeier  
University of Utah, University of Virginia, Portland State University

SCIENTIFIC AMERICAN.

SUSTAINABILITY EDUCATION VIDEO PODCASTS

sciencealert



HUMANIS

**Science's 'Replication Crisis' Has Reached Even The Most Respectable Journals, Report Shows**

MIKE MORAY 27 AUG 2018



## Researcher discusses the the science replica

November 21, 2018 by Emily Velasco, California Institute of Technology

### FiveThirtyEight

Politics Sports **Science & Health** Economics Culture

Our 2018 March Madness

DEC 8, 2018, 8:11:10 AM

## Psychology's Replication Crisis Has Made The Field Better

by [Charlotte Aschewander](#)

## More social science studies just fa replicate. Here's why this is good.

What scientists learn from failed replications: how to do better science.

By Brian Resnick | [@B\\_Resnick](#) | [brian@fivox.com](#) | Aug 27, 2018, 11:00am EDT



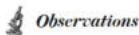
## The Replication Crisis in More social science studies just failed to replicate. Here's why this is good.

What scientists learn from failed replications: how to do better science.

By Brian Resnick | [@B\\_Resnick](#) | [brian@fivox.com](#) | Aug 27, 2018, 11:00am EDT

## SCIENTIFIC AMERICAN

HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCASTS



# (Dis)trust in Science

Can we cure the scourge of misinformation?

By Gleb Tspursky on July 5, 2018



HUMANS

## Science's 'Replication Crisis' Has Reached Even The Most Respectable Journals, Report Shows

MIKE MORSE | 27 AUG 2018

# Replication crisis (RC)

*Nature* 2016 poll of 1,500 scientists

- ▶ 70% had failed to reproduce at least one other scientist's experiment
- ▶ 50% had failed to reproduce one of their own experiments (cf. Fanelli 2009)

2009 meta-analysis of surveys on science fraud (Fanelli 2009)

- ▶ 2% admitted to falsifying studies at least once
- ▶ 14% admitted to personally knowing someone who did

More importantly: data analysis is often too lengthy/complex to describe in detail...

# Replication crisis (RC)

So, what about *Linguistics*???

## Problem

We just do not know how bad our science is. . .  
(outright forgery, data manipulation, p-hacking, etc.)  
because we do not (or only rarely)  
reproduce and replicate. . .

assuming you are a corpus linguist

# RC in Linguistics

## Good

- ▶ blind peer-review
- ▶ we are open and share if we are asked (sometimes)
- ▶ discussion has begun (cf. e.g. Berez-Kroeker et al. 2018)

## Bad

- ▶ analyses are not reproducible/replicated
- ▶ reliance on tools not scripts
- ▶ reproduction is discouraged  
(if successful: journals are not interested in publishing the same analysis twice/several times;  
if unsuccessful: researchers do not want to threaten the face of other researchers)

## Solutions

### Open access

Access to data sets to enable reproduction

### Scripting

Scripts rather than tools

### Publication

Ability to reproduce/replicate should be mandatory



# Open access

- ▶ Access to data sets to enable replication (see Berez-Kroeker et al. 2018: for a more extensive discussion on this point)
- ▶ Access should be easy (not only for programmers!)
- ▶ (Open) Public Repositories  
data sets/corpora/raw data should be made available for replication (within ethical boundaries)
- ▶ Corpora should be treated as publications and should be cited as such (increases citations and makes it more attractive to publish data sets/corpora)
- ▶ Papers that rely on data that is not available should not be published in journals (pressure on publishing houses or other outlets)

# Scripting

- ▶ Scripts allow exact replication (total transparency)
- ▶ Only practical solutions for true replication (too time consuming to replicate a tool-based analysis)
- ▶ Data analysis is too fine-grained to be described in papers (including all steps the researcher has undertaken)
- ▶ Training programs for basic programming at universities/schools (obligatory for grad programs)

```
#install.packages(Rling) # install Rling library (remove # to activate)
library(Rling) # activate Rling library
library(partykit) # activate partykit library
library(dplyr) # activate dplyr library
options(stringsAsFactors = T) # set options: do not convert strings
options(scipen = 999) # set options: suppress math. notation
options(max.print=10000) # set options
# load data
citdata <- read.delim("data/treedata.txt", header = T, sep = "\t")
head(citdata)
```

# Publication

- ▶ If we want the “Linguistics” community to become more science-like we need to change our practices as a community
- ▶ No publication of non-replicable research!
- ▶ Publication of null results must be encouraged (somehow)
- ▶ Replication should be a common practice especially during BA/MA (students learn how more advanced researchers have handled problems and conducted research)
- ▶ Install best practices
- ▶ “Center for Quality Assurance” or sth. like that where people can voice concerns about research practices
- ▶ Results of any replication should be published (maybe even in open source online venues)

# Why R?



Allows full transparency and replication of research

- ▶ Open source
- ▶ Free-ware
- ▶ Scripts can be shared easily
- ▶ Allows full transparency because all steps of the analysis are available
- ▶ A human/user-centered language ( $\neq$  C and daughters or Java)

For Linguists

- ▶ Usable for many different glyph systems (unicode)
- ▶ Can be used to create and curate corpora
- ▶ Allows complex text analysis/data analysis/data viz (including geo mapping)

# Why R?



Allows full transparency and replication of research

- ▶ One of the fastest growing world's top 10 programming environments
- ▶ Enormous support community (StackOverflow, etc.)
- ▶ Extreme flexibility of methods (thousands of packages)
- ▶ Variability in output (statistics, visualizations, text analysis, speech analysis, websites, slides, apps, netbooks, etc.)
- ▶ Compatibility with other software (PRAAT, MAUS, Office apps, etc.)

# R in HASS



Every journey begins with a first step and, step by step, we can go miles on end!

- ▶ Packages for text analysis are readily available
- ▶ Complex issues can be broken down into simple chunks
- ▶ Very easy to learn (steep or shallow learning curve)
- ▶ Even very basic skills allow performing complex analyses

# Solutions at UQ



- ▶ Training program: workshops on R ✓/X  
(for all levels of expertise *Center for Digital Scholarship/School of Languages and Cultures*)
- ▶ Materials ✓/X  
Language Technology and Data Analysis Laboratory (LADAL) website (data and text analysis with R:  
<https://slcladal.github.io/index.html>)
- ▶ Study program X (beginning to plan a program)  
Digital HASS (BA/MA program including modules on data and text analysis with R)

- Aschwanden, C. (2018). Psychology's replication crisis has made the field better.
- Berez-Kroeker, A. L., L. Gawne, S. S. Kung, B. F. Kelly, T. Heston, G. Holton, P. Pulsifer, D. I. Beaver, S. Chelliah, S. Dubinsky, et al. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1), 1–18.
- Diener, E. and R. Biswas-Diener (2019). The replication crisis in psychology.
- Fanelli, D. (2009). How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLoS One* 4, e5738.
- McRae, M. (2018). Science's 'replication crisis' has reached even the most respectable journals, report shows.
- Resnick, B. (2018). More social science studies just failed to replicate. here's why this is good.what scientists learn from failed replications: how to do better science.
- Velasco, E. (2019). Researcher discusses the the science replication crisis.
- Weir, K. (2015). A reproducibility crisis? the headlines were hard to miss: Psychology, they proclaimed, is in crisis. *Monitor on Psychology* 46, 39.
- Yong, E. (2018). Psychology's replication crisis is running out of excuses. another big project has found that only half of studies can be repeated. and this time, the usual explanations fall flat.



So, what do you think???

Comments? Feedback? Suggestions?