The Replication Crisis and HASS

How Best Practices can Assist in Producing Reliable Research

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

Open Data Day

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 in HASS
- Start a discussion about issues related to Best Practices
- Introduce R as a remedy to some issues related to best practices. . .

What is the Replication Crisis?

Replication Crisis

. . . 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

SCIENTIFIC AMERICAN

HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCAS

Observations

(Dis)trust in Science

Can we cure the scourge of misinformation?

By Gleb Tsipursky on July 5, 2018

Replication Crisis

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

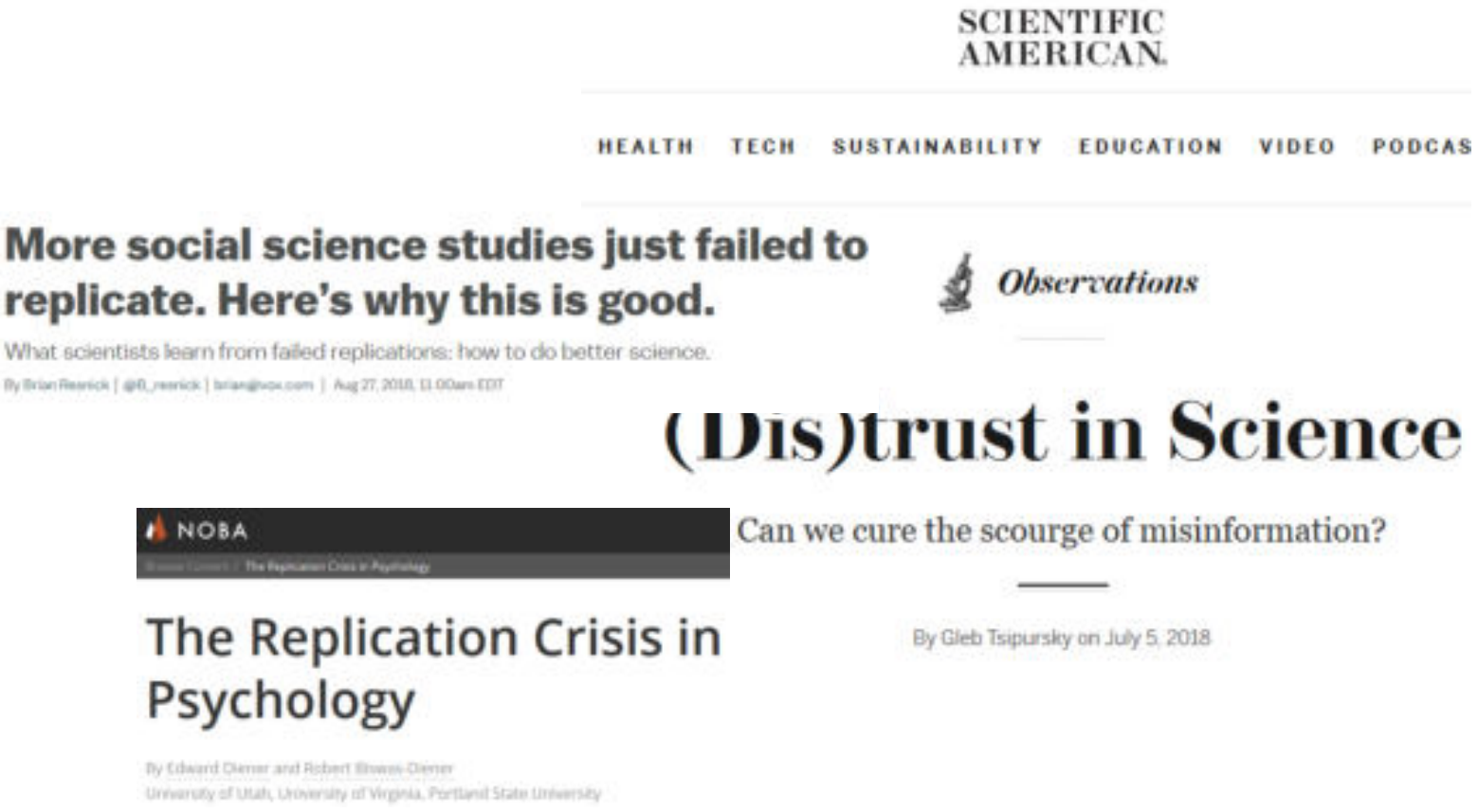
Observations

(Dis)trust in Science

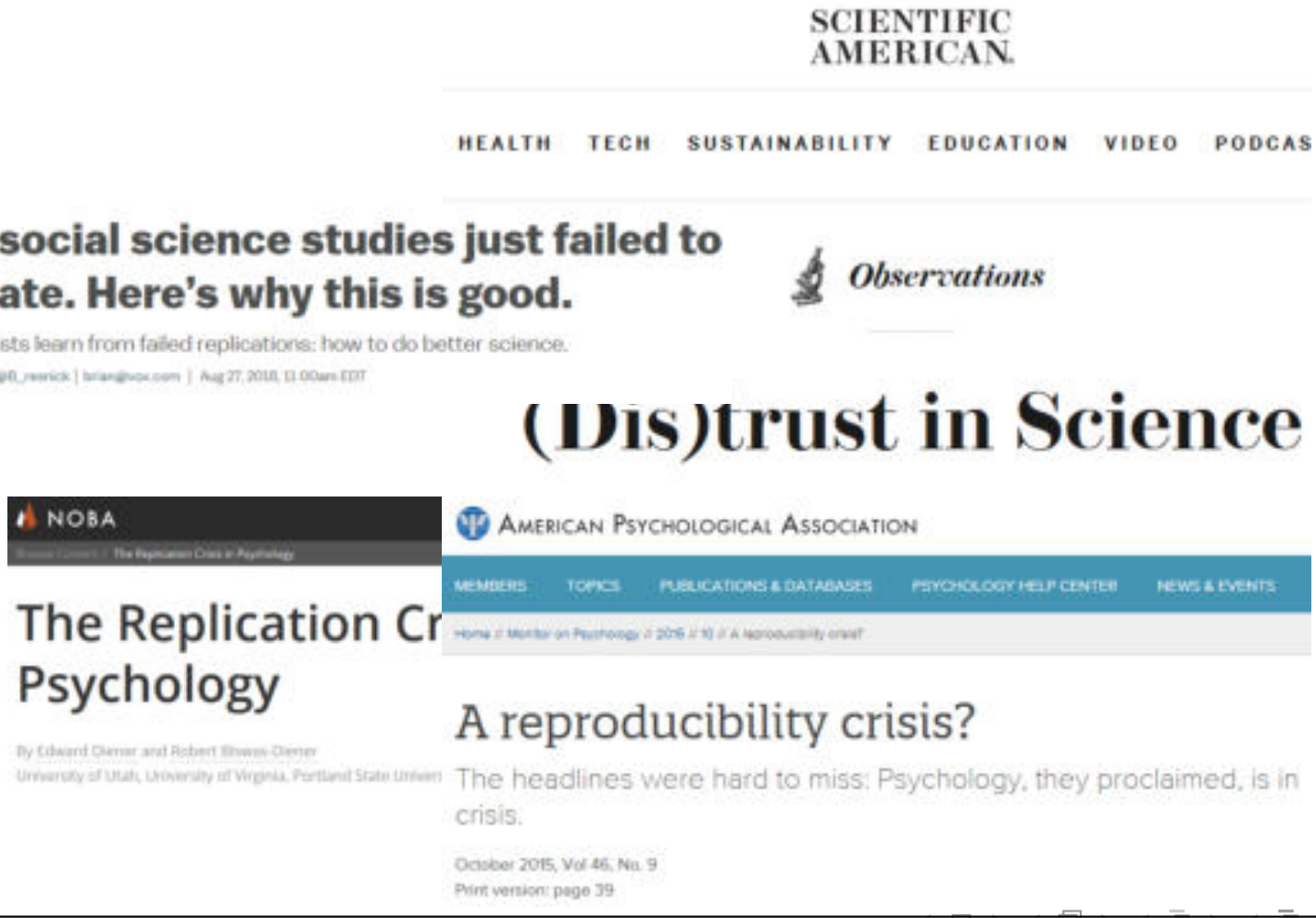
Can we cure the scourge of misinformation?

By Gleb Tsipursky on July 5, 2018

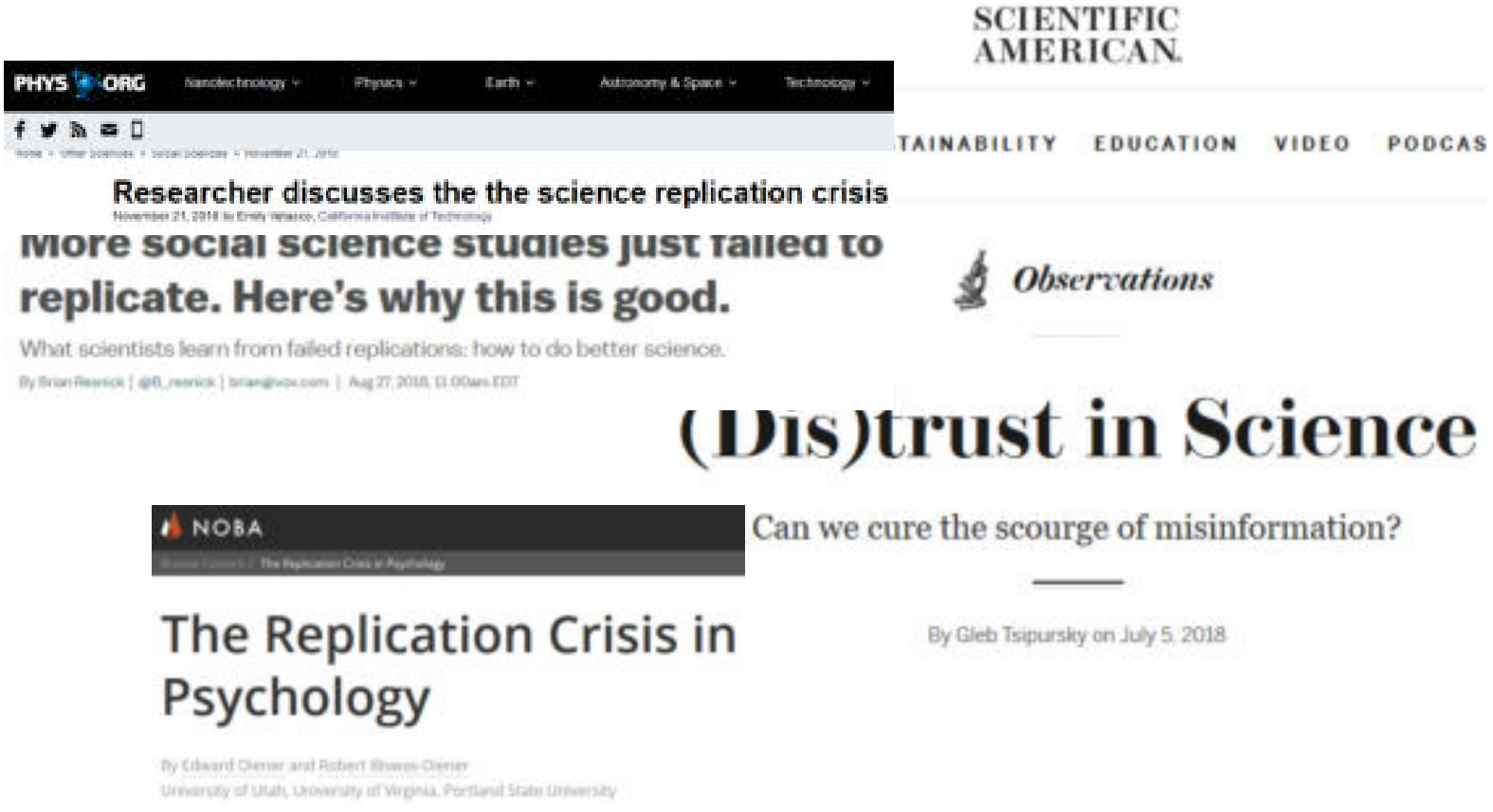
Replication Crisis



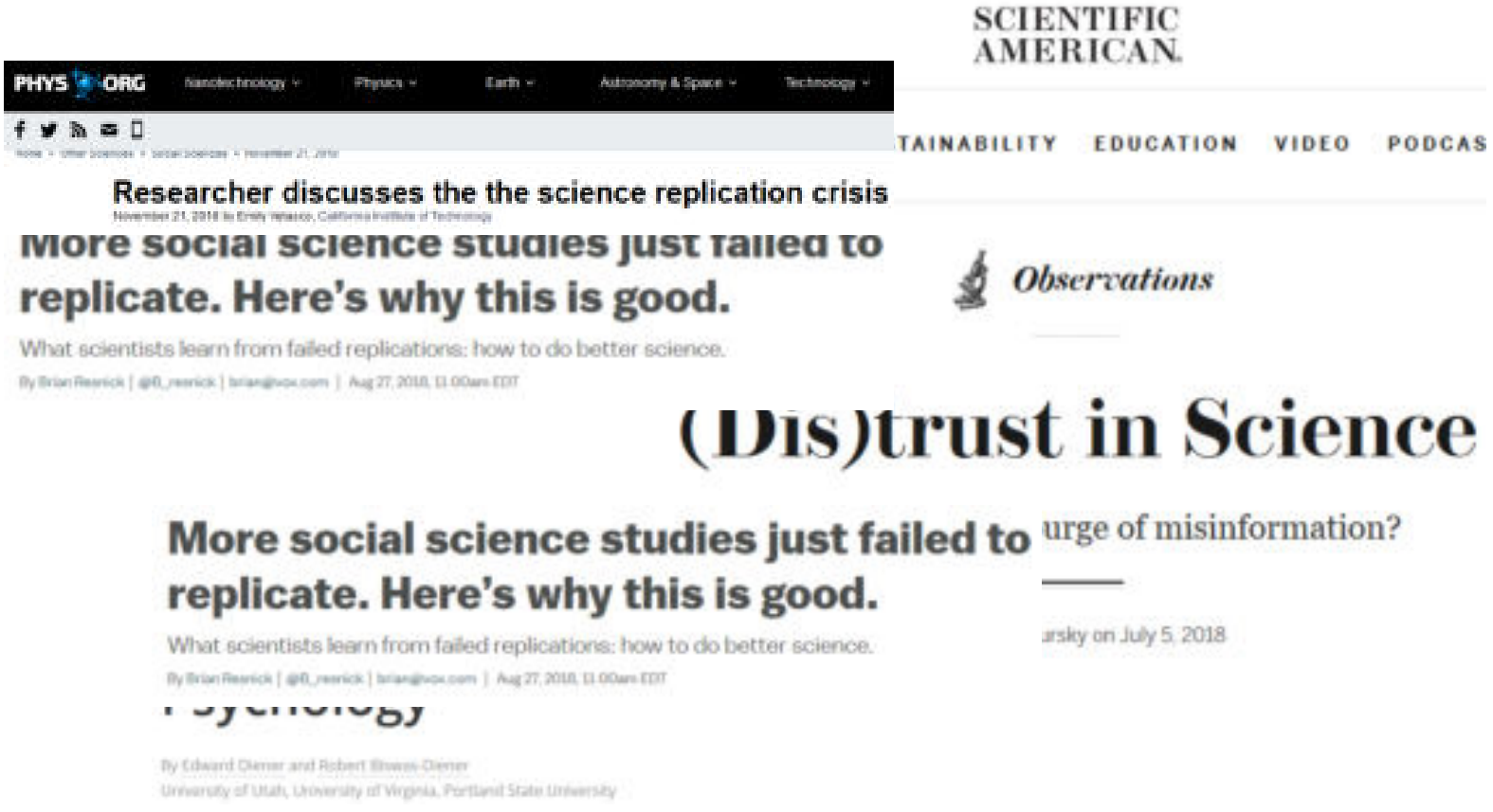
Replication Crisis



Replication Crisis



Replication Crisis



Replication Crisis

PHYS.ORG

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@vox.com | Aug 27, 2018, 11:00am EDT

(Dis)t

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@vox.com | Aug 27, 2018, 11:00am EDT

Psychology

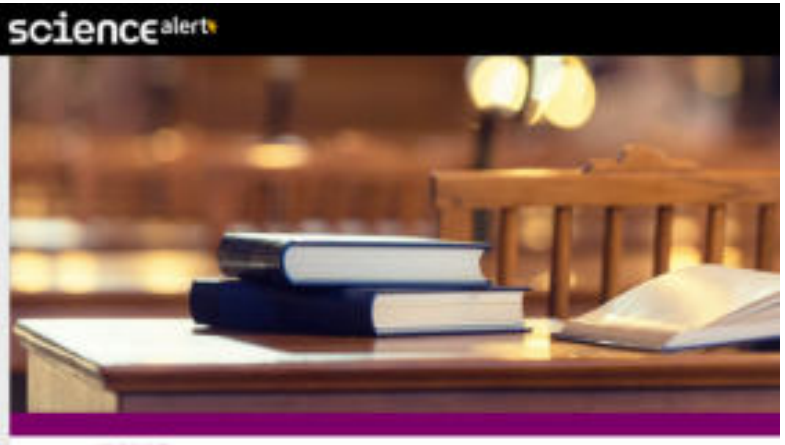
By Edward Diener and Robert Biswas-Diener

University of Utah, University of Virginia, Portland State University

SCIENTIFIC AMERICAN

TAINABILITY EDUCATION VIDEO PODCAS

sciencealert

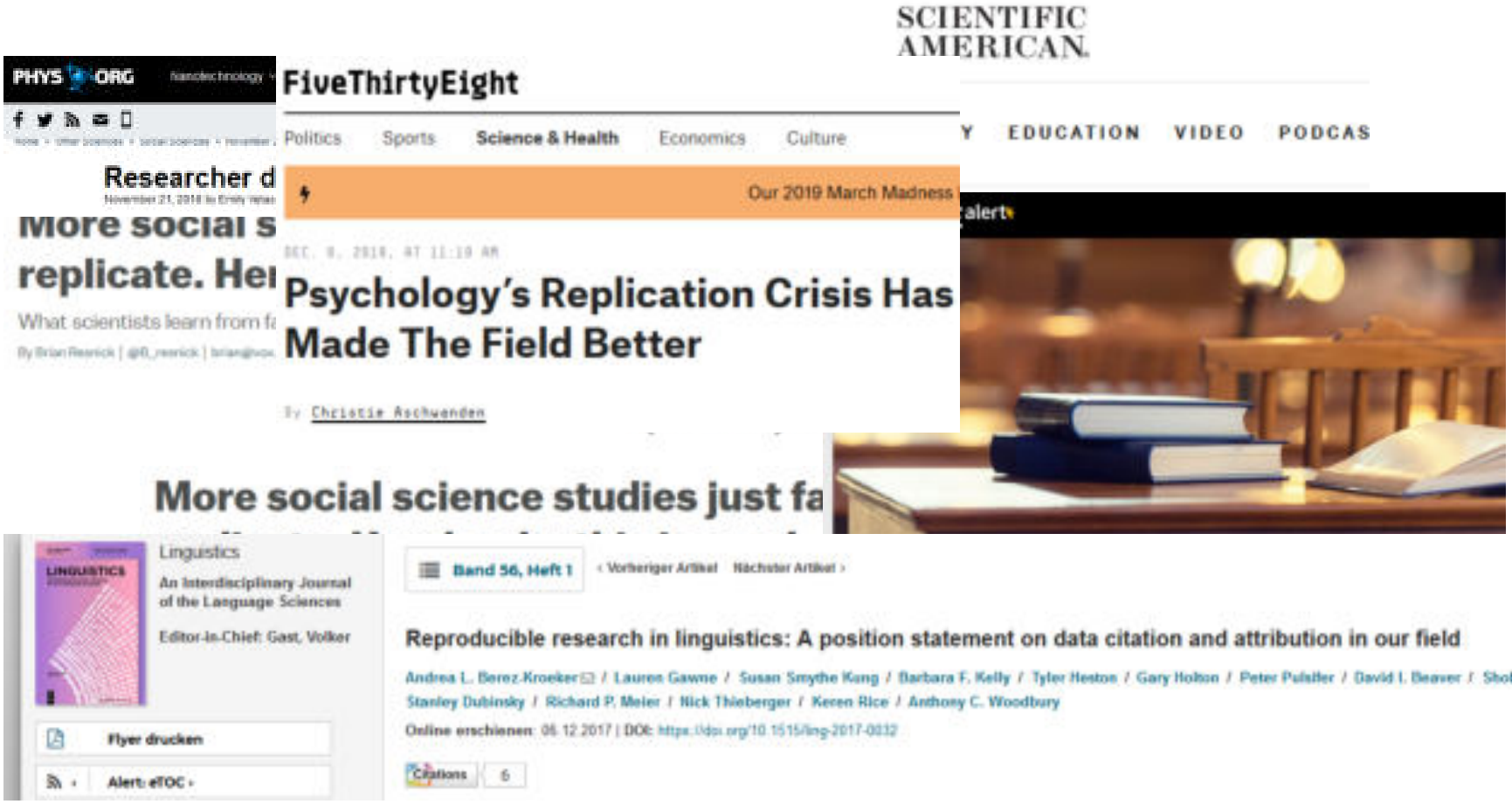


HUMANS

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

MIKE MCRAE 27 AUG 2018

Replication Crisis



Replication Crisis

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

Loss of (public) trust!

What about the Humanities?

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 studying language

Replication Crisis in Language Studies

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)

How can we solve this issue?

Solutions

Open access (FAIR data)
Access to data sets to enable reproduction

Publication
Ability to reproduce/replicate should be mandatory

Scripting / Code
Scripts rather than tools

Open access (FAIR data)

- 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)

Publication

If we want the HASS community to adopt Best Practices we need to change as a community

- No publication of non-replicable research!
- Publication of null results must be encouraged (pre-registration)
- Results of all replications should be published
- Replication should be a common practice especially during BA/MA (students learn how more advanced researchers have handled problems and conducted research)
- Installing best practices: extensive support for training programs
- "Center for Quality Assurance" or sth. like that where people can voice concerns about research practices

Scripting / Code

- 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: supress math. notation
options(max.prAmplified=10000) # set options
# load data
citdata <- read.delim("data/treedata.txt", header = T, sep = "\t")
head(citdata)
```

Why R?

Allows full transparency and replication of research

- Open source free-ware ($\neq$ SPSS)
- Scripts can be shared easily (easily connected to Git)
- Allows full transparency because all steps of the analysis are available
- A human/user-centered language ($\neq$ the C family or Java)
- Fully-fledged programming environment
- Not teaching researchers to use software but to create software $\rightarrow$ flexibility, independence, employability!

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 packages that are common in Language Studies (PRAAT, MAUS, etc.)

Why R?

For HASS (Language Studies)

- Combines the advantages of Python, Stata/MatLab and Gephi:

  Offers the same functionality of Python (NLP) but is (arguably) better at complex data analysis (Stata/MatLab/SPSS) and data viz (including geo mapping) (e.g. Gephi)

- Is already wide-spread in the community!
- Usable for many different glyph systems (unicode)
- Can be used to create and curate corpora

R in HASS

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

- Packages for text analysis/NLP/data viz/statistics 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.

The Replication Crisis and HASS

How Best Practices can Assist in Producing Reliable Research

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

Open Data Day