Best Practices in Corpus Linguistics

What lessons should we take from the Replication Crisis and how can we guarantee high quality in our research?

Dr. Martin Schweinberger (m.schweinberger@uq.edu.au)

(see also www.martinschweinberger.de)

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

CREATE CHANGE

Aims, definition, and the current state of affairs

This presentation aims to

- Raise awareness for *Best Practices* in Corpus Linguistics
- Discuss issues related to Best Practices and Replicability
- Propose improvements to current research practices
- Offer solutions on how best practices can be implemented

Aims, definition, and the current state of affairs

What are best practices?

A **best practice** is a method or technique that is superior to alternatives because it produces results that are more reliable, transparent, replicable, and in compliance with legal or ethical requirements.

Replication Crisis

Aims, definition, and the current state of affairs

Best practices have come into focus as a result of the **Replication Crisis** (RC). The Replication Crisis is an ongoing methodological crisis primarily affecting parts of the social and life sciences beginning in the late 2000s.

Nature 2016 poll of 1,500 scientists:

- 70% failed to reproduce at least one other scientist's experiment
- 50% failed to reproduce one of their own experiments

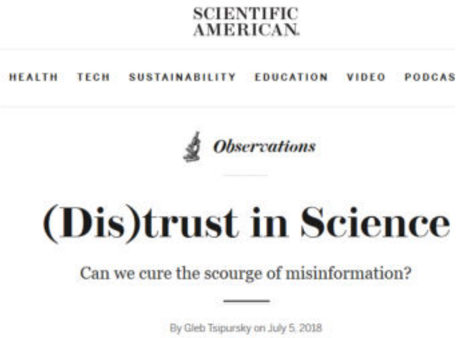
Meta-analysis of surveys on science fraud (Fanelli 2009)

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

Repercussions

As a consequence of the RC, there is growing awareness...

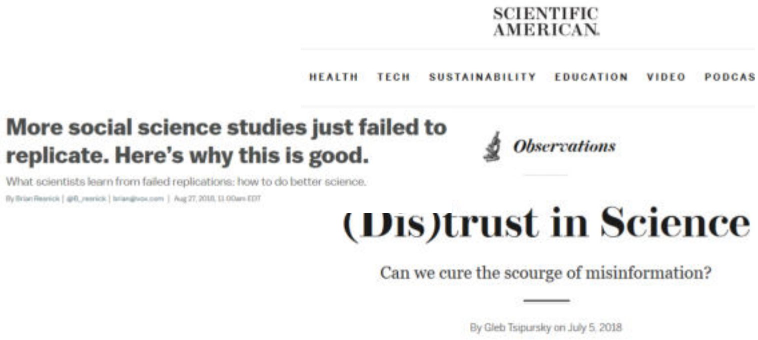
- of a problem: currently most research is difficult to replicate/reproduce!
- that reproducibility is an essential part of the scientific method
- that the inability to replicate has potentially grave consequences as significant theories are grounded on unreproducible work
- that there is substantial loss of trust in science, its results, and its proponents.



(Dis)trust in Science

Can we cure the scourge of misinformation?

By Gleb Tsipursky on July 5, 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.

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

SCIENTIFIC AMERICAN.

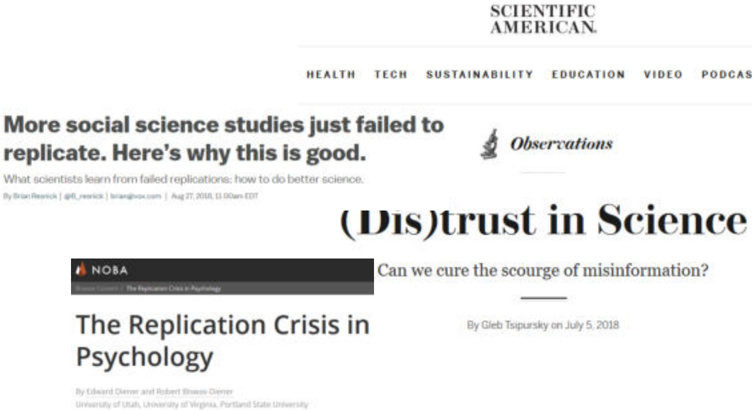
HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCAS

Observations

(Dis)trust in Science

Can we cure the scourge of misinformation?

By Gleb Tsipursky on July 5, 2018



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

NOBA

Browse Content / The Replication Crisis in Psychology

The Replication Crisis in Psychology

By Edward Diener and Robert Biswas-Diener
University of Utah, University of Virginia, Portland State University

SCIENTIFIC
AMERICAN.

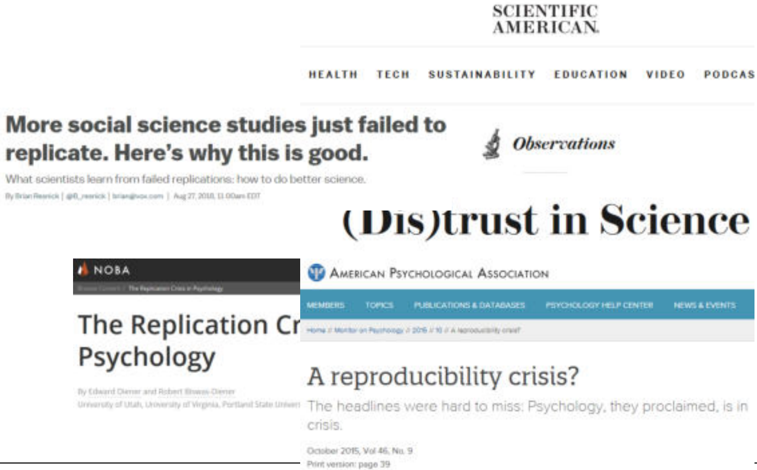
HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCAS

Observations

(Dis)trust in Science

Can we cure the scourge of misinformation?

By Gleb Tsipursky on July 5, 2018



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

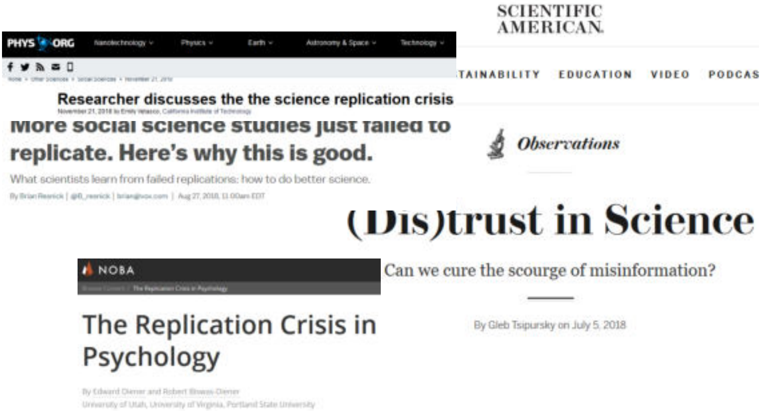
The Replication Cr
Psychology

By Edward Diener and Robert Biswas-Diener
University of Utah, University of Virginia, Portland State Univers

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



Researcher discusses the the science replication crisis

November 21, 2018 by Emily Velasco, 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

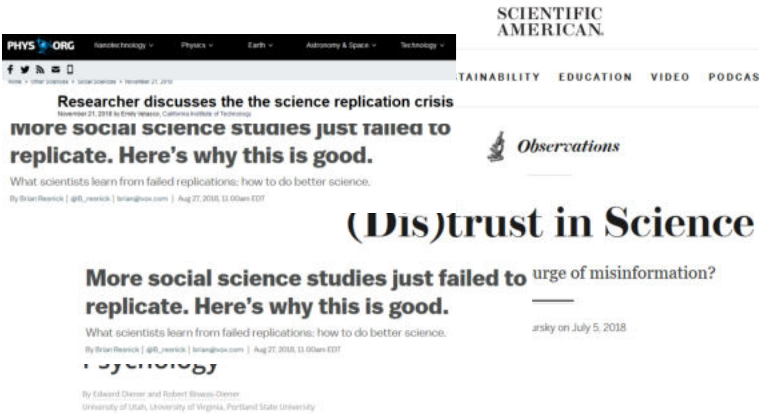
(Dis)trust in Science

Can we cure the scourge of misinformation?

By Gleb Tsipursky on July 5, 2018

The Replication Crisis in Psychology

By Edward Diener and Robert Biswas-Diener
University of Utah, University of Virginia, Portland State University



Home • Other Sciences • Social Sciences • November 21, 2018

Researcher discusses the the science replication crisis

November 21, 2018 by Emily Velasco, 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

(Dis)trust in Science

urge of misinformation?

ursky on July 5, 2018

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

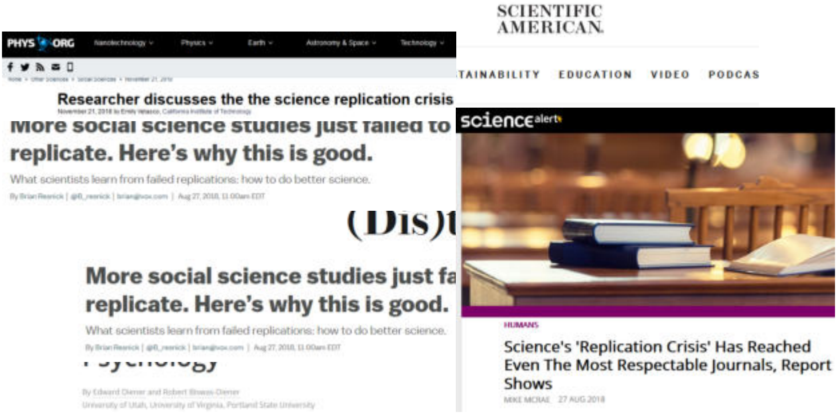
What scientists learn from failed replications: how to do better science.

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

PHYS ORG Nanotechnology Physics Earth Astronomy & Space Technology

Home • Other Sciences • Social Sciences • November 21, 2018

Researcher discusses the the science replication crisis

November 21, 2018 by Emily Velasco, 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

(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

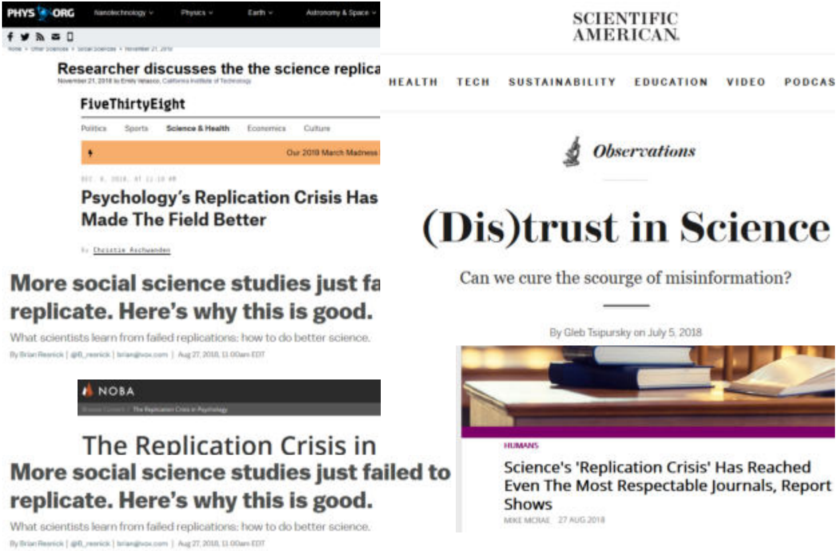
University of Utah, University of Virginia, Portland State University

sciencealert

HUMANS

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

MIKE MCRAE 27 AUG 2018



Researcher discusses the the science replica

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

Psychology's Replication Crisis Has Made The Field Better

By Christie Aschwanden

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

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@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

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

MIKE MCRAE 27 AUG 2018

Best Practices in Corpus Linguistics

Best Practices in Corpus Linguistics

As a community, we endorse *blind peer-review*, we are *open to sharing* (if we are asked), and we have *begun with a discussion around best practices* and replication (Berez-Kroeker et al. 2018; Ruhi et al. 2014).

The E-MELD School of

Best Practices in Digital Language Documentation

Your Location: School | E-MELD School of Best Practice

View comments | Add a comment | Print page

Entrance

Case Studies

Classroom

Reading Room

Work Room

Tool Room

Ask an Expert

Endangered Languages / Endangered Documentation

Classroom

Case Studies

Work Room

Welcome

This site promotes Best Practices in preserving language documentation. Mouse-over an image to learn more about each School room.

Reading Room

Tool Room

Ask-an-Expert

Linguistics | Band 56: Heft 1

Reproducible research in linguistics: A position statement on data citation and attribution in our field

Andrea L. Berez-Kroeker, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice and Anthony C. Woodbury

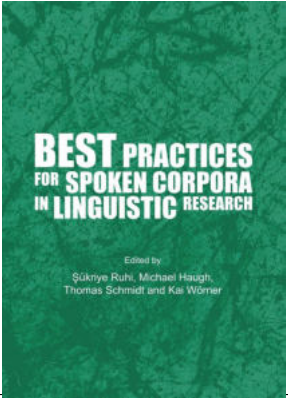
DOI: https://doi.org/10.1515/ling-2017-0032 | Online veröffentlicht: 06.12.2017

OPEN ACCESS

ZEITSCHRIFT + HEFTE

Linguistics

Band 56: Heft 1



BEST PRACTICES FOR SPOKEN CORPORA IN LINGUISTIC RESEARCH

Edited by

Şükriye Ruhi, Michael Haugh, Thomas Schmidt and Kai Wörner

Best Practices in Corpus Linguistics

However, we could be better. . .

- *analyses often not reproducible*
- *over-reliance on tools*
- *reproduction is discouraged*
  - (i) journals are not interested in publishing the same analysis twice;
  - (ii) researchers fear repercussions if they criticize the research of others (face-threatening).

Best Practices in Corpus Linguistics

While replicability has improved with the rise of natural language corpora, **we just do not know how bad our research is** (mistakes in using statistical methods or data processing, outright forgery, data manipulation, p-hacking, etc.) because . . .

1. researchers do not (or only rarely) reproduce and replicate
2. researchers do not know about best practices or what they are
3. researchers do not know how to make their research comply with best practices
4. lack of training in best practices and how to make research reproducible

Suggestions to make our research more replicable

For individual researchers and teams

- **FAIR principles**
  data should be ***F****indable*, ***A****ccessible*, ***I****nteroperable*, ***R****eusable* (FAIR) (Wilkinson et al. 2016)
- **Data a publication**
  - i clear example for how to cite your data
  - ii publish it on an online repository (this way your data is a proper publication)
  - iii assign a *Digital Object Identifier* (DOI) to your data.
- **Notebooks and version control**
  - i use R or Jupyter Notebooks
  - ii share your projects on GitHub
    (this way, your research is fully transparent and reproducible)

For individual researchers and teams

- **Scripts over tools**
  R rather than ready-made tools (tools are black-boxes that hinder replication due to limited accessibility and/or time-restraints)
- **Documentation**
  document what you do, where you find stuff, and who to ask for help
- **Archiving**
  use online repositories (*GitHub*, etc.) to avoid data loss and various versions of a single document or file

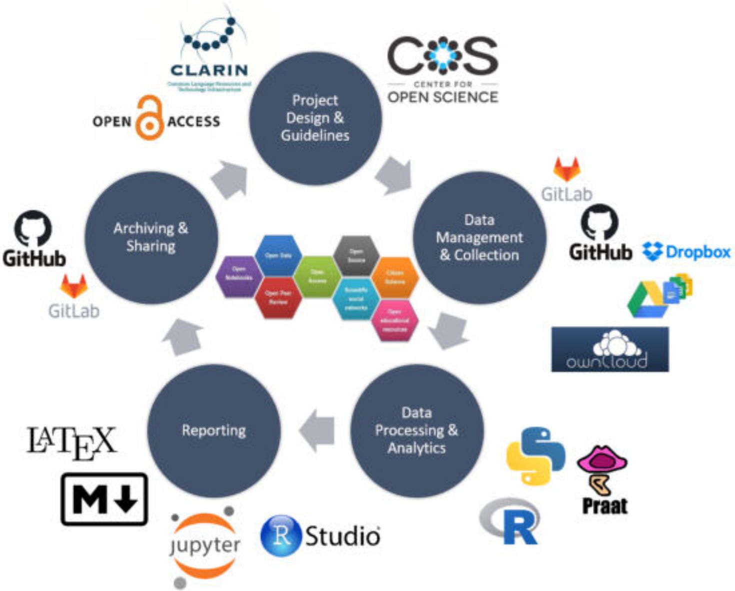
DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
JORGE CHAM © 2014
WWW.PHDCOMICS.COM

For the community

Endorse *Open Science*
Open Data + Open Access + Open Methodology + Open Educational Resources

- Ask for data and scripts when reviewing papers
- Cite appropriately (rewards publication of corpora)
- Promote and support replication
- Invest in/support training in data management and transparent data analysis options
  (*R*, *Git*, *Markdown*, *wikis*, etc.)

Continue the discussion and talk to colleagues about *Best Practices/Replication*


CLARIN
OPEN ACCESS
Project Design & Guidelines
COS CENTER FOR OPEN SCIENCE
GitLab
Data Management & Collection
GitHub
Dropbox
owncloud
Data Processing & Analytics
Praat
R
Reporting
Studio
jupyter
LATEX
M↓
Archiving & Sharing
GitHub
GitLab
Open Notebooks
Open Data
Open Peer Review
Open Access
Open Source
Scientific social networks
Citizen Science
Open educational resources

Please contact me if you are interested in setting up a network of researchers who are interested in pursuing this further!

Live long and replicate!

Thank you!

Acknowledgements

I would like to thank. . .

Amanda Miotto
for sharing materials about Replicability in Research
(https://github.com/amandamiotto/Reproducible-Research-Things)

my colleagues at UQ

for comments and their feedback on earlier versions of this talk

Aschwanden, C. (2018). Psychology's replication crisis has made the field better. https://fivethirtyeight.com/features/psychologys-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. https://nobaproject.com/modules/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. https://www.sciencealert.com/replication-results-reproducibility-crisis-science-nature-journals.

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.

Ruhi, Ş., M. Haugh, and T. Schmidt (2014). *Best practices for spoken corpora in linguistic research*. Cambridge: Cambridge Scholars Publishing.

Velasco, E. (2019). Researcher discusses the the science replication crisis. https://phys.org/news/2018-11-discusses-science-replication-crisis.html.

Weir, K. (2015). A reproducibility crisis? the headlines were hard to miss: Psychology, they proclaimed, is in crisis. *Monitor on Psychology 46*, 39.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data 3*. https://www.nature.com/articles/sdata201618.

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. https://www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/.

Best Practices in Corpus Linguistics

What lessons should we take from the Replication Crisis and how can we guarantee high quality in our research?

Dr. Martin Schweinberger (m.schweinberger@uq.edu.au)

(see also www.martinschweinberger.de)

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

CREATE CHANGE