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) available under CC license (see also www.martinschweinberger.de)
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
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.
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 Reiser | @B_reiser | brislawx.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 C. Diener and Robert Biswas-Diener
University of Utah, University of Virginia, Portland State University
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 Reinsick | @B_reinsick | brianthepwox.com | Aug 27, 2018, 13:00am EDT

(Dis)trust in Science

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
Dr. Martin Schweinberger (m.schweinberger@uq.edu.au), available under CC license (See also www.martinschweinberger.de)
Researcher discusses 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 Reinsick | @b_reinsick | briang@vox.com | Aug 27, 2018, 11:00am EDT

(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.

arsky on July 5, 2018

By Edward Olemić and Robert Biswas-Diener
University of Utah, University of Virginia, Portland State University
Researcher discusses 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.

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.
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).
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).
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

Dr. Martin Schweinberger (m.schweinberger@uq.edu.au), available under CC license (See also www.martinschweinberger.de)
Suggestions to make our research more replicable
For individual researchers and teams

- **FAIR principles**
  data should be *Findable, Accessible, Interoperable, Reusable* (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
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*
Best Practices in Corpus Linguistics

Dr. Martin Schweinberger (m.schweinberger@uq.edu.au), available under CC license (See also www.martinschweinberger.de)
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


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.


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) available under CC license (see also www.martinschweinberger.de)