Contact
Martin Schweinberger
The University of Queensland, Australia
School of Languagesa dn Cultures
Gordon Greenwood Building (32-503A)
St Lucia Qld 4072 Australia
E: m.schweinberger@uq.edu.au
H: www.martinschweinberger.de

# Best Practices in Corpus Linguistics

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

## Aims, definition, and the current state of affairs

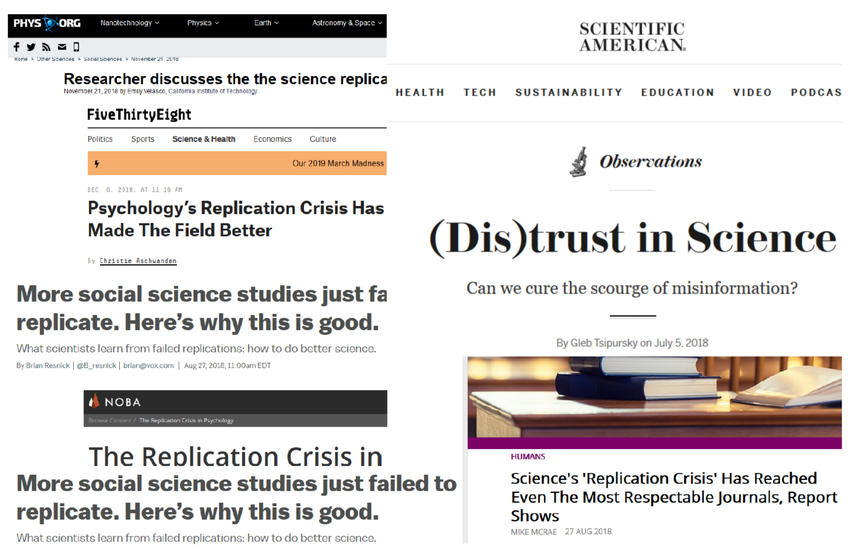
This presentation aims to
- Raise awareness for „Best Practices“ in (Corpus) Linguistics
- Start a discussion about issues related to Best Practices and Replicability
- Propose improvements to current research practices
- Offer solutions on how best practices can be implemented

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.

Best practices have come into focus as a result of the **Replication Crisis** (RC) which is a ongoing methodological crisis primarily affecting parts of the social and life sciences beginning in the early 2010s.

Nature 2016 poll of 1,500 scientists:
- 70% had failed to reproduce at least one other scientists experiment
- 50% had failed to reproduce one of their own experiments (Fanelli 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




Examples for media outlets reporting on the Replication Crisis.

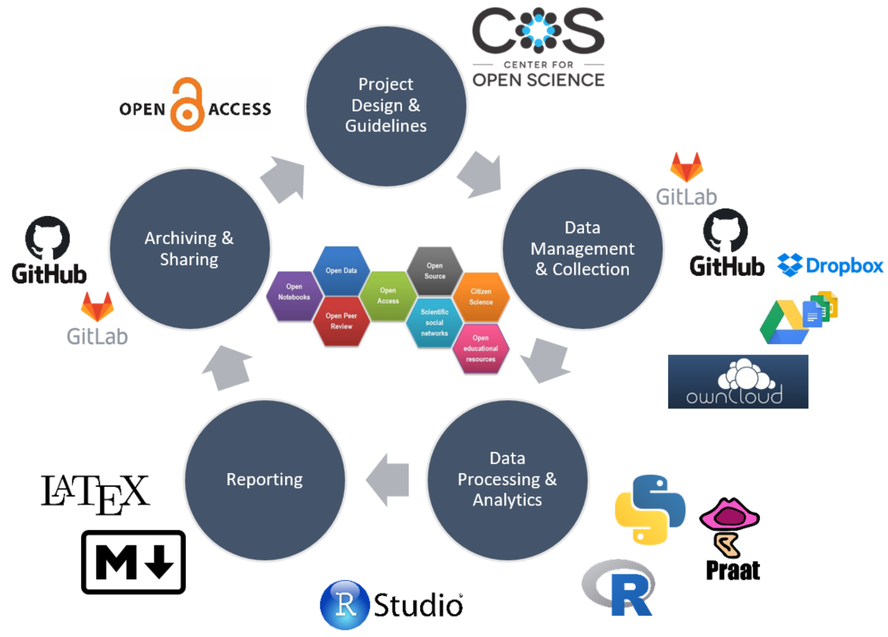
As a consequence of the RC, there is growing awareness. . .
- of a problem: currently most research is difficult to replicate/reproduce!
- that reproducibility is an essential part of the scientific method
- that the inability to replicate has potentially grave consequences as significant theories are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

## Best Practices and Replication in 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).

However, we could be better because *our analyses are not reproducible*, we have an *over-reliance on tools*, and *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

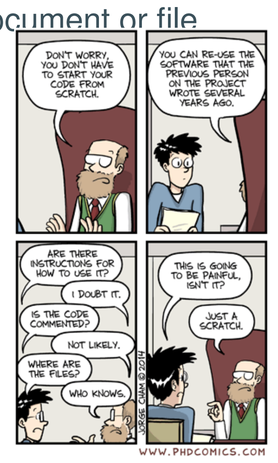



Research circle with software options to make your research more reproducible.

## Suggestions to make our research more replicable

### For individual researchers and teams

- **FAIR principles**: share and make your data ***F****indable*, ***A****ccessible*, ***I****nteroperable*, ***R****eusable* (FAIR) (Wilkinson et al. 2016). Also assign a *Digital Object Identifier* (DOI) to your data and provide a clear example for how your data should be cited (this way your data is a proper publication)
- **Archiving**: use online repositories (e.g. *GitHub*, *GitLab*, *CouldStore*, *MyDrive*, *Dropbox*) to avoid data loss and various versions of a single document or file
- **Scripts over tools**: use R rather than ready-made software tools because such apps are black-boxes that hinder replication and transparency (due to limited accessibility and/or time-consuming replication)
- **Documentation**: write down what you do and where you store all relevant elements of your project
- **Folder templates**: think about a schematic folder structure and use it for all your projects, e.g. always using subfolders for *data*, *tables*, and *images* for research projects or *slides*, *exercises*, *student materials*, and *assignments* for courses (ideally implement a policy in your team so that all team members use the same folder template)
- **Notebooks and version control**: make your research fully transparent and reproducible by doing your analyses in R Notebooks and sharing entire projects on GitHub




### For the community

- Endorse *Open Science*
  Open Data + Open Access + Open Methodology + Open Educational Resources)
- Only accept papers that have made data (and scripts) available
- Require data to be cited appropriately (this serves as both a reward and an incentive to publish corpora)
- Promote replication and support replication studies to be published
- Invest in and support training for staff and students in data management and other options that help make research more transparent (*R*, *Git*, *Markdown*, *wikis*, etc.)
- Continue the discussion and talk to your colleagues about *Best Practices* and *Replication*

## References

Aschwanden, C. (2018). Psychologys 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. heres 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. 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). Psychologys 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/.