The Replication Crisis

And what to do about it

Daniela Palleschi

2024-04-23

Table of contents

Learning Objectives

Today we will learn about…

- the replication crisis
- replication in language sciences
- requirements for replication

Resources

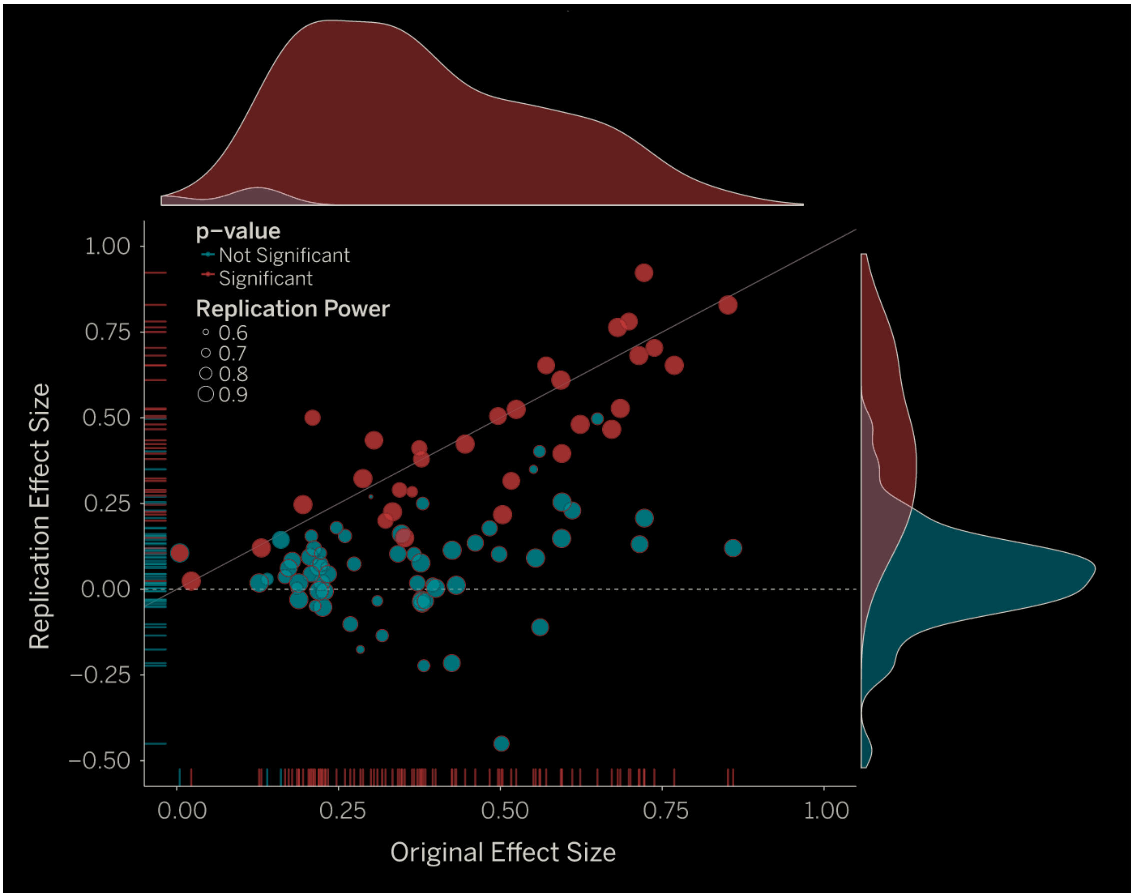
- this lecture covers Sönning & Werner (2021)
- introduction article of a special issue of the Journal *Linguistics*
  - *The replication crisis: Impications for linguistics*
- contains several articles on the topic, some of which we'll read later

Replication crisis

- data-based claims turned out to be less reliable than previously believed
  - statistical claims could not be replicated with new data
- large-scale replications brought attention to the issue
  - e.g., Nieuwland et al. (2018); Open Science Collaboration (2015)
- this has also led to distrust in findings in academia and the public

Most Published Research Findings are False

- the issue became more widespread with Ioannidis (2005)
  - defined bias in terms of design, analysis, and presentation factors
  - focussed on issues with *p*-values and statistical power
- Open Science Collaboration (2015) ran replications of 100 studies
  - 36% of replications found significant effects
  - 47% of original effects fell within 95% CIs of replication effect
- in essence: fewer significant findings and smaller effects in replications
- how is can this be?



Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Figure 1: Source: Open Science Collaboration (2015) (all rights reserved)

The problem with p-values

- issues relating to reported findings:
  - misuse of misinterpretation of p-values in Null Hypothesis Significant Testing (NHST; e.g., Ioannidis, 2005)
  - study design
  - improper use of statistical methods
    - stemming from inadequate teaching
  - selective reporting
- in other words, HARKing and p-hacking (whether consciously done or not)

Solving the problem

- these could be mitigated with Open Science practices
  - *transparency* in writing, analyses, planning/hypothesising stages
  - *reproducibility* of analyses
  - greater value given to *replication* studies
  - embracing and addressing *uncertainty* (Vasishth & Gelman, 2021)
- in sum: "conscientious practice" (Sönning & Werner, 2021, p. 1182)

The garden of forking paths

- or 'researcher degrees of freedom' (Simmons et al., 2011)
- the problem: there are many plausible ways to analyse any given data set
- there are many choices researchers make in:
  - experimental design
  - data collection
  - data preprocessing
  - data analyses
  - reporting
- the path we happen to go down can seem pre-determined (Gelman & Loken, 2014)
  - but can amount to HARKing, *p*-hacking, fishing
- the fastest solution: share everything and write transparently

The current state of quantitative linguistics

- there is a trend towards empirical methods throughout linguistics
    - we should pay attention to methodological discussions in related fields
- we also find ourselves in a state of methdological crisis

Kuhn's structure of scientific revoluations

- Thomas Kuhn's *The Theory of Scientifc Revoluations* (1962)
    - based on socio-historical observation
    - the evolution of scientificy theory is cycical
    - crisis leads to revolution
- also applies to research methodology

Three recurrent phases:

1. **normal science**
    - little controversy over theoretical underpinnings
    - researchers work on small problems within a theory
2. **crisis**
    - contradictions between theory and evidence
    - questioning of conventionally accepted theory
3. **revolution**
    - overthrowing of previous norms in favour of a new paradigm
    - leads to new normal science

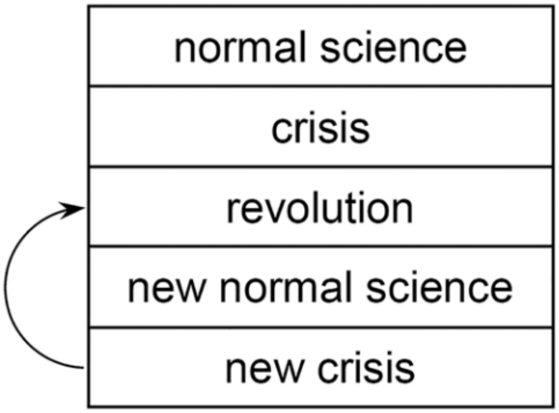


Figure 2: Download GitHub repositiory

Previous cycles of statistical analyses

- proprietary, point-and-click software (e.g., SPSS)
  - move to open source programming languages (e.g., R, Python, Julia)
- ANOVAs
  - move to linear regression
  - then linear mixed models
    - random-intercepts only models
    - maximal models
    - parsimonious models
- now a trend towards Bayesian regression

What do statisticians say?

- Wasserstein et al. (2019): list of Do's and Don't from statisticians
  - *Don't* base conclusions on p-values
  - *Do* think about ATOM: Accept uncertainty, be Thoughtful, Open, and Modest
- Wasserstein & Lazar (2016): the American Statistical Association's statement on p-values
  - p-values are often misused and misinterpreted
  - good statistical practice is part of good scientific practice
    - as such, relies on good study design and conduct
    - interpretation in context
    - complete and transparent reporting

Overwhelmed?

- we're in a state of crisis with a wealth of possible statistical paths
  - but no current "gold standard"
- this can lead to anxiety among researchers
  - which analysis should I run? am I doing it correctly?
- just keep in mind ATOM
  - strive for honesty, not perfection

Revolution

- methodological anxiety stems from shifting sands, but leads to revolution
- revolution usually comes from young newcomers
  - resistance to change usually comes those with more invested in the prior ways
- the good news: the revolution is underway
  - leads to an increase in resources and courses on e.g., multi-level models
- one suggested reform: Open Science!

The old vs. the new

- changes refer to not only statistical analyses
  - but also emphasise transparency
- ideally, we (as a field) would up our analysis game
  - but a good first step is moving towards Open Science
  - share data, code
  - transparently map out your route in the ‘garden of forking paths’
- these are steps we’ll cover in this course
  - pre-registration
  - data and code sharing
  - reproducible workflow
  - transparent writing

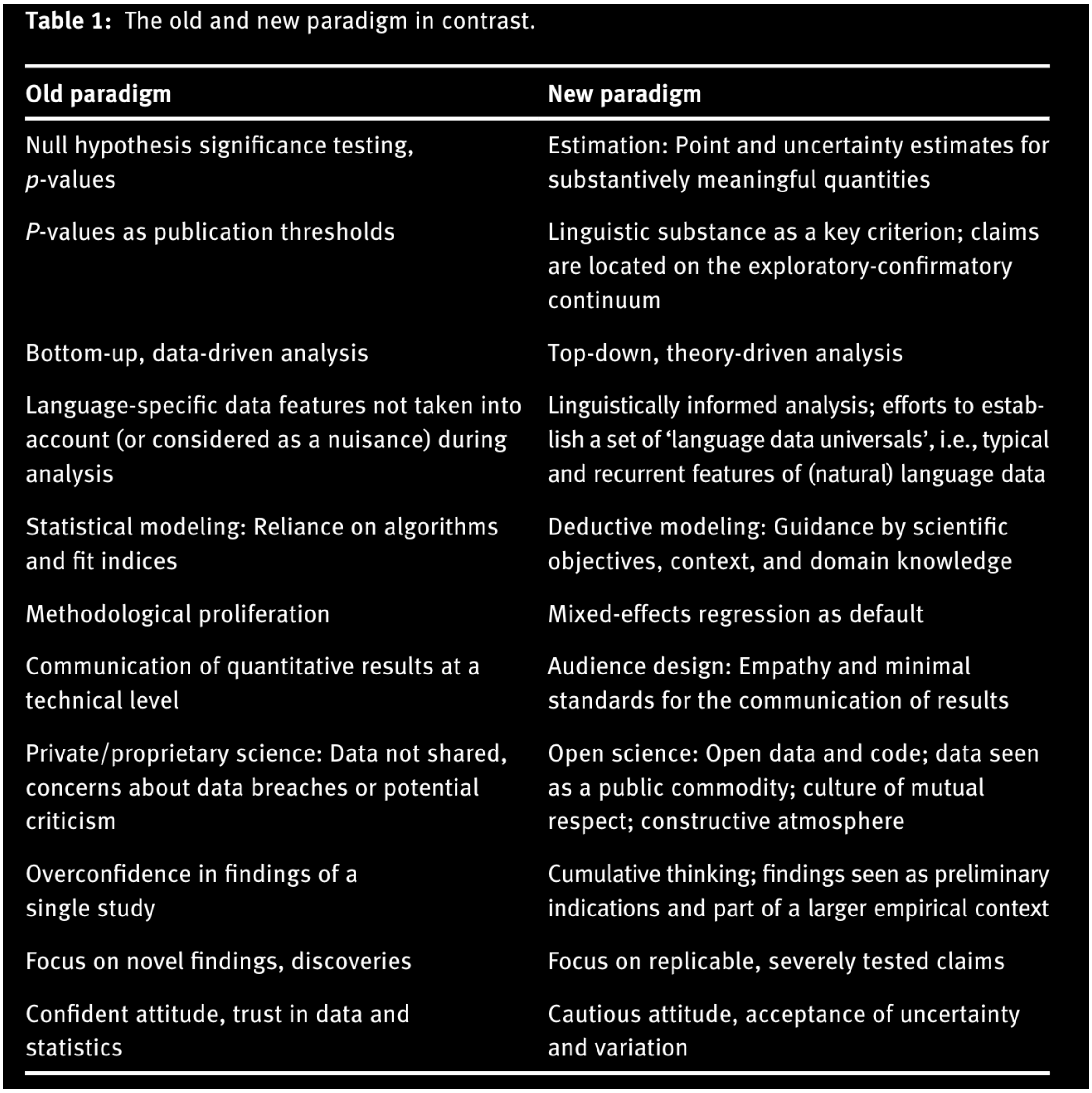


Table 1: The old and new paradigm in contrast.

Old paradigm	**New paradigm**
Null hypothesis significance testing, *p*-values	Estimation: Point and uncertainty estimates for substantively meaningful quantities
P-values as publication thresholds	Linguistic substance as a key criterion; claims are located on the exploratory-confirmatory continuum
Bottom-up, data-driven analysis	Top-down, theory-driven analysis
Language-specific data features not taken into account (or considered as a nuisance) during analysis	Linguistically informed analysis; efforts to establish a set of 'language data universals', i.e., typical and recurrent features of (natural) language data
Statistical modeling: Reliance on algorithms and fit indices	Deductive modeling: Guidance by scientific objectives, context, and domain knowledge
Methodological proliferation	Mixed-effects regression as default
Communication of quantitative results at a technical level	Audience design: Empathy and minimal standards for the communication of results
Private/proprietary science: Data not shared, concerns about data breaches or potential criticism	Open science: Open data and code; data seen as a public commodity; culture of mutual respect; constructive atmosphere
Overconfidence in findings of a single study	Cumulative thinking; findings seen as preliminary indications and part of a larger empirical context
Focus on novel findings, discoveries	Focus on replicable, severely tested claims
Confident attitude, trust in data and statistics	Cautious attitude, acceptance of uncertainty and variation

Figure 3: Source: Sönning & Werner (2021)

Simple fixes

- planning and design
  - large sample sizes
  - establish pre-processing/analysis steps a priori
- methodologically
  - select variables based on theory and research questions
  - model non-independence of data points (mixed models)
  - move towards estimation and away from arbitrary significance thresholds
- writing
  - be transparent about choices made

Words of comfort

> Less experienced scholars must not fear methodological attacks on their analyses, which are instead seen as informing interim interpretations that may require future modification.

— Sönning & Werner (2021), p. 1199

- in some sub-fields linear mixed models are still not considered the standard
  - so you're well situated despite the doom around *p*-values
- moving from frequentist (NHST) framework to the Bayesian framework is relatively painless
  - in this class we will run a LMM with `lme4` (Frequentist) and with `brms` (Bayesian)

Running replications: what to replicate

- what makes a study 'worth' replicating?
  - suggestions from Isager (2020):

  1. value/interest of the topic
  2. uncertainty about the claim
  3. quality of proposed replication
     - or ability to reduce uncertainty
  4. costs and feasibility

- what makes a replication study ‘worthy’ for publication?
  - theoretical impact of the replicated finding
  - statistical power of the replication

Replication value

- replication value (RV): “the expected utility of a finding before replication” (Isager, 2020, p. 6)
  - (scientific) value of the research claim
    - importance to the field, to policy, health etc.
  - the uncertainty of our knowledge about the claim
    - validity of study design, statistical power, bias, etc.
- replication aims to reduce uncertainty
  - which also increases utility of the claim

Quantifying RV

- how to quantify *value* and *uncertainty*?
- Isager et al. (2021) suggest using…
  - average yearly citation count to estimate *value*
    - the more citations, the higher the impact the original study had
  - and sample size to estimate *uncertainty* ($\frac{1}{\sqrt{n}}$)
    - the higher the sample size, the more precise the estimate
    - the lower the sample size, the greater the uncertainty
    - i.e., n is inversely correlated with *uncertainty*

RV_{Cn}

Isager et al. (2021):

$$RV_{Cn} = value\ x\ uncertainty$$

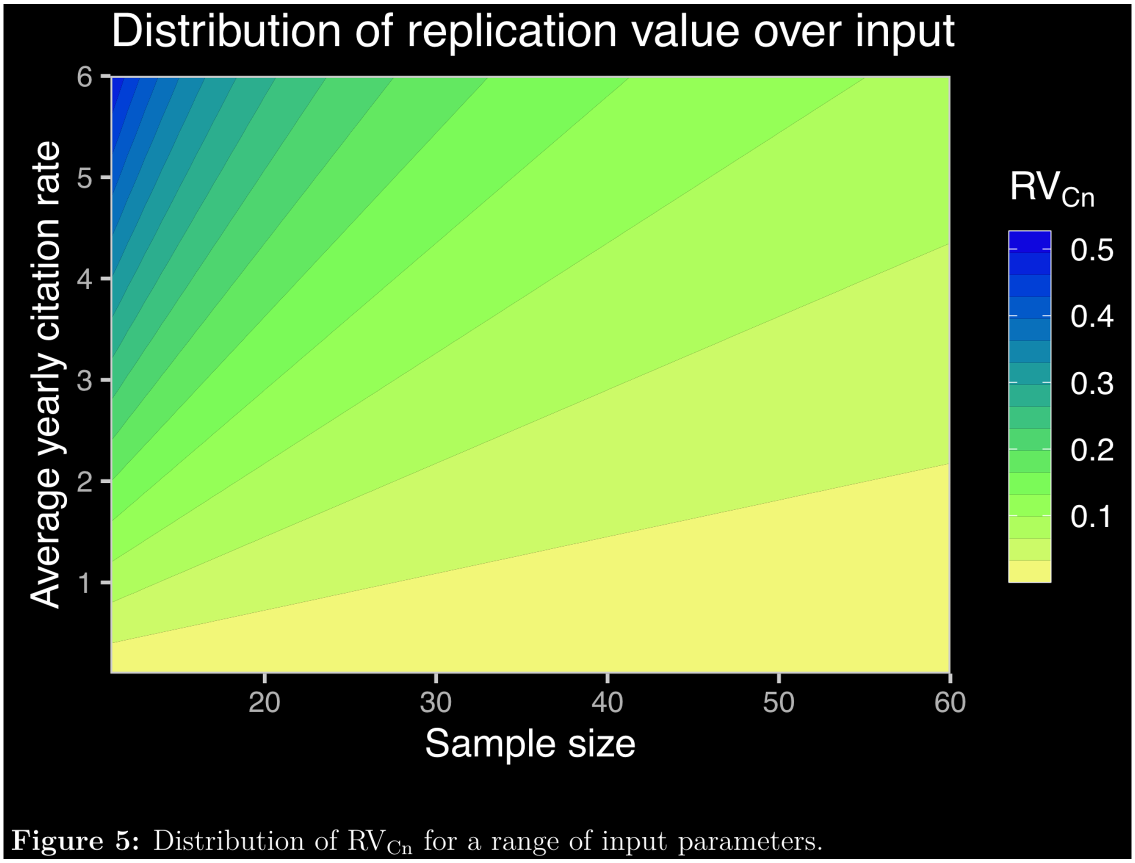


Figure 5: Distribution of $\mathrm{RV}_{\mathrm{Cn}}$ for a range of input parameters.

Figure 4: Source: Isager et al. (2021), p. 25 (all rights reserved)

Student replications

> It is peculiar that undergraduate students can be taught about the perils of underpowered studies in formal statistical instruction and simultaneously be required to perform research that is almost inevitably underpowered

— Quintana (2021), p. 1117

- a possible solution: student thesis replications
  - hands-on experience in open science practices
  - e.g., cumulative replication studies run by multiple groups
- some resources for students interested in replications
  - Student Theses Replication Network Linguistics (STReNeL)
  - Collaborative REplications and Education Project (CREP)
  - Framework for Open and Reproducible Research Training (FORRT)
  - German Reproducibility Network (DERN)

Exercise

- Moodle: Quiz 'Kobrok & Roettger (2022)'
  - scan the article to answer the questions
  - this is not graded

Learning objectives

Today we learned…

- the replication crisis
- replication in language sciences
- requirements for replication

Important terms

References

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460–465.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), 2–8. https://doi.org/10.1371/journal.pmed.0020124

Isager, P. M. (2020). *Deciding what to replicate: A formal definition of "replication value" and a decision model for replication study selection.* MetaArXiv. https://doi.org/10.1037/met0000438

Isager, P. M., Van 'T Veer, A. E., & Lakens, D. (2021). *Replication value as a function of citation impact and sample size.* https://doi.org/10.31222/osf.io/knjea

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Quintana, D. S. (2021). Replication studies for undergraduate theses to improve science and education. *Nature Human Behaviour*, *5*(9), 1117–1118. https://doi.org/10.1038/s41562-021-01192-8

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Sönning, L., & Werner, V. (2021). The replication crisis, scientific revolutions, and linguistics. *Linguistics*, *59*(5), 1179–1206. https://doi.org/10.1515/ling-2019-0045

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, *59*(5), 1311–1342. https://doi.org/10.31234/osf.io/zcf8s

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond " $p <$ 0.05." *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913