Reproducible Data Science: How do you know if published scientific findings are replicable?

Adam Thomas
Data Science and Sharing Team, Intramural Research Program, NIMH

National Institute of Mental Health

NIMH IRP Data Science and Sharing Team

Adam Thomas
Team Lead

Dustin Moraczewski
Data Scientist

Arshitha Basavaraj
Data Engineer

Jessica Dafflon
Data Scientist

Eric Earl
Data Scientist

Anthony Galassi
Software Engineer

Carl Harris
Post-bac IRTA

http://cmn.nimh.nih.gov/dsst

Poll #1

- What percentage of peer-reviewed papers in the biomedical literature present results that can be replicated by another scientist in the same field?

1. > 80%
2. 50-80%
3. 20-50%
4. < 20%

Poll #1

- What percentage of peer-reviewed papers in the biomedical literature present results that can be replicated by another scientist in the same field?

1. > 80%
2. 50-80%
3. 20-50%
4. < 20%

Less than 50%

PLoS Medicine | www.plosmedicine.org

August 2005 | Volume 2 | Issue 8 | e124

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern
current published research fi
false. The probability that a re
is true may depend on study
bias, the number of other stu
same question, and, importan
of true to no relationships am
relationships probed in each
field. In this framework, a rese
is less likely to be true when t
conducted in a field are small
effect sizes are smaller; when
greater number and lesser pr
of tested relationships; where
greater flexibility in designs, d
outcomes, and analytical mod
there is greater financial and
interest and prejudice; and w
teams are involved in a scient
in chase of statistical significa
Simulations show that for mo
designs and settings, it is mor

Table 4. PPV of Research Findings for Various Combinations of Power (1 − β), Ratio of True to Not-True Relationships *(R)*, and Bias *(u)*

1 − β	*R*	*u*	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study.
RCT, randomized controlled trial.
DOI: 10.1371/journal.pmed.0020124.t004

Ioannidis on Reproducability Feb 2021, https://www.youtube.com/watch?v=L1-fNM0uJQA

Reproducibility of Cancer Studies

HEALTHCARE & PHARMA MARCH 28, 2012 / 1:10 PM / UPDATED 9 YEARS AGO

In cancer science, many "discoveries" don't hold up

During a decade as head of global cancer research at Amgen, C. Glenn Begley identified 53 “landmark” publications -- papers in top journals, from reputable labs -- for his team to reproduce. Begley sought to double-check the findings before trying to build on them for drug development.

Result: 47 of the 53 could not be replicated. He described his findings in a commentary piece published on Wednesday in the journal Nature.

Reproducibility in Psychology

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

SCIENCE sciencemag.org 28 AUGUST 2015 • VOL 349 ISSUE 6251

We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available.

Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results

https://www.theatlantic.com/science/archive/2015/08/psychology-studies-reliability-reproducability-nosek/402466/
https://science.sciencemag.org/content/349/6251/aac4716

Poll Question #2

- Papers published in prestigious journals are more likely to report reproducible findings.

1. True
2. False

Poll Question #2

- Papers published in prestigious journals are more likely to report reproducible findings.

1. True
2. False

Reproducibility of Cancer Studies

HEALTHCARE & PHARMA MARCH 28, 2012 / 1:10 PM / UPDATED 9 YEARS AGO

In cancer science, many "discoveries" don't hold up

During a decade as head of global cancer research at Amgen, C. Glenn Begley identified 53 "landmark" publications -- papers in top journals, from reputable labs -- for his team to reproduce. Begley sought to double-check the findings before trying to build on them for drug development.

Result: 47 of the 53 could not be replicated. He described his findings in a commentary piece published on Wednesday in the journal Nature.

REPRODUCIBILITY OF RESEARCH FINDINGS

Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.
*Source of citations: Google Scholar, May 2011.

Reproducibility of Cancer Studies

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

nature reviews drug discovery

Published: 31 August 2011

Surprisingly, even publications in prestigious journals or from several independent groups did not ensure reproducibility. Indeed, our analysis revealed that the reproducibility of published data did not significantly correlate with journal impact factors, the number of publications on the respective target or the number of independent groups that authored the publications.

Poll Question #3

Since Ioannidis's explosive 2005 paper, the field has become more vigilant, and the number of unreproducible findings published in the literature has decreased.

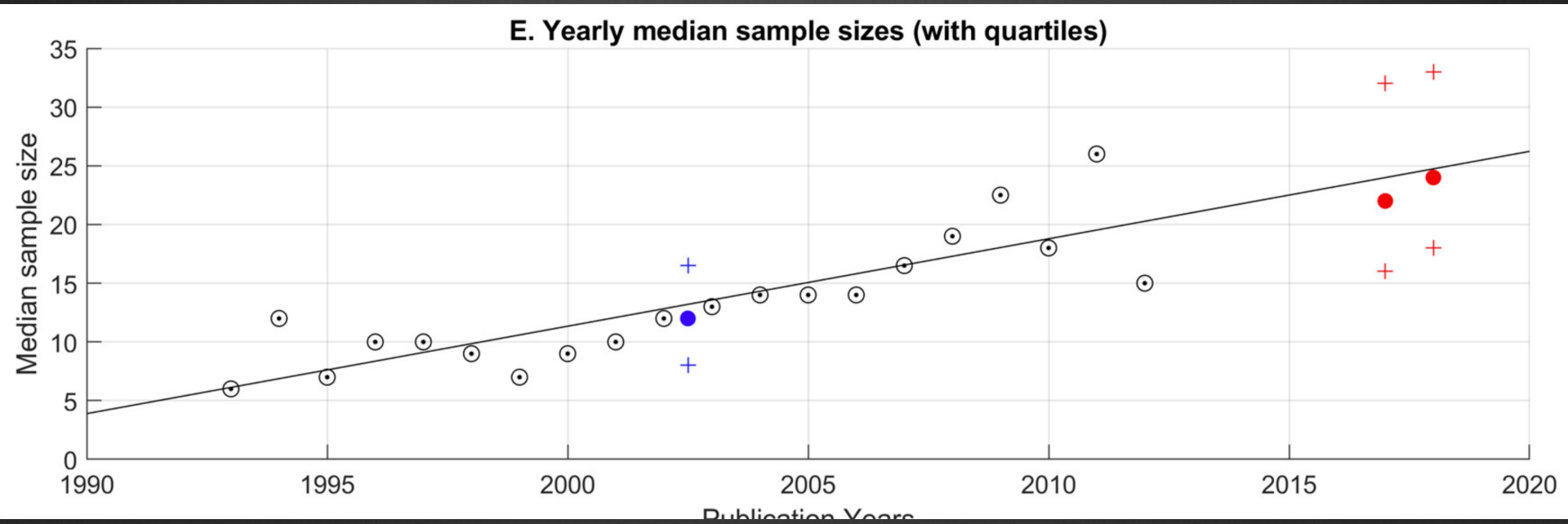
- True
- False

Is it getting better?

Maybe, slowly

Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals

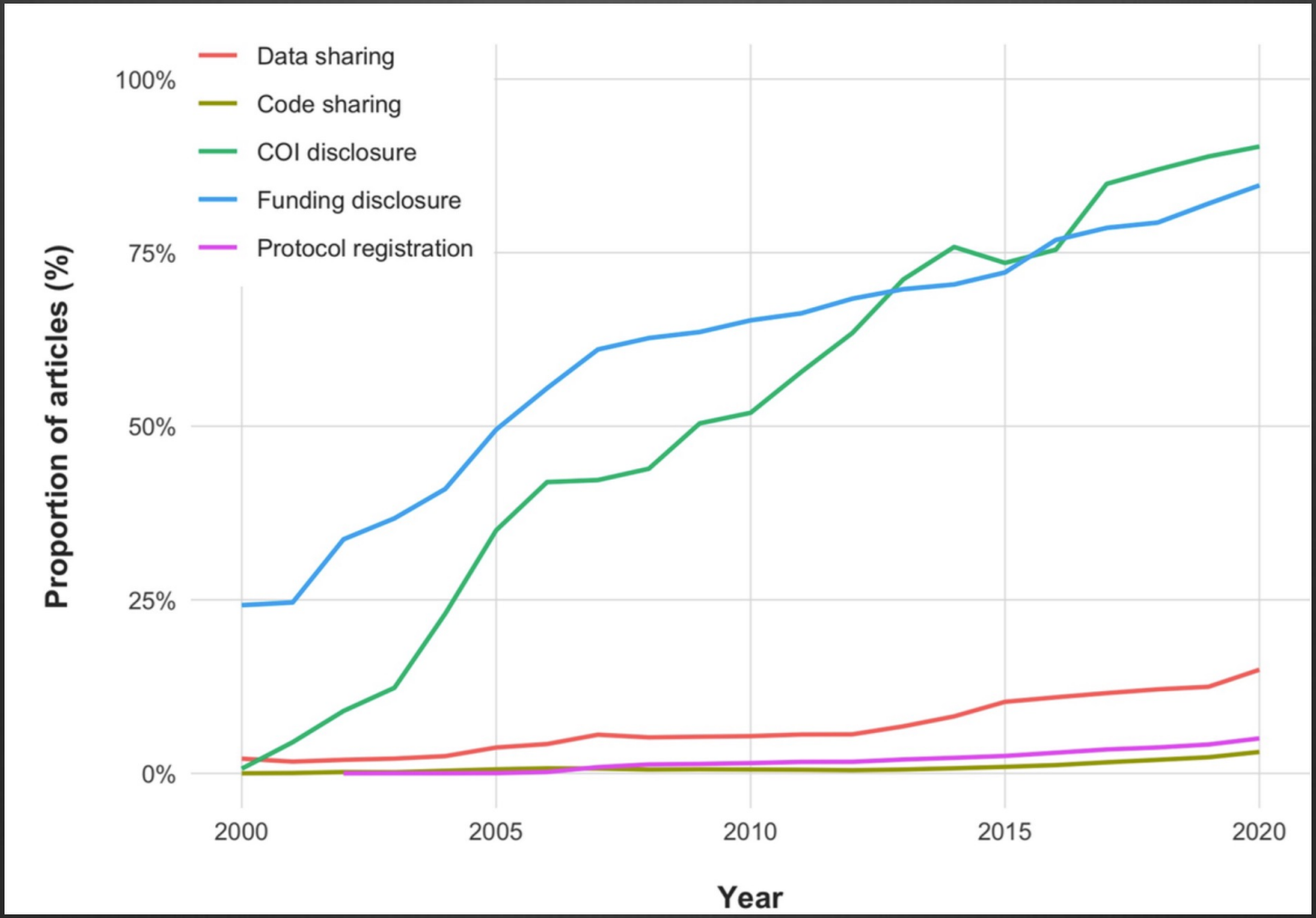
Denes Szucs [a,*], John PA. Ioannidis [b]



https://www.vox.com/future-perfect/21504366/science-replication-crisis-peer-review-statistics

Is it getting better?

Maybe, slowly



Serghiou et al. (2021) Assessment of transparency indicators across the biomedical literature: How open is open?. PLOS Biology
https://doi.org/10.1371/journal.pbio.3001107

Reproducibility in Mental Health

SCIENCE

The Atlantic

A Waste of 1,000 Research Papers

Decades of early research on the genetics of depression were built on nonexistent foundations. How did that happen?

By Ed Yong

MAY 17, 2019

https://www.theatlantic.com/science/archive/2019/05/waste-1000-studies/589684/
https://doi.org/10.1176/appi.ajp.2018.18070881

Reproducibility in Mental Health

- A 1996 study of 1,000 subjects claimed that a candidate gene, SLC6A4, was found to influence a person's risk of clinical depression.
- 17 other candidate genes were subsequently linked to depression risks, using similar methods and sample sizes
- Together, over 1,000 papers were published on these genes.
- A 2019 study using 10s to 100s of thousands of subjects showed none of the 18 genes were related to depression risk.
- "How on Earth could we have spent 20 years and hundreds of millions of dollars studying pure noise?" – Matthew Keller, lead author

https://www.theatlantic.com/science/archive/2019/05/waste-1000-studies/589684/
https://doi.org/10.1176/appi.ajp.2018.18070881

Poll Question #4

Scientists are pretty good at predicting (guessing) which findings will be reproducible in published papers.

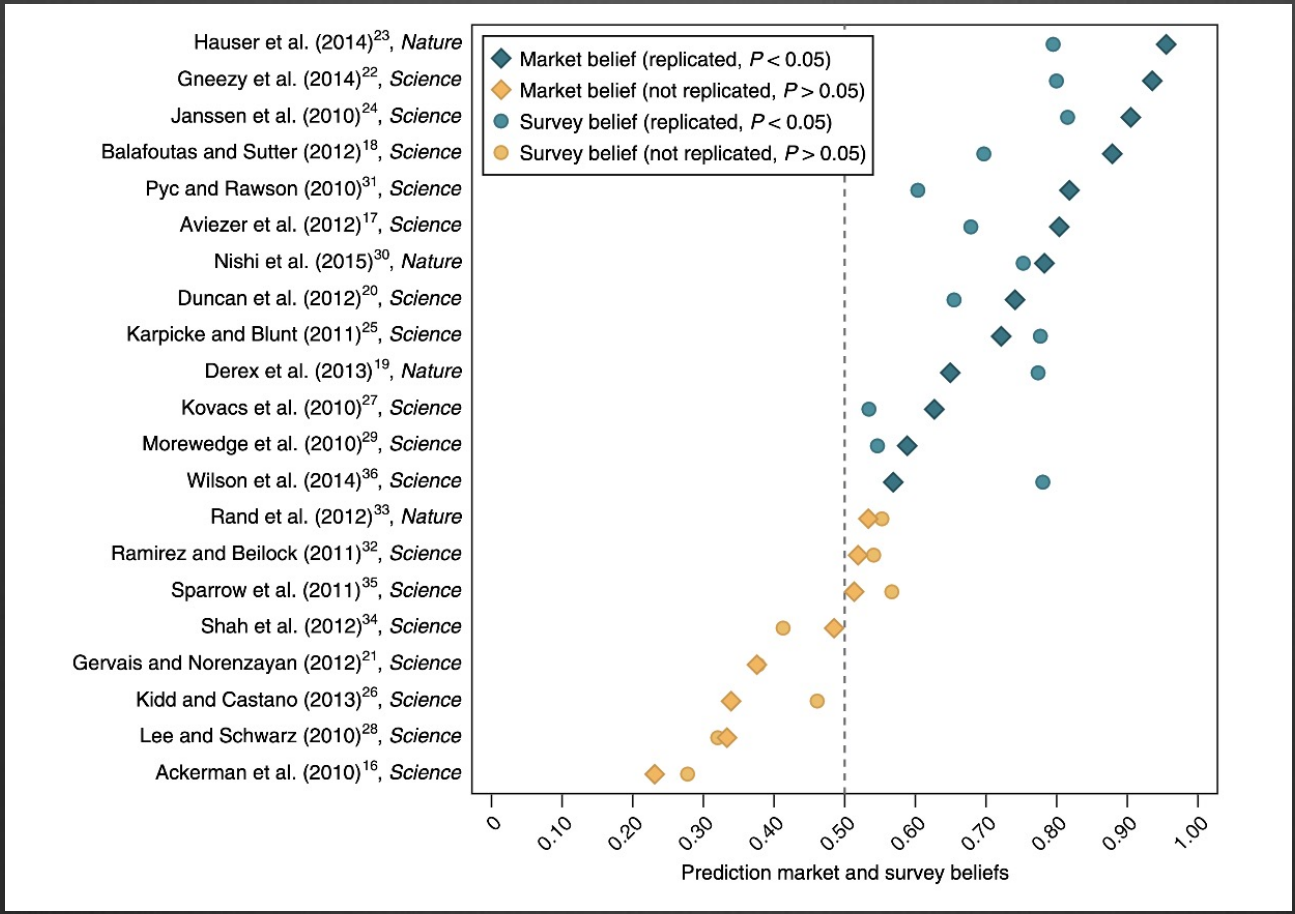
- True
- False

Poll Question #4

Scientists are pretty good at predicting (guessing) which findings will be reproducible in published papers.

- True
- False

Unreproducible results are predictable



https://doi.org/10.1038/s41562-018-0399-z

Poll Question #5

Which factor is the most predictive of findings that will be reproducible?

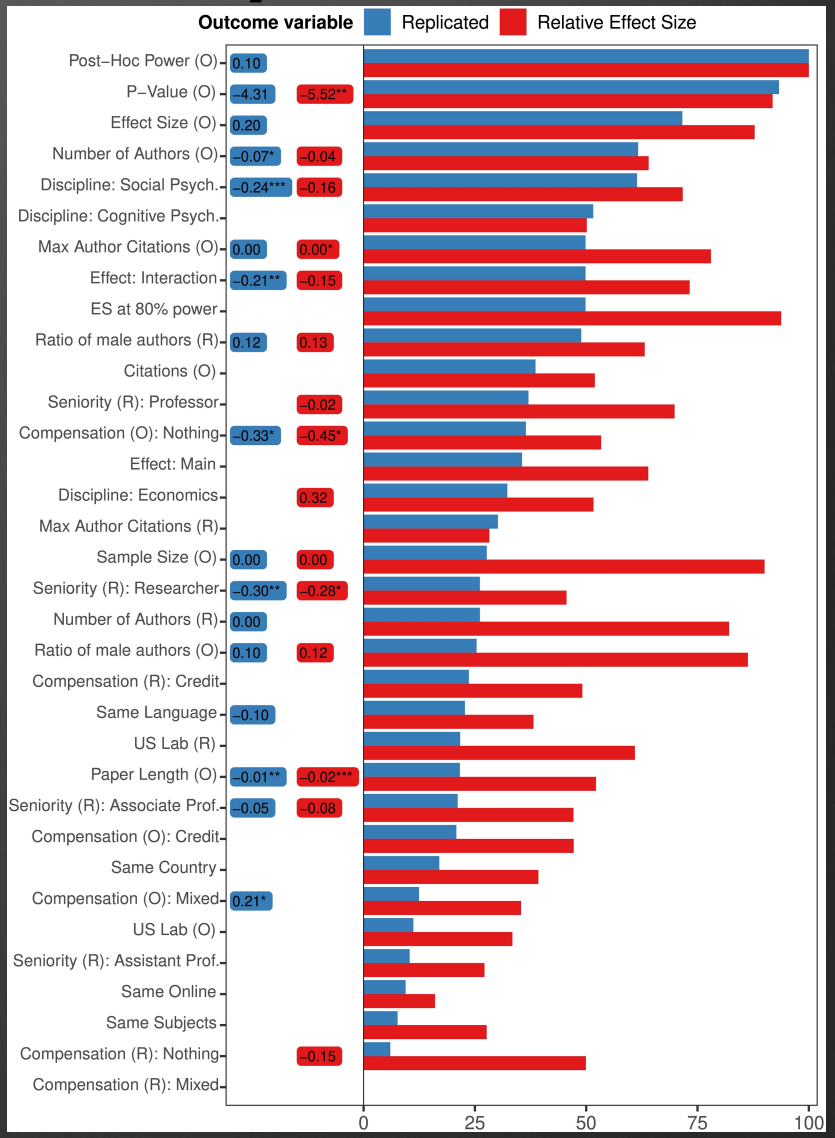
1. Large effect size
2. Published in prestigious journal
3. Large sample size
4. Well respected lab/institution
5. Transparency/availability of data/code/methods

What factors are predictive of reproducibility?

RESEARCH ARTICLE

Predicting the replicability of social science lab experiments

Adam Altmejd[1,2]*, Anna Dreber[1,3], Eskil Forsell[1], Juergen Huber[3], Taisuke Imai[4], Magnus Johannesson[1], Michael Kirchler[3], Gideon Nave[5], Colin Camerer[6]

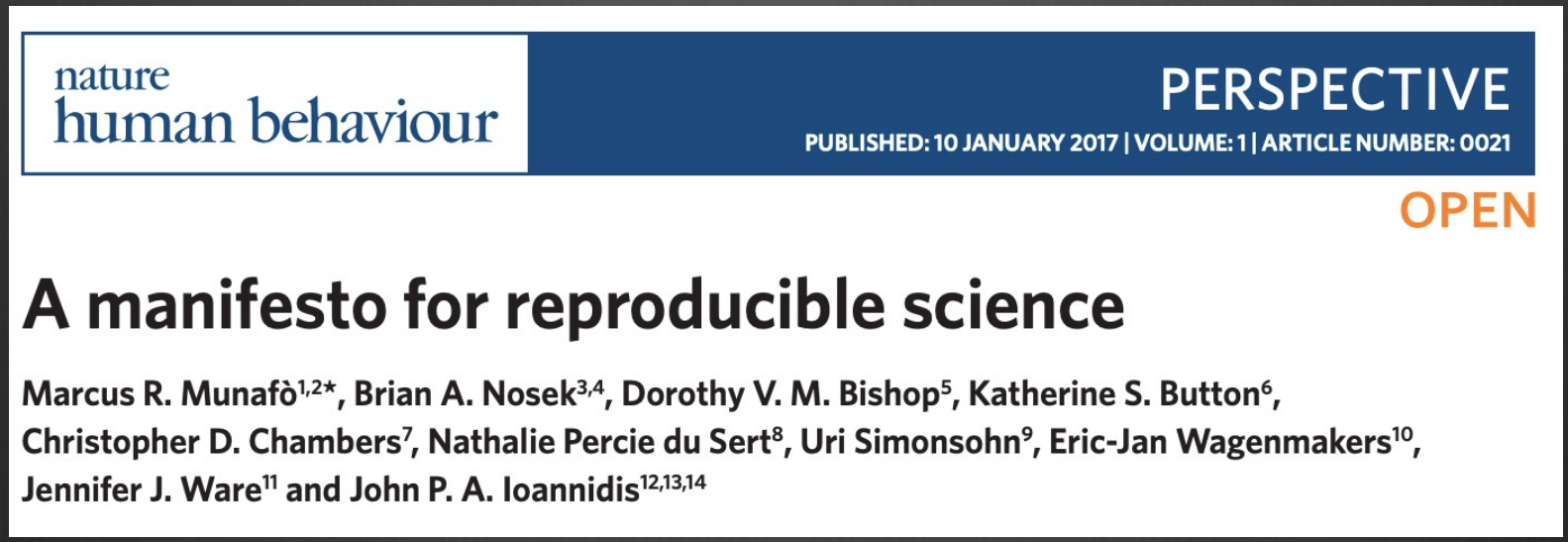


https://doi.org/10.1371/journal.pone.0225826

What features correlate with unreproducible findings?

1. Large effect size
2. ~~Published in prestigious journal~~
3. Large sample size
4. Well respected lab/institution
5. Transparency/availability of data/code/methods

What features correlate with unreproducible findings?



nature human behaviour

PERSPECTIVE

PUBLISHED: 10 JANUARY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0021

OPEN

A manifesto for reproducible science

Marcus R. Munafò[1,2*], Brian A. Nosek[3,4], Dorothy V. M. Bishop[5], Katherine S. Button[6], Christopher D. Chambers[7], Nathalie Percie du Sert[8], Uri Simonsohn[9], Eric-Jan Wagenmakers[10], Jennifer J. Ware[11] and John P. A. Ioannidis[12,13,14]

- Transparency/availability of data/code/methods
- Well respected lab/institution?
  - Isolated and siloed researchers are more likely to produce irreproducible results (many eyeballs)
  - However, high ranking labs may be blindly trusted and subject to less scrutiny.
  - Groupthink is common in science. Creative, out-of-the-box ideas can come from labs outside the establishment.

Summary so far (Part 1)

- The number of unreproducible findings in the biomedical literatures remains alarmingly high
- Some irreproducibility is just part of science, but the majority is preventable
- There has been some progress in changing norms and behavior, but it is slow
- Questions before Part 2?

Part 2: If all biomedical science has been mired in a reproducibility crisis for nearly 20 years, is there anything a data science minded summer intern can do about it?

Yes!

In fact, trainees are the key to fixing this mess!

How can you improve reproducibility?

- Pre-registration
- Planning for data sharing
- Version control & code sharing (git/github)
- Dealing with errors (Unit tests / continuous integration)
- Making your code reusable (Containers)
- Reporting and dissemination (Pre-prints & checklist)

Pre-registration

PNAS

The preregistration revolution

Brian A. Nosek[a,b,1], Charles R. Ebersole[b], Alexander C. DeHaven[a], and David T. Mellor[a]

[a]Center for Open Science, Charlottesville, VA 22903; and [b]Department of Psychology, University of Virginia, Charlottesville, VA 22904

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved August 28, 2017 (received for review June 15, 2017)

- Pre-registration is the simple process of making a record of experimental predictions
- It does not prohibit or prevent exploratory research
- It does help the investigator and the reader distinguish between predictions and "postdiction" (HARKing)
- Pre-registration also combat publication bias

https://doi.org/10.1073/pnas.1708274114

Planning ahead for Data Sharing

- Human subjects: Ensure your protocol has acceptable data sharing language:
  - https://open-brain-consent.readthedocs.io
- When designing, collecting, and analyzing consult with standards documents:
  - Enhancing Quality and Transparency of Health Research (EQUATOR) http://www.equator-network.org
  - Best Practices in Data Analysis and Sharing in Neuroimaging using MRI (COBIDAS) http://dx.doi.org/10.1101/054262
- Think about which repository might be best for sharing your data. What structure/format will it require?

https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

Version Control

Version control systems allows you to:

- Store all your analysis in a central repository
- Keep a history of "snapshots" of your evolving analysis
- Quickly switch between different versions of your analysis
- Adopt and modify code from other scientists
- Collaborate

Poll Question #6

Studies have shown that computer code written by scientists has an average of:

1. less than 1 defect per 1000 lines of code
2. 1-10 defects per 1000 lines of code
3. 10-20 defect per 1000 lines of code
4. Greater than 20 defects per lines of code

Poll Question #6

Studies have shown that computer code written by scientists has an average of:

1. ~~less than 1 defect per 1000 lines of code~~
2. ~~1-10 defects per 1000 lines of code~~
3. ~~10-20 defect per 1000 lines of code~~
4. Greater than 20 defects per lines of code

Coding Errors are a fact of life

- One 2003 study looked at coding errors from professional software developers make an average of 1 to 7.5 defects per lines of code.
- Other estimates put the number at 10 to 50 defects per 1000 lines of code
- It's not a question of whether your code has errors (it does), it's a question of when you find them and how they impact your results

https://reproducibility.stanford.edu/coding-error-postmortem/

Coding Errors are a fact of life

- One 2003 study looked at coding errors from professional software developers make an average of 1 to 7.5 defects per lines of code.
- Other estimates put the number at 10 to 50 defects per 1000 lines of code
- It's not a question of whether your code has errors (it does), it's a question of when you find them and how they impact your results

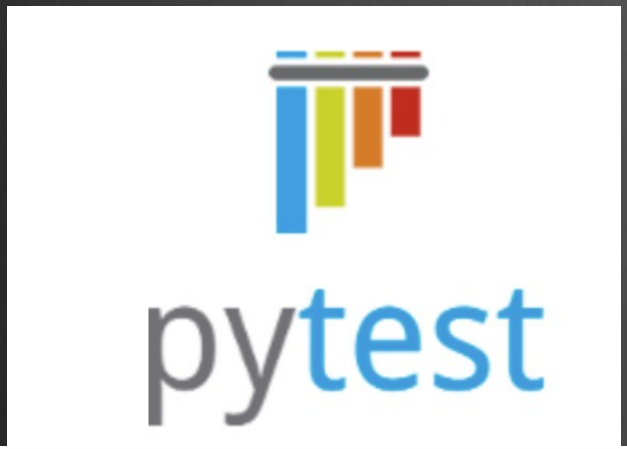
https://reproducibility.stanford.edu/coding-error-postmortem/

Typical approach to code testing in science

- Write some code
- Run code on some of your data
- Debug code until the output looks reasonable
- Repeat!

- "The first principle is that you must not fool yourself and you are the easiest person to fool." –Richard Feynman, 1974

Testing your code



pytest: helps you write better programs

The `pytest` framework makes it easy to write small tests, yet scales to support complex functional testing for applications and libraries.

An example of a simple test:

```
# content of test_sample.py
def inc(x):
    return x + 1


def test_answer():
    assert inc(3) == 5
```

```
$ pytest
============================= test session starts ==============================
test_sample.py F                                                        [100%]

================================= FAILURES =================================
_____________________________ test_answer _____________________________

    def test_answer():
>       assert inc(3) == 5
E       assert 4 == 5
E        +  where 4 = inc(3)

test_sample.py:6: AssertionError
========================== short test summary info ==========================
FAILED test_sample.py::test_answer - assert 4 == 5
============================== 1 failed in 0.12s ===============================
```

Testing your code

- A basic pytest tutorial for data science
  https://github.com/poldrack/pytest_tutorial
- Ask github to run your tests every time you commit (continuous integration)
- Automated assessment how much of your code has been tested (`pip install coverage`)
- Unit tests in R
  https://towardsdatascience.com/unit-testing-in-r-68ab9cc8d211

Dealing with errors

- Even in the best labs, sometimes errors slip through


SCIENTIFIC PUBLISHING

A Scientist's Nightmare: Software Problem Leads to Five Retractions

22 DECEMBER 2006 VOL 314 **SCIENCE**

In 2007, Chang and coauthors retracted five previously published papers describing the structures of three multidrug transporter proteins after another research group published a widely differing structure, which led to the discovery of a critical bug in the Chang group's custom software tools.


Coding error postmortem

August 10, 2020

By Russ Poldrack, McKenzie Hagen, and Patrick Bissett

"We had posted a preprint describing some issues that we had identified with the stop-signal task in the ABCD Study, along with the code used for all of the analyses. The ABCD stop-signal team performed a detailed review our code and notified us of an error in the code that resulted in inaccurate estimation of one of the basic behavioral measures on the task"

Dealing with errors

- Even in the best labs, sometimes errors slip through


SCIENTIFIC PUBLISHING

A Scientist's Nightmare: Software Problem Leads to Five Retractions

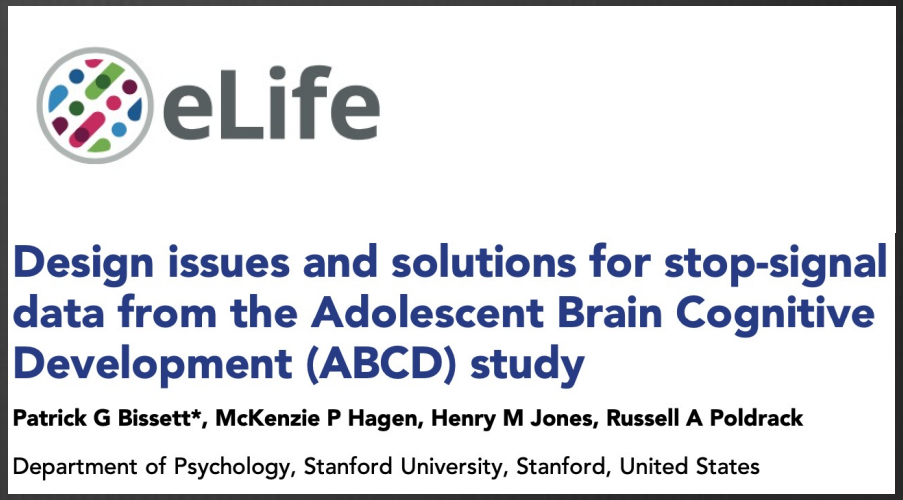
22 DECEMBER 2006 VOL 314 **SCIENCE**


Coding error postmortem

August 10, 2020

By Russ Poldrack, McKenzie Hagen, and Patrick Bissett

Since then, Chang has published several papers in the field of structural biology, and has been awarded a EUREKA grant, "for exceptionally innovative research projects that could have an extraordinarily significant impact on many areas of science," from the NIH and is a full professor at USCD.


eLife

Design issues and solutions for stop-signal data from the Adolescent Brain Cognitive Development (ABCD) study

Patrick G Bissett*, McKenzie P Hagen, Henry M Jones, Russell A Poldrack

Department of Psychology, Stanford University, Stanford, United States

Dealing with errors

- Even in the best labs, sometimes errors slip through
- Sharing code is critical for discovering error. Don't let embarrassment about ugly code stop you from sharing.
- Own your mistakes and do what's necessary to correct them. Good PIs will applaud you for it.
- Making errors and correcting them is part of science.
- Normalizing errors is a critical step in addressing the reproducibility crisis.
- Never mock or jeer another person's code, no matter how sloppy.

Reproducible workflows with Containers

Tutorial

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and Practices in Psychological Science
April-June 2021, Vol. 4, No. 2, pp. 1–18

DOI: 10.1177/25152459211017853
www.psychologicalscience.org/AMPPS
SAGE

Leveraging Containers for Reproducible Psychological Research

Kristina Wiebels and David Moreau
School of Psychology and Centre for Brain Research, University of Auckland, Auckland, New Zealand

Containers (with Docker or Singularity) allow packaging up all code and dependencies to ensure that analyses run reliably across a range of operating systems and software versions.
https://doi.org/10.1177/25152459211017853

Containers at NIH are easy!
https://hpc.nih.gov/apps/singularity.html

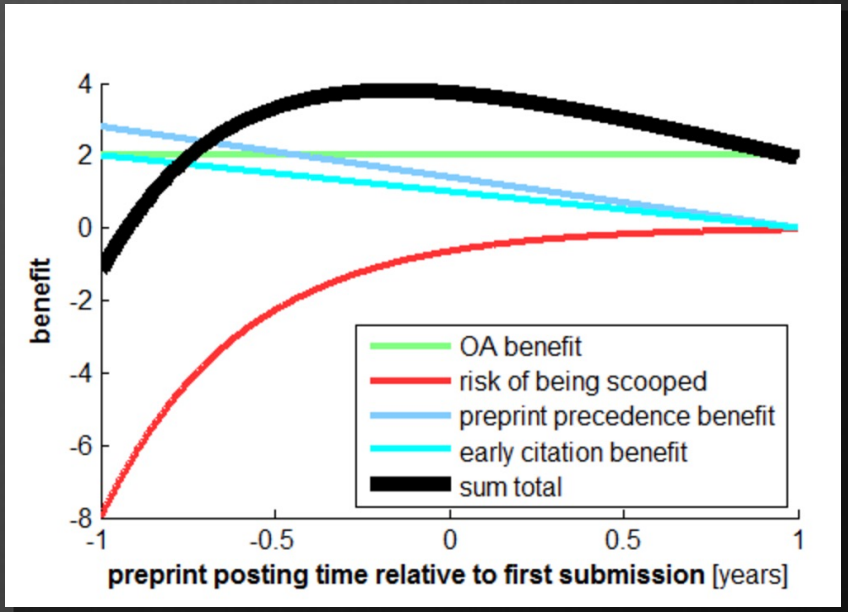
Preprint posting

arXiv.org

bioRxiv beta

THE PREPRINT SERVER FOR BIOLOGY

- Benefits:
  - Open access
  - Catch errors
  - Earlier citation
  - Earlier precedence, prevent scooping
  - Speed and improve final submission



Kriegeskorte (2016) DOI:10.15200/winn.145838.88372

Standards – EQUATOR & COBIDAS

Checklists

CONSORT 2010 checklist of information to include when reporting a randomised trial*

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	
	2b	Specific objectives or hypotheses	
Methods			
Trial design	3a	Description of trial design (such as p	
	3b	Important changes to methods after	
Participants	4a	Eligibility criteria for participants	
	4b	Settings and locations where the dat	
Interventions	5	The interventions for each group wit actually administered	
Outcomes	6a	Completely defined pre-specified pri were assessed	
	6b	Any changes to trial outcomes after	
Sample size	7a	How sample size was determined	
	7b	When applicable, explanation of any	
Randomisation:			

Table D.1. Experimental Design Reporting

Aspect	Notes	Mandatory
Number of subjects	*Elaborate each by group if have more than one group.*	
Subjects approached		N
Subjects consented		N
Subjects refused to participate	Provide reasons.	N
Subjects excluded	Subjects excluded after consenting but before data acquisition; provide reasons.	N
Subjects participated and analyzed	Provide the number of subjects scanned, number excluded after acquisition, and the number included in the data analysis. If they differ, note the number of subjects in each particular analysis.	Y
Inclusion criteria and descriptive statistics	*Elaborate each by group if have more than one group.*	
Age	Mean, standard deviation and range.	Y
Sex	Absolute counts or relative frequencies.	Y
Race & ethnicity	Per guidelines of NIH or other relevant agency.	N

Take Homes

- A large fraction of the people listening to this will author an unreproducible finding
- Some of reproducibility is just part of science, but a much larger fraction is preventable
- YOU can influence your PI to do more reproducible science:
  - Ask to pre-register the study
  - Make your code available, readable, re-runnable
  - Plan for coding errors and include test to catch them
  - Don't be embarrassed of your code and don't mock others
  - Report your work thoroughly using checklists and broadly using pre-print servers

Thanks!

See online slides for more URLs and references:

https://github.com/agt24

http://cmn.nimh.nih.gov/dsst

Questions?

Extra Slides

What is Open Science?

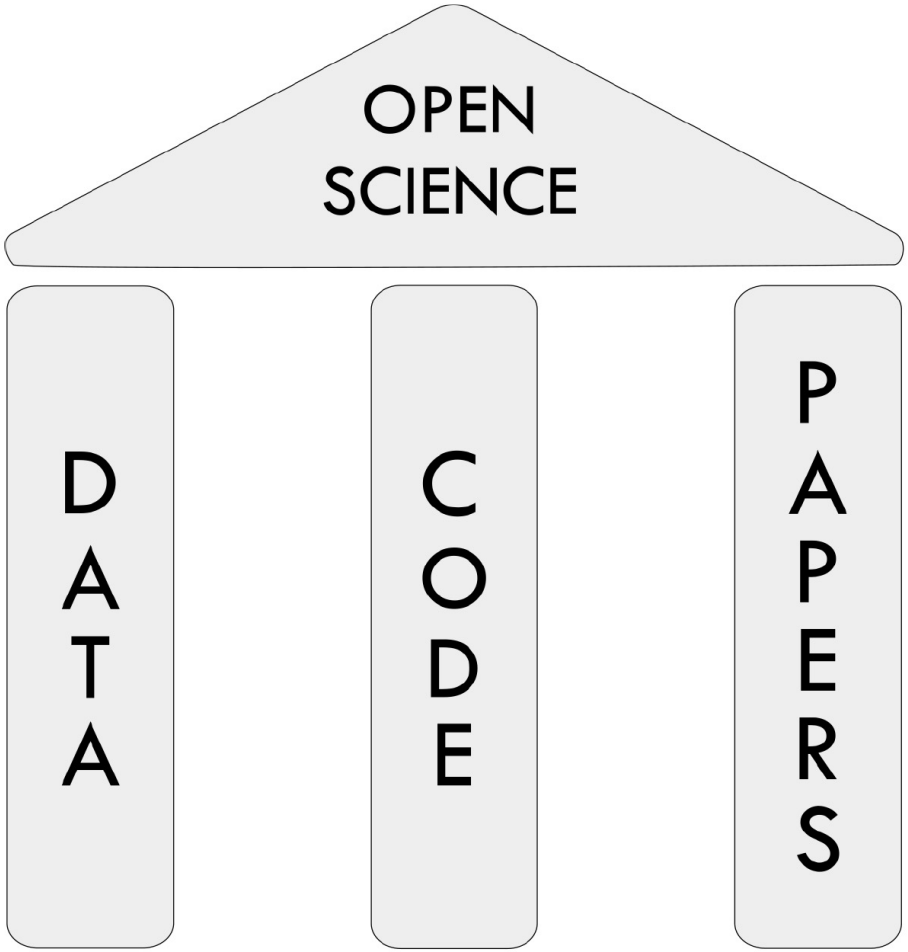


Fig 1. Three pillars of Open Science: data, code, and papers.

doi:10.1371/journal.pbio.1002506.g001

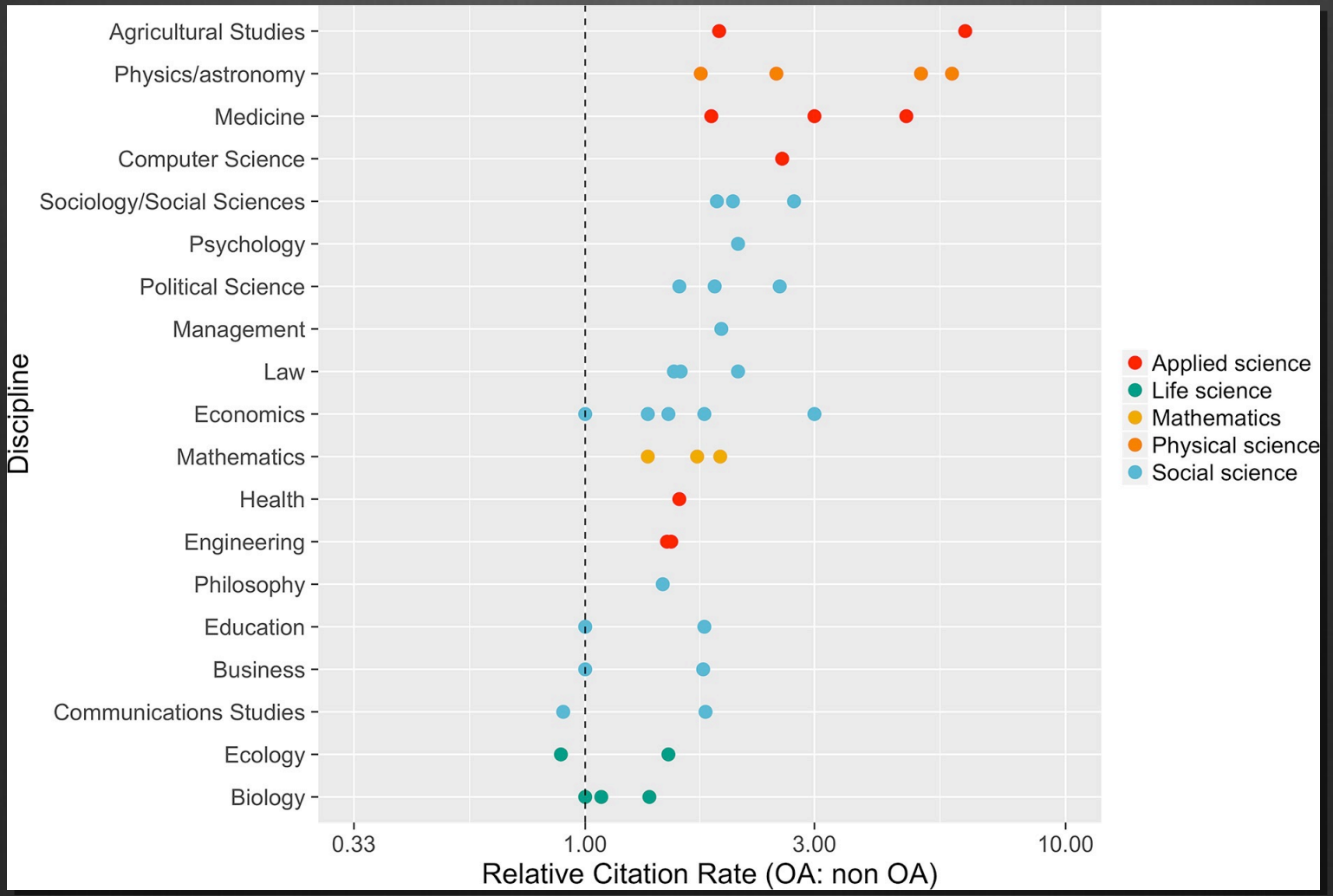
What does Reproducibility even mean?

Varies by field

- Political Science, Economics: The terms are used with no distinction between them.
- Statistics & Most Biomedical Research: "Reproducibility" refers to instances in which the original researcher's data and computer codes are used to regenerate the results, while "replicability" refers to instances in which a researcher collects new data to arrive at the same scientific findings as a previous study.
- Microbiology, Immunology, Computer Science: "Reproducibility" refers to independent researchers arriving at the same results using their own data and methods, while "replicability" refers to a different team arriving at the same results using the original author's artifacts.

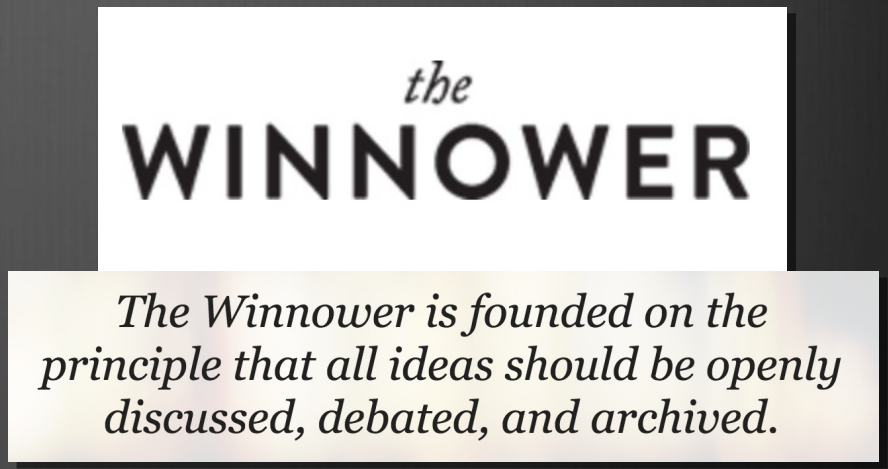
https://www.ncbi.nlm.nih.gov/books/NBK547546/
https://arxiv.org/pdf/1802.03311.pdf

Open Access
Open access publication are cited more



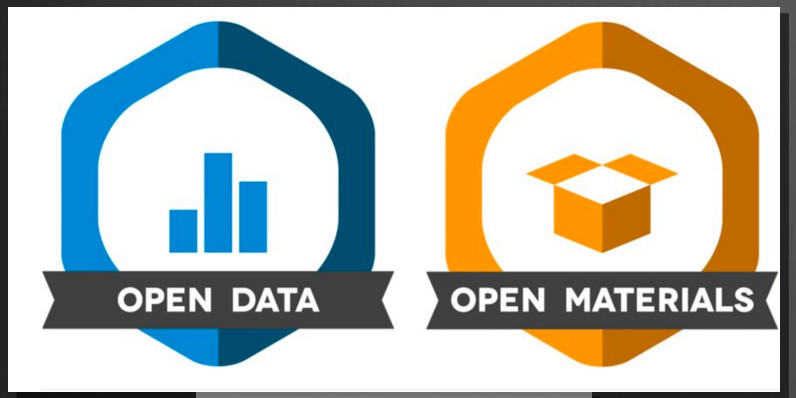
mean citation rate of OA articles divided by mean citation rate of non-OA articles

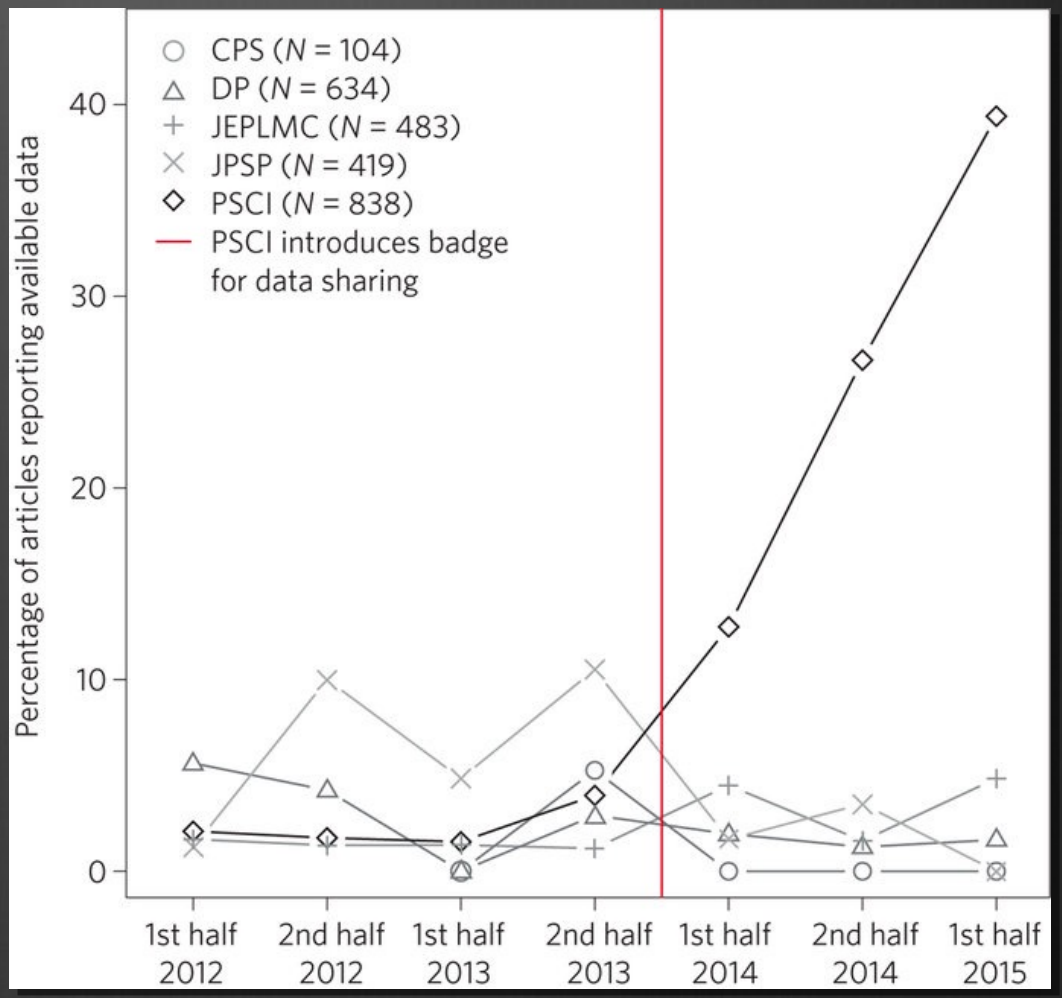
Open Review



- Public discussion of pros and cons of submission
- Optional anonymity
- Prevent low-quality and or biased review

Incentives: Badges





How to be Open – Choose your battles

Be open when you can, as you can

Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

	LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3
Citation standards	Journal encourages citation of data, code, and materials—or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used, consistent with journal's author guidelines.	Article is not published until appropriate citation for data and materials is provided that follows journal's author guidelines.
Data transparency	Journal encourages data sharing—or says nothing.	Article states whether data are available and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Analytic methods (code) transparency	Journal encourages code sharing—or says nothing.	Article states whether code is available and, if so, where to access them.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Research materials transparency	Journal encourages materials sharing—or says nothing	Article states whether materials are available and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Design and analysis transparency	Journal encourages design and analysis transparency or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Preregistration of studies	Journal says nothing.	Journal encourages preregistration of studies and provides link in article to preregistration if it exists.	Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Preregistration of analysis plans	Journal says nothing.	Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.	Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
Replication	Journal discourages submission of replication studies—or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts blind review of results.	Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.