NYU HEALTH SCIENCES LIBRARY

Replicability I

Fred LaPolla, MLS
Alisa Surkis, PhD, MLS
NYU Health Sciences Library

Topics

- What is reproducibility
- The role of reproducibility today
- Study Design

Outcomes: Students will be able to...

- Identify key types of reproducibility/replicability
- Explain why our understanding of concepts like "power" and "alpha" impact replicability
- Discuss how blinding and randomization can impact outcomes

What does "reproducibility" mean to you?

NYU HEALTH SCIENCES LIBRARY

0

Terms

- Reproducibility
- Replicability
- Repeatability

NYU HEALTH SCIENCES LIBRARY

Science Translational Medicine

What does research reproducibility mean?

Steven N. Goodman*, Daniele Fanelli and John P. A. Ioannidis

frontiers in Neuroinformatics

Reproducibility vs. Replicability: A Brief History of a Confused Terminology

Hans E. Plesser[1,2*]

arXiv.org > cs > arXiv:1802.03311

Computer Science > Digital Libraries

Terminologies for Reproducible Research

Lorena A. Barba

CSH Cold Spring Harbor Laboratory

bioRχiv

A statistical definition for reproducibility and replicability

Prasad Patil, Roger D. Peng, Jeffrey Leek

nature methods

Correspondence | Published: 30 July 2015

Clarifying the terminology that describes scientific reproducibility

Ron S Kenett & Galit Shmueli

NYU Langone Health

THE WALL STREET JOURNAL.

World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate

IDEAS | THE SATURDAY ESSAY

The Breakdown in Biomedical Research

Contaminated samples, faulty studies and inadequate training have created a
laboratories and industry, slowing the quest for new treatments and cures

npr WNYC RADIO news arts & life music programs shop

HIDDEN BRAIN A CONVERSATION ABOUT LIFE'S UNSEEN PATTERNS

28:17

When Great Minds Think Unlike: Inside Science's 'Replication Crisis'

May 24, 2016 - 12:10 AM ET

The Economist World politics Business & finance Economics Science & technology Culture

Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 19th 2013 | From the print edition

The New York Times

Many Psychology Findings Not as Strong as Claimed, Study Says

SundayReview

Why Do So Many Studies Fail to Replicate?

National Academies of Sciences 2019 Report:

- Reproducibility: Obtaining consistent results from the same data, computation and analysis (**computational reproducibility**)
- Replicability: Obtaining consistent results across studies on the same question
- Generalizability: The extent that one study applies to other contexts

Ask:

- If someone uses my code to process and analyze my raw data, will I get an identical answer?
- If I repeat the same experiment, will I get a result that is consistent with my original result?
- If someone else tries to replicate my experiment, will they get a result that is consistent with my result?
- Will someone else replicating the experiment draw a conclusion that is consistent with the original?

For this class: Reproducibility vs Replicability

Reproducibility

- Typically expect bitwise agreement
- Exact reproduction of result does not guarantee it's correct

Replicability

- Do not expect every study to replicate
- Lack of replication does not necessarily indicate flaws in the experimental process

Factors Limiting Replicability: The Good

- Complexity of the system under study
- Intrinsic variation in nature
- Variables outside the scope of current scientific knowledge
- Limitations of current technologies
- Prior probability of the scientific hypothesis (i.e. unexpected results more likely not to replicate)

Factors Limiting Replicability: The Bad

- Poor study design
- Poor execution
- Misuse/understanding of statistics
- Researcher bias
- Publication bias

?

Factors Limiting Replicability: The Ugly

- Fraud
- Relatively rare

?

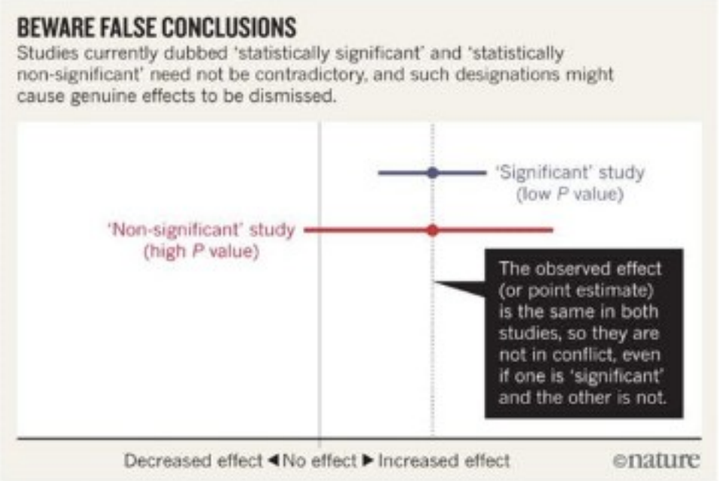
A 2017 Replicability Study

NATURE | COMMENT

A long journey to reproducible results

Gordon J. Lithgow, Monica Driscoll & Patrick Phillips

- Aimed to learn about aging in worms
- Lab technique variability slowed progress
- Found that there were actually cohorts of worms that partition into short and long life



Source: V. Amrhein *et al.*

No litmus test for replicabiilty

- What is being replicated: direction of effect? magnitude?
- Significance can be problematic
- Inherent uncertainty in statistics

NYU HEALTH SCIENCES LIBRARY

NATURE | NEWS

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Response:

PLOS | ONE

A Bayesian Perspective on the Reproducibility Project: Psychology

Alexander Etz , Joachim Vandekerckhove

Published: February 26, 2016 • DOI: 10.1371/journal.pone.0149794

"…analysis revealed no obvious inconsistencies between the original and replication results."

"…apparent failure of the Reproducibility Project to replicate many target effects can be adequately explained by overestimation of effect sizes… due to **small sample sizes** and **publication bias** in the psychological literature.

Do you feel that there is a reproducibility crisis?

NYU HEALTH SCIENCES LIBRARY

0	0	0	0
Yes, a serious problem	Yes, but it is a minor issue	No, this is not a real issue	Unsure

0

Why reproducibility matters

NIH Guidelines as of 2016

- Premise
- Rigorous study design
- Relevant biological variables
- Authentication of resources

The NIH Names Factors Contributing to Poor Reproducibility

- Poor training in experimental design
- Focus on headline grabbing statements
- Lack of detail in publications
- Publication bias
- Misinterpretation of hypothesis-generating research

NYU HEALTH SCIENCES LIBRARY

2015 Review Article

Review

Reproducibility in Science

Improving the Standard for Basic and Preclinical Research

C. Glenn Begley, John P.A. Ioannidis

Table 1. Examples of Some Reported Reproducibility Concerns in Preclinical

Author	Field	Reported Conc
Ioannidis et al (2009)[22]	Microarray data	16/18 studies unable to be reproduced in principle from
Baggerly et al (2009)[23]	Microarray data	Multiple; insufficient data/poor documentation
Sena et al (2010)[24]	Stroke animal studies	Overt publication bias: only 2% of the studies we
Prinz (2011)[1]	General biology	75% to 80% of 67 studies were not reproduc
Begley & Ellis (2012)[2]	Oncology	90% of 53 studies were not reproduced
Nekrutenko & Taylor(2012)[25]	NGS data access	26/50 no access to primary data set
Perrin (2014)[26]	Mouse, in-vivo	0/100 reported treatments repea
Tsilidis et al (2013)[27]	Neurological studies	Too many significant results
Lazic & Essioux (2013)[28]	Mouse VPA model	Only 3/34 used correct experime
Haibe-Kains et al (2013)[29]	Genomics/cell line analysis	Direct comparison of 15 drugs and 4
Witwer (2013)[30]	Microarray data	93/127 articles were not MIAME compliant
Elliott et al (2006)[31]	Commercial antibodies	Commercial antibodies detect wrong antigens
Prassas et al (2013)[32]	Commercial ELISA	ELISA Kit identified wrong antigen
Stodden et al (2013)[33]	Journals	Computational biology: 105/170 journals noncompliant with
Baker et al (2014)[34]	Journals	Top tier fail to comply with agreed standards for animal studies
Vaux (2012)[35]	Journals	Failure to comply with their own statistical guidelines

ALS indicates amyotrophic lateral sclerosis; MIAME, minimum information about a microarray experiment; NGS, next generation sequencing; a
(model of autism).

Table 2. Additional Basic Science Fields Where Concerns Regarding Reproducibility Have Been Raised

Discipline	Issues Raised	Author
Neuroscience	Low statistical power; small sample size	Button et al (2013)[36]
Pharmacology	Lack of training, lack of statistical power, blinding, hypothesis, requisite PK studies, randomization, dose-response, controls, prospective plan, validation, independent replication, and selection of doses that are not tolerable in humans	Henderson et al(2013)[37]; Kenakin et al (2014)[38]; McGonigle et al (2014)[39]; Winquist et al (2014)[40]; Marino (2014)[41]
Genomics/bioinformatics	Irreproducibility of high-profile studies	Sugden et al (2013)[42]
Stem cell biology	Lack of reliable, quality data	Plant & Parker (2013)[43]
Oncology, in vitro testing	Use of clinically unachievable concentrations	Smith & Houghton(2013)[44]
Chemistry lead-discovery	Artifacts; false positives and negatives	Davis & Erlanson (2013)[45]
Computational biology	10 common errors	Sandve et al (2013)[46]
Pathology/Biomarkers	Biospecimen quality	Simeon-Dubach et al (2012)[47]
Organizational psychology	Suppression of negative studies	Kepes & McDonald (2103)[48]
Observational research	0/52 hypotheses confirmed in randomized Trials	Young & Karr (2011)[11]

https://www.ahajournals.org/doi/pdf/10.1161/CIRCRESAHA.114.303819

NYU HEALTH SCIENCES LIBRARY

nature International weekly journal of science

Drug development: Raise standards for preclinical cancer research

C. Glenn Begley & Lee M. Ellis

Scientists in hematology and oncology departments at Amgen tried to confirm findings from 53 “landmark” studies

How many of the 53 landmark studies were confirmed?

NYU HEALTH SCIENCES LIBRARY

0	0	0	0	0	0
0-10	11-20	21-30	31-40	41-52	All of them!

0

Preclinical Research

- In 6 (11%) studies, findings were confirmed
- 25% were consistent enough to continue research (per the standards of Amgen)
- https://www.nature.com/articles/483531a

What do scientists think?

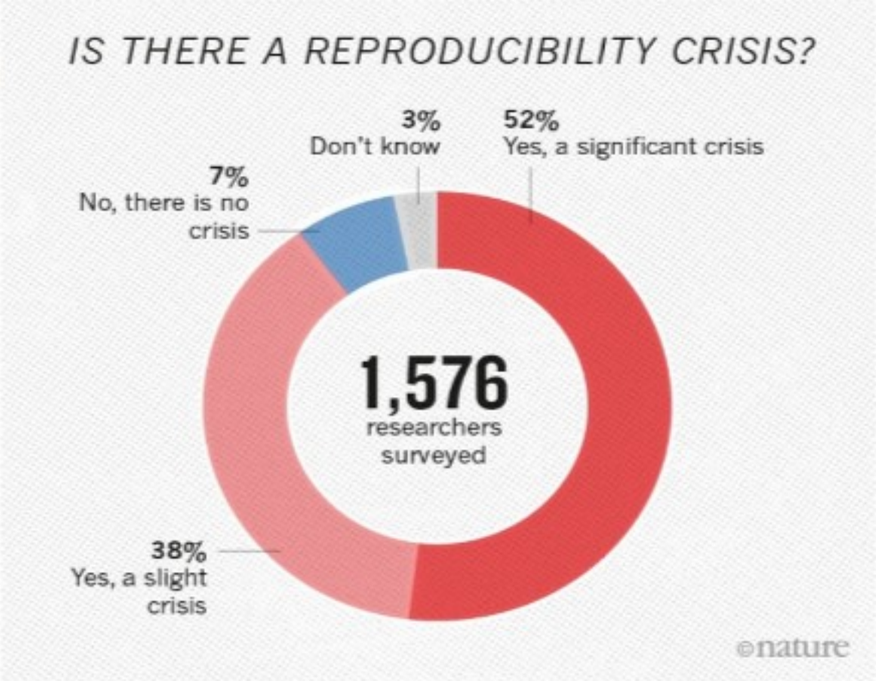
NATURE | NEWS FEATURE

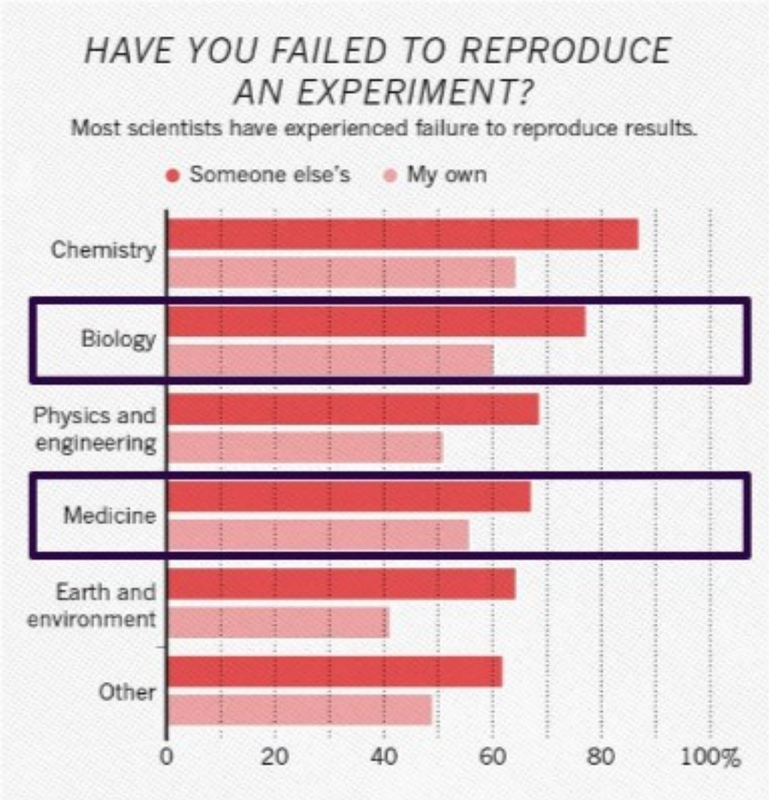
1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

Monya Baker

25 May 2016 |





http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Have you ever tried to replicate someone else's study?

NYU HEALTH SCIENCES LIBRARY

0	0	0
Yes, and I was able to	Yes, but I couldn't get consistent results	Never tried

0

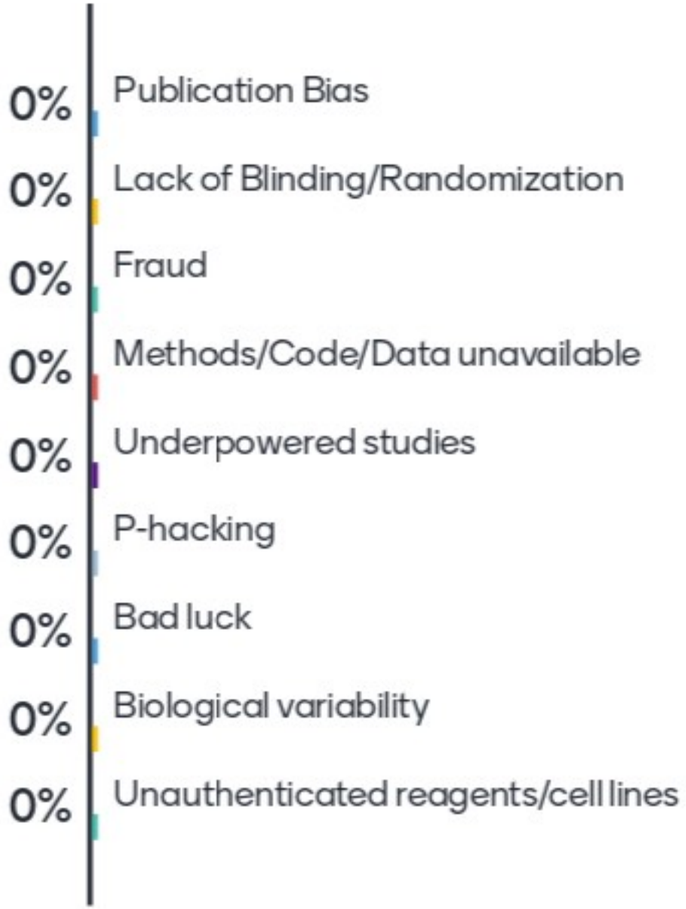
If you tried and were unable to replicate: what were the barriers and what was learned?

NYU HEALTH SCIENCES LIBRARY

0

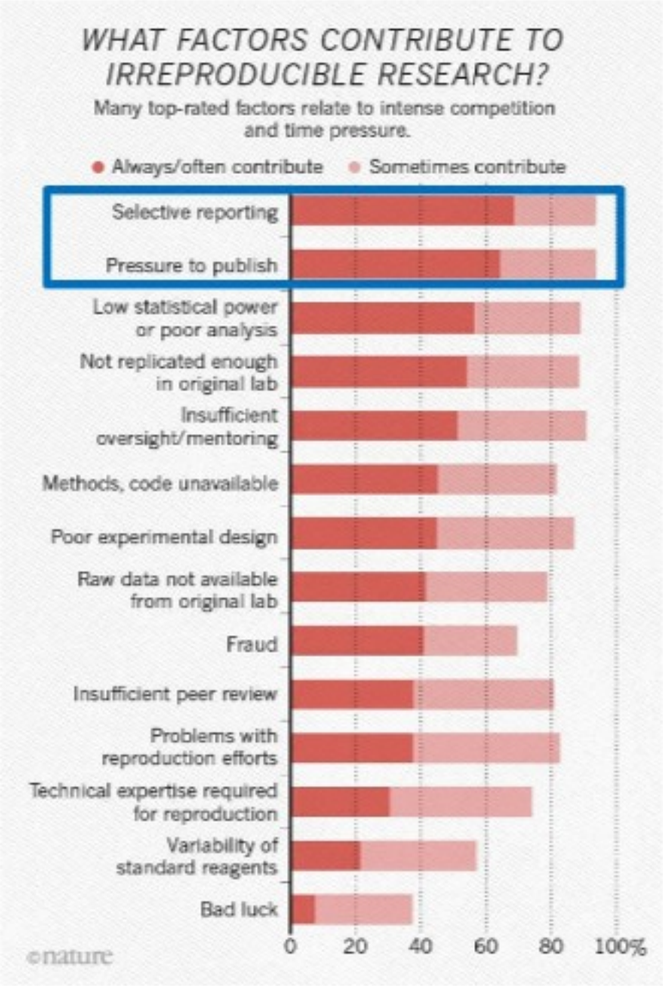
Assign points to each based on how big a factor you think it is in the current reproducibility/replicability crisis

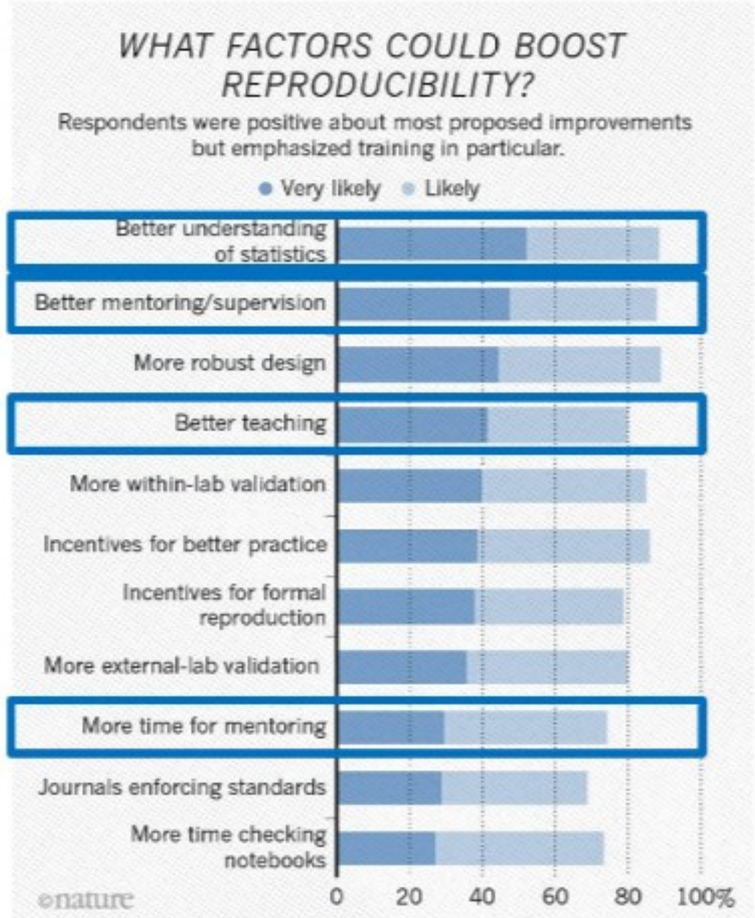
NYU HEALTH SCIENCES LIBRARY



0

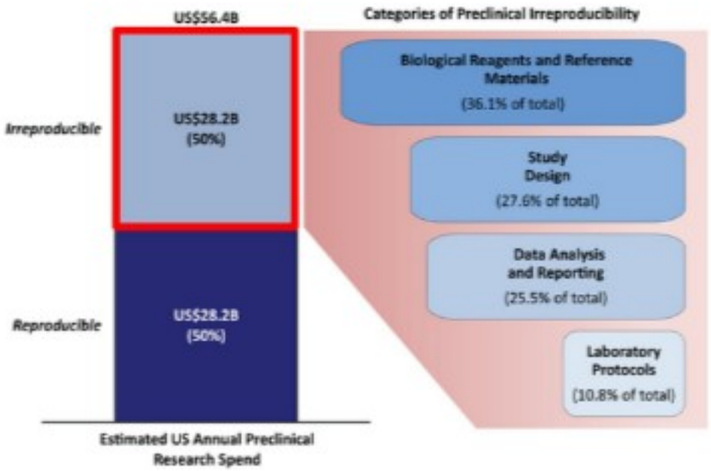
What do scientists think?





http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

The Economics of Reproducibility in Preclinical Research
PLoS Biology, 2015



https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002165

Fig 2. Estimated US preclinical research spend and categories of errors that contribute to irreproducibility.

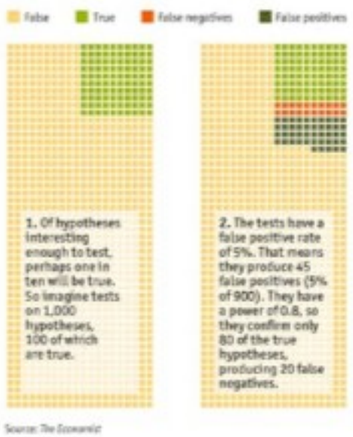
Estimated $28 Billion on Irreproducible Research

- Looked at study design, biological resources, protocols and analysis
- Estimated an upper and lower bound of of impact on reproducibility
- Estimate based on $56 Billion total on preclinical research

Study Design

Power!

- Statistical power is the probability a test will discover differences when they exist
- Higher power lowers the probability of Type II Error (Power is the probability of avoiding Type II error)
- 0.8 is a commonly used benchmark
- Power is influenced by sample size, effect size and variance
- Larger samples are needed to detect smaller effects



A thought experiment

- 100/1000 hypotheses are true
- If Power is 0.8, we expect to find 80 true positives will be detected
- If alpha = .05, we can expect to find 0.5*900 = 45 false positives
- Of the 125 positive findings, 36% (45) may be false

nature REVIEWS NEUROSCIENCE

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson & Marcus R. Munafò

Nature Reviews Neuroscience **14**, 365–376 (2013) | Download Citation

A 2013 study looked at the power levels in neuroscience studies

In what range do you think the median statistical power fell?

NYU HEALTH SCIENCES LIBRARY

0	0	0	0
< 0.25	>= 0.25 & < 0.5	>= 0.5 & < 0.75	>= 0.75

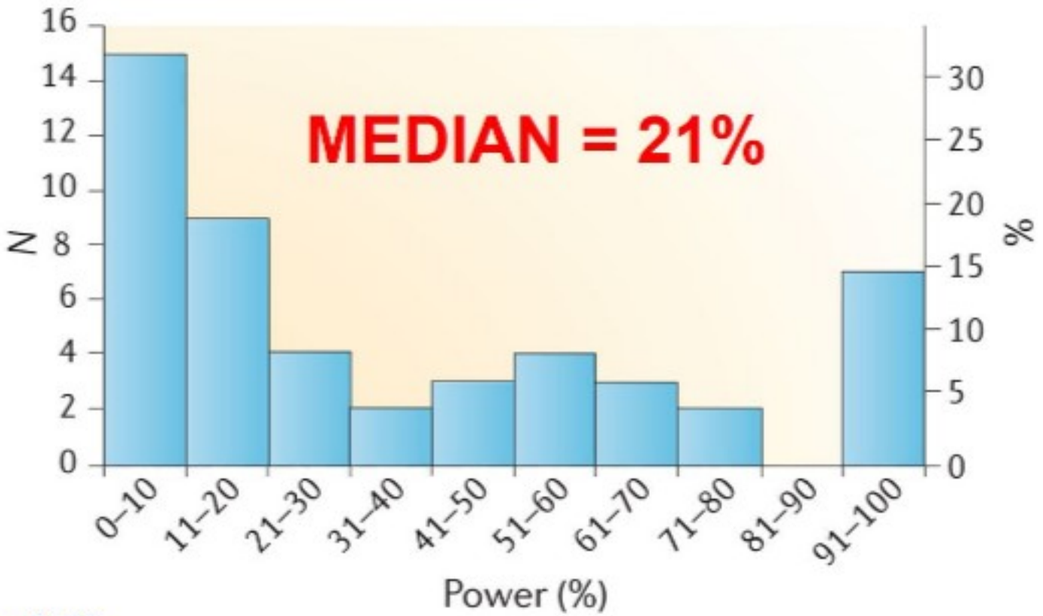
0

nature REVIEWS NEUROSCIENCE

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson & Marcus R. Munafò

Nature Reviews Neuroscience **14**, 365–376 (2013) | Download Citation



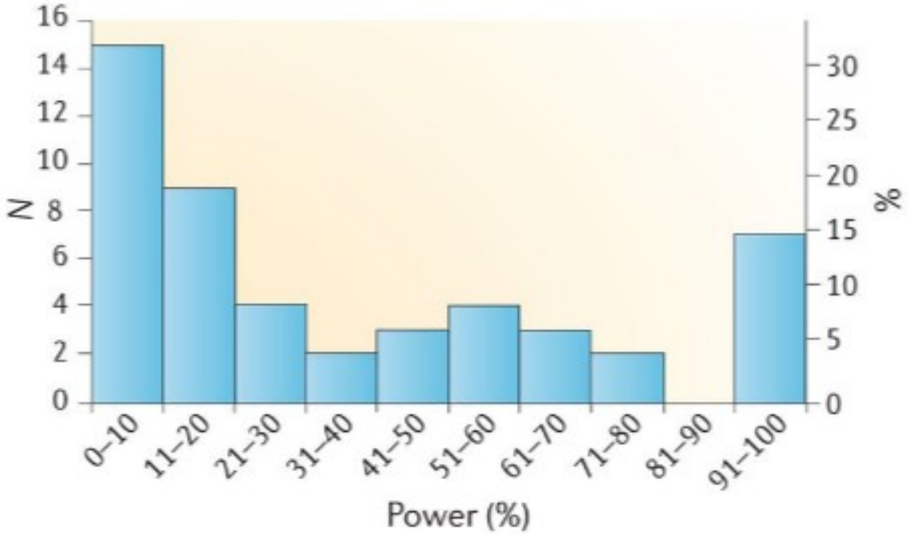
https://www.nature.com/articles/nrn3475

A power of .21 greatly increases the likelihood of type II error

For power of 0.20, how many times out of 100 would we expect to detect a difference between groups when one is present?

NYU HEALTH SCIENCES LIBRARY

0



- 100/1000 hypotheses are true
- Power = 0.2

- 0.2*100 = power to detect 20 True Positives
- 900*.05 = possibility to miss 45 False Positives
- 45/65 = 69% of findings may be False Positives

"Winner's Curse"

- For small study, a smaller actual effect size will not reach statistical significance
- Therefore, low powered study only large effects will be "significant"
- "Lucky" scientist who discovers effect in a small study is cursed by finding an inflated effect

Ways to determine a meaningful effect size

- Look at effect in pilots
- Clinical/subject expertise
- Variation
- Consult with a statistician (NYU Biostatistics Resource)

Bias

In the context of science, what is bias?

NYU HEALTH SCIENCES LIBRARY

0

Bias:

The **systematic** introduction of error

COGNITIVE FALLACIES IN RESEARCH

HYPOTHESIS MYOPIA

Collecting evidence to support a hypothesis, not looking for evidence against it, and ignoring other explanations.

TEXAS SHARPSHOOTER

Seizing on random patterns in the data and mistaking them for interesting findings.

ASYMMETRIC ATTENTION

Rigorously checking unexpected results, but giving expected ones a free pass.

JUST-SO STORYTELLING

Finding stories after the fact to rationalize whatever the results turn out to be.

DEBIASING TECHNIQUES

DEVIL'S ADVOCACY

Explicitly consider alternative hypotheses — then test them out head-to-head.

PRE-COMMITMENT

Publicly declare a data collection and analysis plan before starting the study.

TEAM OF RIVALS

Invite your academic adversaries to collaborate with you on a study.

BLIND DATA ANALYSIS

Analyse data that look real but are not exactly what you collected — and then lift the blind.

go.nature.com/nqyohl

Lack of blinding of outcome assessors in animal model experiments implies risk of observer bias

Blinding:

- 2014 meta-analysis looked at impact of not blinding outcome assessors on estimates of intervention effects in 10 animal studies (2,450 animals)
- Found that unblinded studies exaggerated odds ratios of effect by 59%
- Important to note this does not imply bad intent
- https://www.sciencedirect.com/science/article/abs/pii/S0895435614001577

Randomization

- 2014 systematic review found failure to randomize leads to overestimation of treatment effect
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098856

Do you feel failure to blind or randomize may result in bias? Why or why not?

NYU HEALTH SCIENCES LIBRARY

0

Blinding and Randomization Not Happening

- 2019 study of 574 papers found 56% reported if randomization happened and blinding of outcomes assessors happened in 31%
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0215221
- 2014 study of 2280 papers found 25% randomized, 15-24% blinded
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0089981

In studies you have done, have the experiments

NYU HEALTH SCIENCES LIBRARY

0	0	0	0
Been both blinded and randomized	Blinded to outcomes only	Randomized subjects only	Neither blinded nor randomized

0

Review!

List factors that limit replicability

NYU HEALTH SCIENCES LIBRARY

0

Why does low power limit replicability?

Why might a lack of blinding and randomization limit replicability?

NYU HEALTH SCIENCES LIBRARY

0

Homework

Watch this video: https://www.youtube.com/watch?v=a4fUU85ABwc

|

Write a paragraph (~200 words) addressing the following:

Have labs you've worked in generally employed randomization? Do you think randomization are generally employed in your area of study? If you have employed randomization in a study, did you consider all the factors discussed in the video?

|

Due in Brightspace by **9:00am 2/12/2020**

?

Bibliography

1. National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press.
2. Lithgow GJ, Driscoll M, Phillips P. A long journey to reproducible results. Nature. 2017;548(7668):387-8.
3. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019 Mar;567(7748):305-307.
4. Baker, M. Over half of psychology studies fail reproducibility test. Nature. 2015 August 27.
5. Etz A, Vandekerckhove J. A Bayesian Perspective on the Reproducibility Project: Psychology. PLOS ONE. 2016;11(2):e0149794.
6. Stark PB. Before reproducibility must come preproducibility. Nature. 2018;557(7707):613.
7. Freedman LP, Cockburn IM, Simcoe TS. The Economics of Reproducibility in Preclinical Research. PLOS Biology. 2015;13(6):e1002165.
8. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. Circulation research. 2015;116(1):116-26.
9. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012 Mar 28;483(7391):531-3. doi: 10.1038/483531a.
10. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery. 2011;10:712.
11. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533(7604):452-4.
12. Nuzzo R. How scientists fool themselves - and how they can stop. Nature. 2015 Oct 8;526(7572):182-5.
13. Avey MT, Moher D, Sullivan KJ, Fergusson D, Griffin G, et al. (2016) The Devil Is in the Details: Incomplete Reporting in Preclinical Animal Research. PLOS ONE 11(11): e0166733.
14. Ioannidis JPA. Why Most Published Research Findings Are False. PLOS Medicine. 2005;2(8):e124.
15. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nature Râeviews Neuroscience. 2013;14:365.
16. Bello S, Krogsbøll LT, Gruber J, Zhao ZJ, Fischer D, Hróbjartsson A. Lack of blinding of outcome assessors in animal model experiments implies risk of observer bias. J Clin Epidemiol. 2014 Sep;67(9):973-83.
17. Bebarta V, Luyten D, Heard K. Emergency medicine animal research: does use of randomization and blinding affect the results? Acad Emerg Med. 2003 Jun;10(6):684-7.
18. Hirst JA, Howick J, Aronson JK, Roberts N, Perera R, et al. (2014) The Need for Randomization in Animal Trials: An Overview of Systematic Reviews. PLOS ONE 9(6): e98856.
19. Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, et al. Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. PLOS ONE. 2009;4(11):e7824.
20. van Luijk J, Bakker B, Rovers MM, Ritskes-Hoitinga M, de Vries RB, Leenaars M. Systematic reviews of animal studies; missing link in translational research? PLoS One. 2014 Mar 26;9(3):e89981.
21. Fergusson DA, Avey MT, Barron CC, Bocock M, Biefer KE, et al. (2019) Reporting preclinical anesthesia study (REPEAT): Evaluating the quality of reporting in the preclinical anesthesiology literature. PLOS ONE 14(5): e0215221.