Reproducible Research: Why and How

Part 1: Integrity Problems

Sam Harper

SER Pre-Conference Workshop
2020-10-30

1. Scientific Integrity Problems

1.1 Mertonian norms

1.2 Significance testing

1.3 Non-replication

1.4 Incentive structure

1. Scientific Integrity Problems

1.1 Mertonian norms

1.2 Significance testing

1.3 Non-replication

1.4 Incentive structure

Mertonian Norms in Science

Core Values of Scientific Research

1. Universalism
2. Communality
3. Disinterestedness
4. Organized Skepticism

A NOTE ON SCIENCE AND DEMOCRACY

by ROBERT K. MERTON

SCIENCE, as any other activity involving social collaboration, is subject to shifting fortunes. Difficult as the very notion may appear to those reared in a culture which grants science a prominent if not a commanding place in the scheme of things, it is evident that science is not immune from attack, restraint and repression. Writing a scant thirty-five years ago, Veblen could observe that the faith of western culture in science was unbounded, unquestioned, unrivalled. The revolt from science which then appeared so improbable as to concern only the timid academician who would ponder all contingencies, however remote, has now been forced upon the attention of scientist and layman alike. Local contagions of anti-intellectualism threaten to become epidemic.

Merton (1942), Christensen et al. (2019)

Norms

- *Universalism*: Evaluate research only on its merit.
- *Communality*: Openly share new findings.
- *Disinterestedness*: Motivated by the desire for knowledge and discovery.
- *Skepticism*: Consider all new evidence, hypotheses, theories, and innovations, even those that challenge or contradict their own work.

Counternorms

- *Particularism*: New knowledge from reputation or group.
- *Secrecy*: Protect own findings for private gain.
- *Self-interestedness*: Colleagues are competitors.
- *Dogmatism*: Protecting one's own findings, resisting alternatives.

Potential sources of "bias" in published research

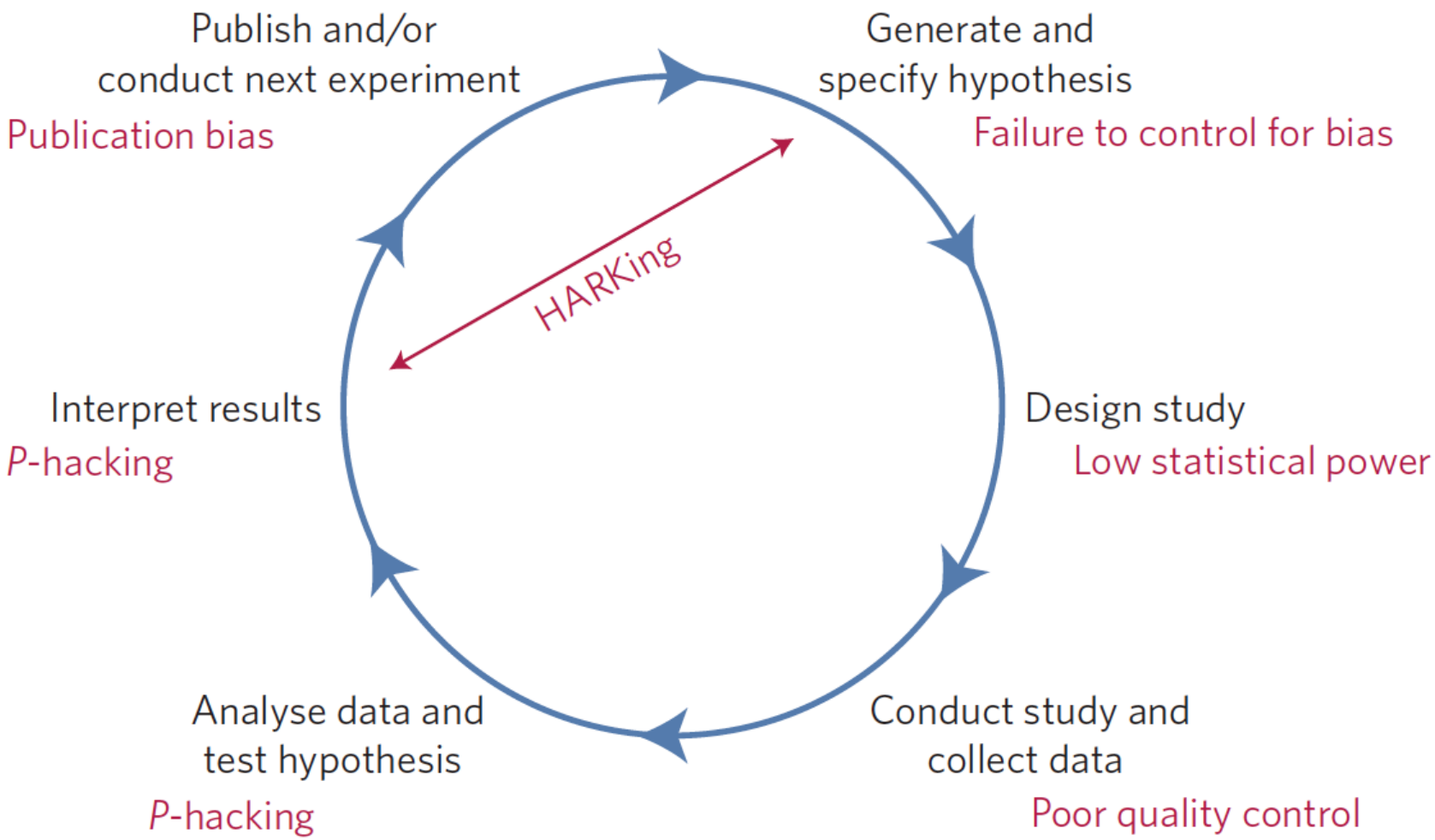
Usual explanations

Confounding, measurement error,
selection bias, model misspecification, etc.

Problems with integrity

- Fraud/data manipulation/fabrication.
- Poor design / inadequate power.
- NHST: Publication bias.
- NHST: P-hacking.
- Financial ties/ideological commitments.
- Careerism.
- Lack of transparency.

Affects the entire research lifecycle



Munafo et al. (2017)

How do we know that science isn't working?

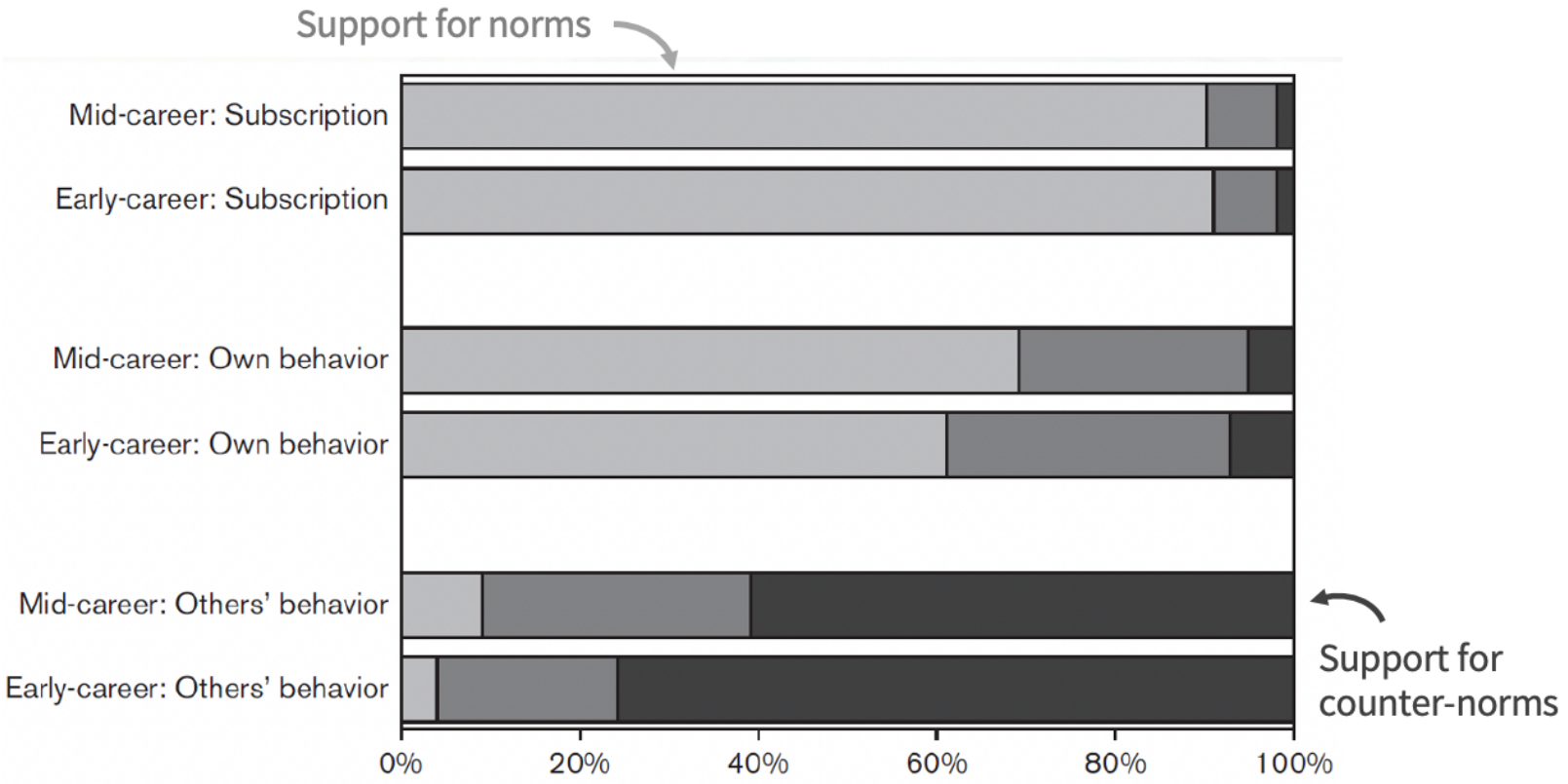
Ask scientists.

Norm support:

"In theory"

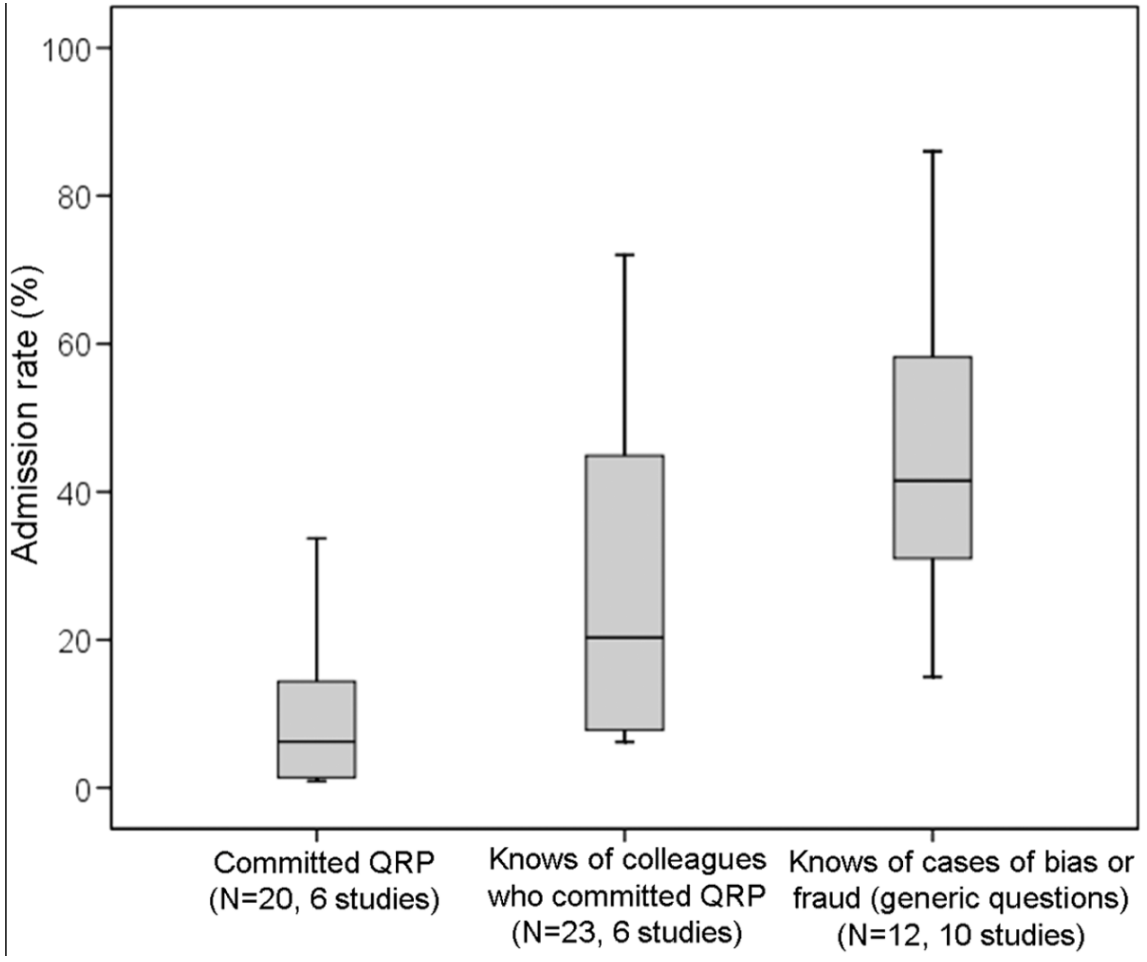
"Me"

"Others"



Christensen et al. (2019) surveyed 3247 US researchers funded by NIH

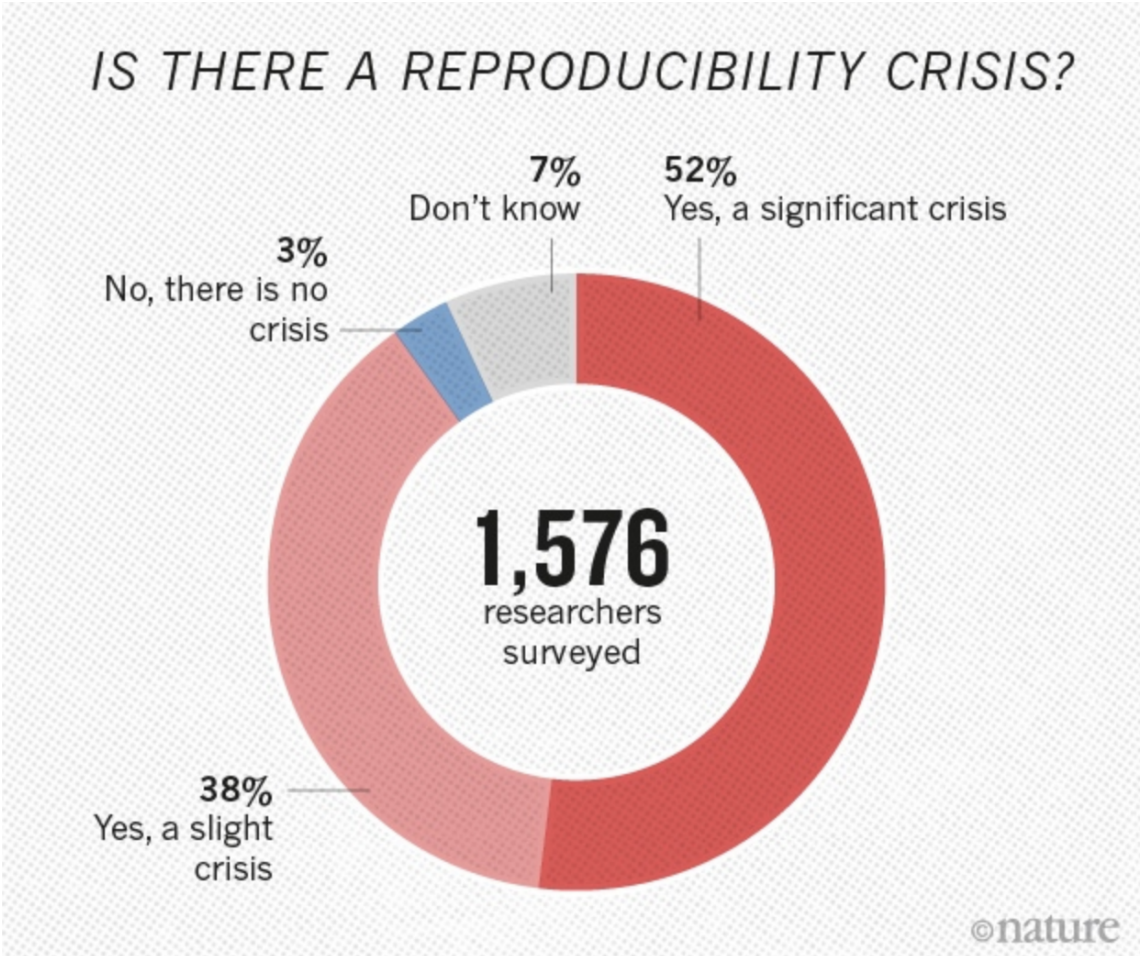
Scientists admit to engaging in questionable research practices.



Fanelli *PLoS ONE* (2011)

Scentists think there is a "reproducibility" crisis

or a "slight" crisis? 🤔



Baker *Nature* (2018)

1. Scientific Integrity Problems

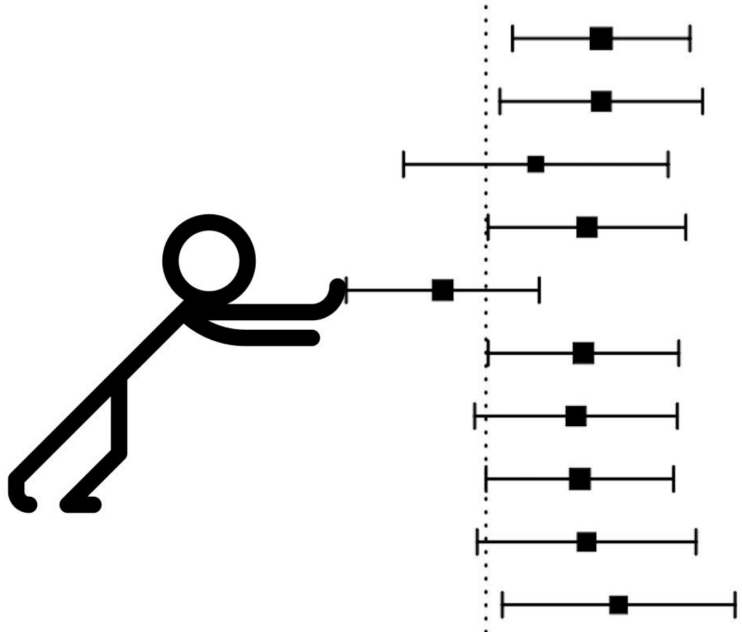
1.1 Mertonian norms

1.2 Significance testing

1.3 Non-replication

1.4 Incentive structure

A lot of irreproducible or unreliable research stems from Null Hypothesis Significance Testing (NHST).



https://mobile.twitter.com/wviechtb/status/1228327958810648576/photo/1

Researcher "degrees of freedom" are difficult to control

How are analyses conducted?

- collect the data over many months.
- finish recording and merging.
- run *one* regression.
- new regression, different controls.
- now a different functional form.
- new regression, different measures.
- yet another regression on subset.
- have 100 or 1000 estimates.
- 1 or maybe 5 results in the paper.

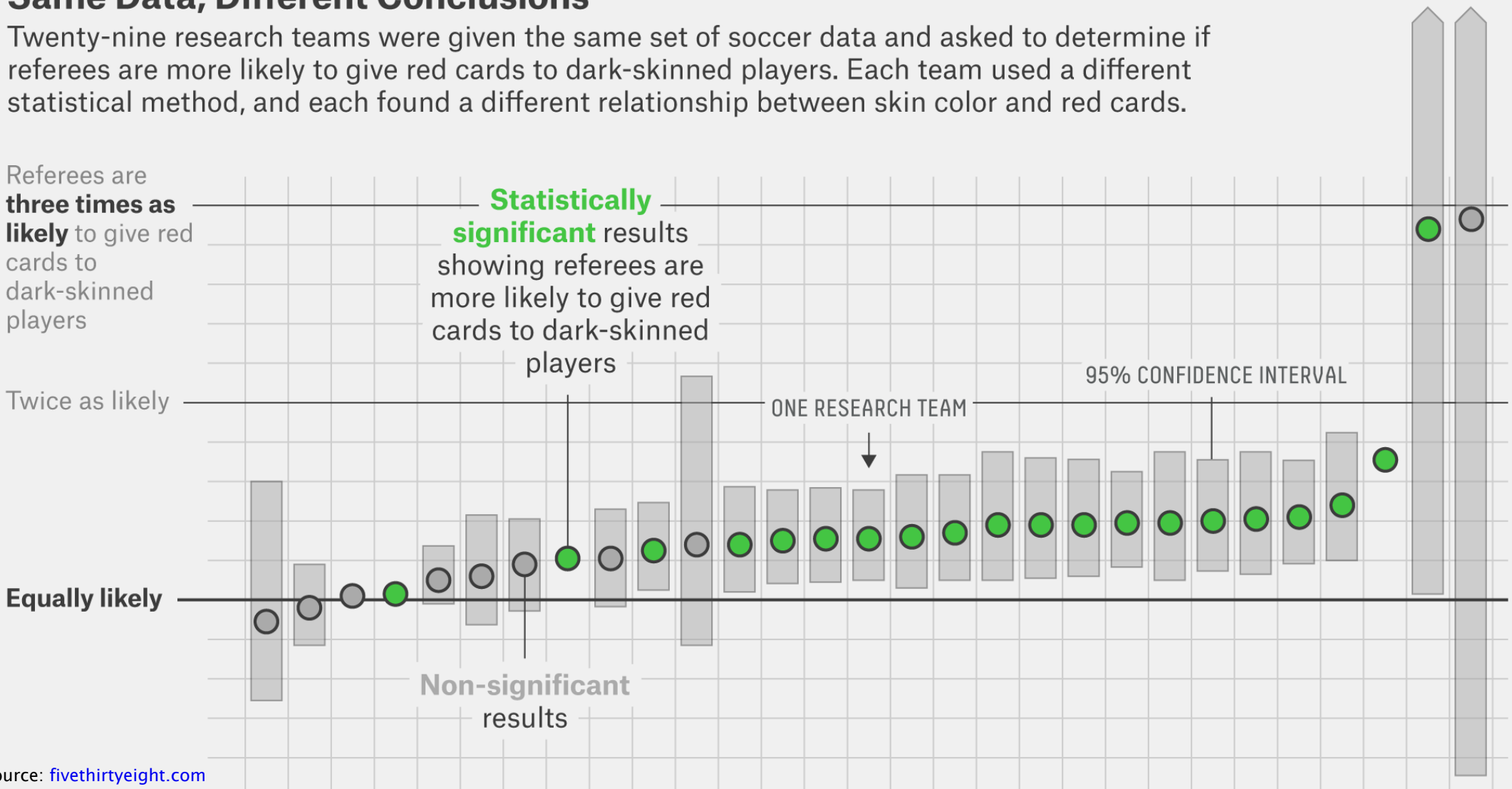
What's the problem?

- Some result is designated as the “correct” one, only *after* looking at the estimates.
- Is this a true test of a hypothesis or just confirmation bias?
- This is "p-hacking"

Source: Gary King

Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.



Source: fivethirtyeight.com

FIVETHIRTYEIGHT

SOURCE: BRIAN NOSEK ET AL.

Let's do some hacking!

Go to https://projects.fivethirtyeight.com/p-hacking/ and answer this question:

Will next week's election affect the economy?

03:00

1 CHOOSE A POLITICAL PARTY

Republicans | **Democrats**

2 DEFINE TERMS

Which politicians do you want to include?

- [x] **Presidents**
- [x] **Governors**
- [x] **Senators**
- [x] **Representatives**

How do you want to measure economic performance?

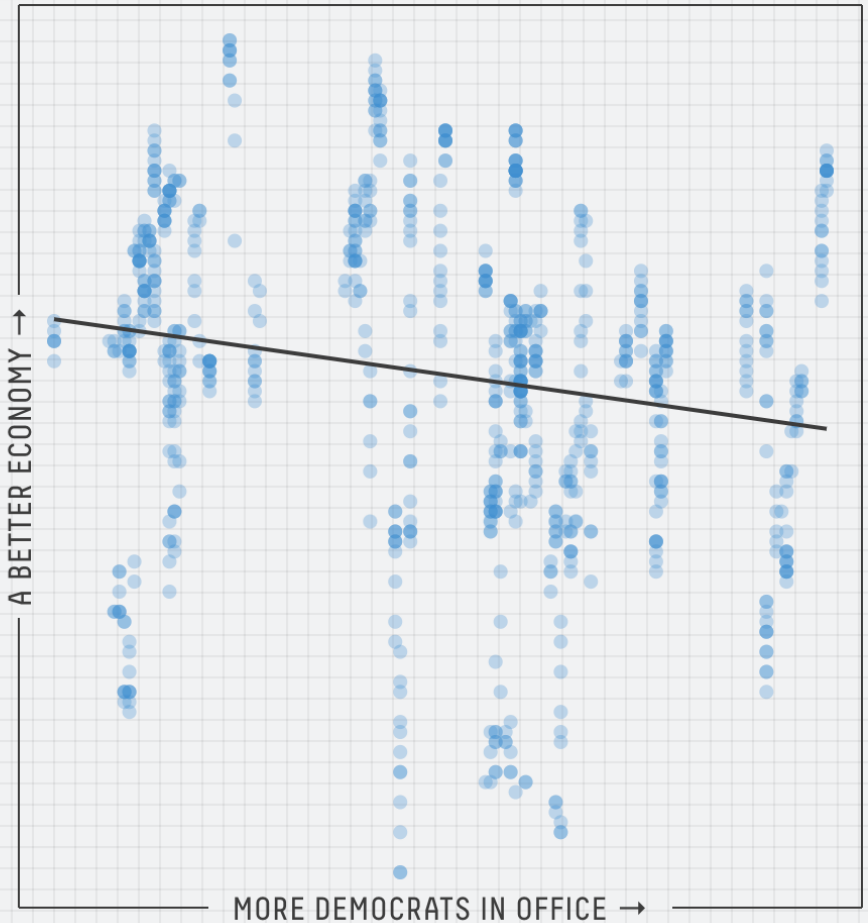
- [x] **Employment**
- [ ] **Inflation**
- [ ] **GDP**
- [ ] **Stock prices**

Other options

- [ ] **Factor in power**
  Weight more powerful positions more heavily
- [ ] **Exclude recessions**
  Don't include economic recessions

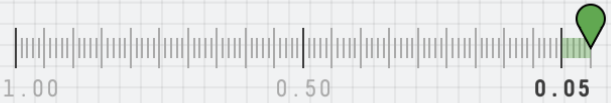
3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in office? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.



Result: Publishable

You achieved a p-value of **less than 0.01** and showed that **Democrats** have a **negative** effect on the economy. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

1 CHOOSE A POLITICAL PARTY

Republicans | **Democrats**

2 DEFINE TERMS

Which politicians do you want to include?

- [x] **Presidents**
- [x] **Governors**
- [x] **Senators**
- [x] **Representatives**

How do you want to measure economic performance?

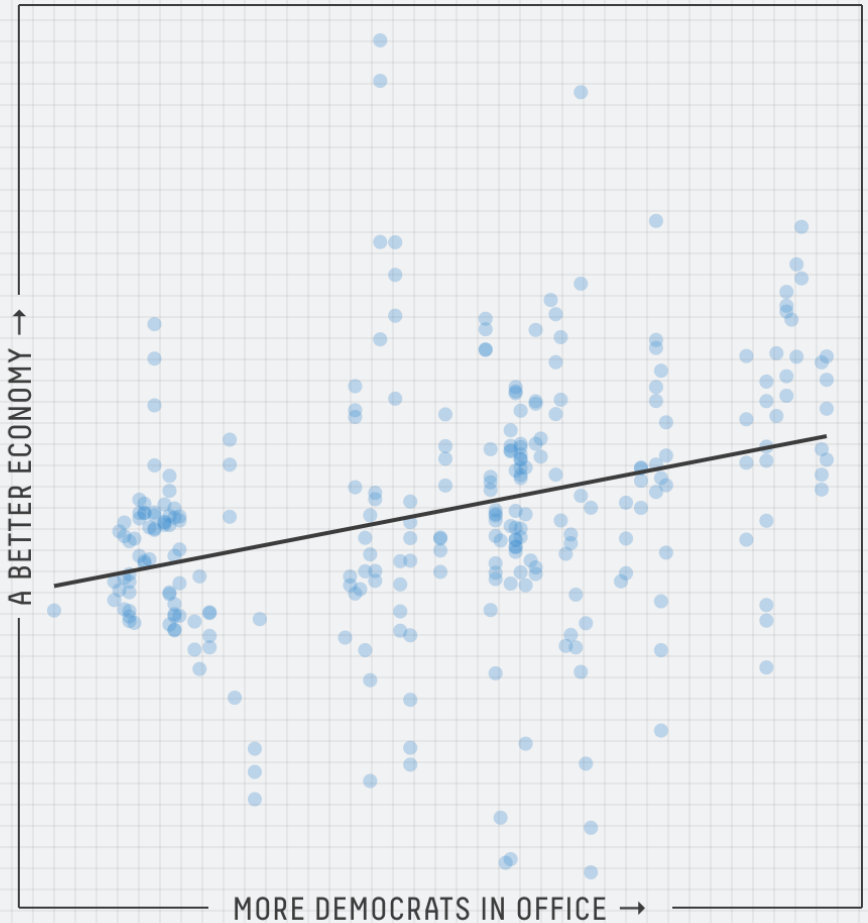
- [ ] **Employment**
- [ ] **Inflation**
- [x] **GDP**
- [ ] **Stock prices**

Other options

- [ ] **Factor in power**
  Weight more powerful positions more heavily
- [ ] **Exclude recessions**
  Don't include economic recessions

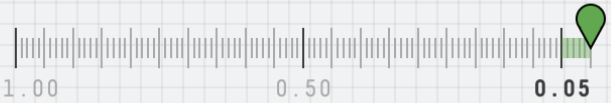
3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in office? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.



Result: Publishable

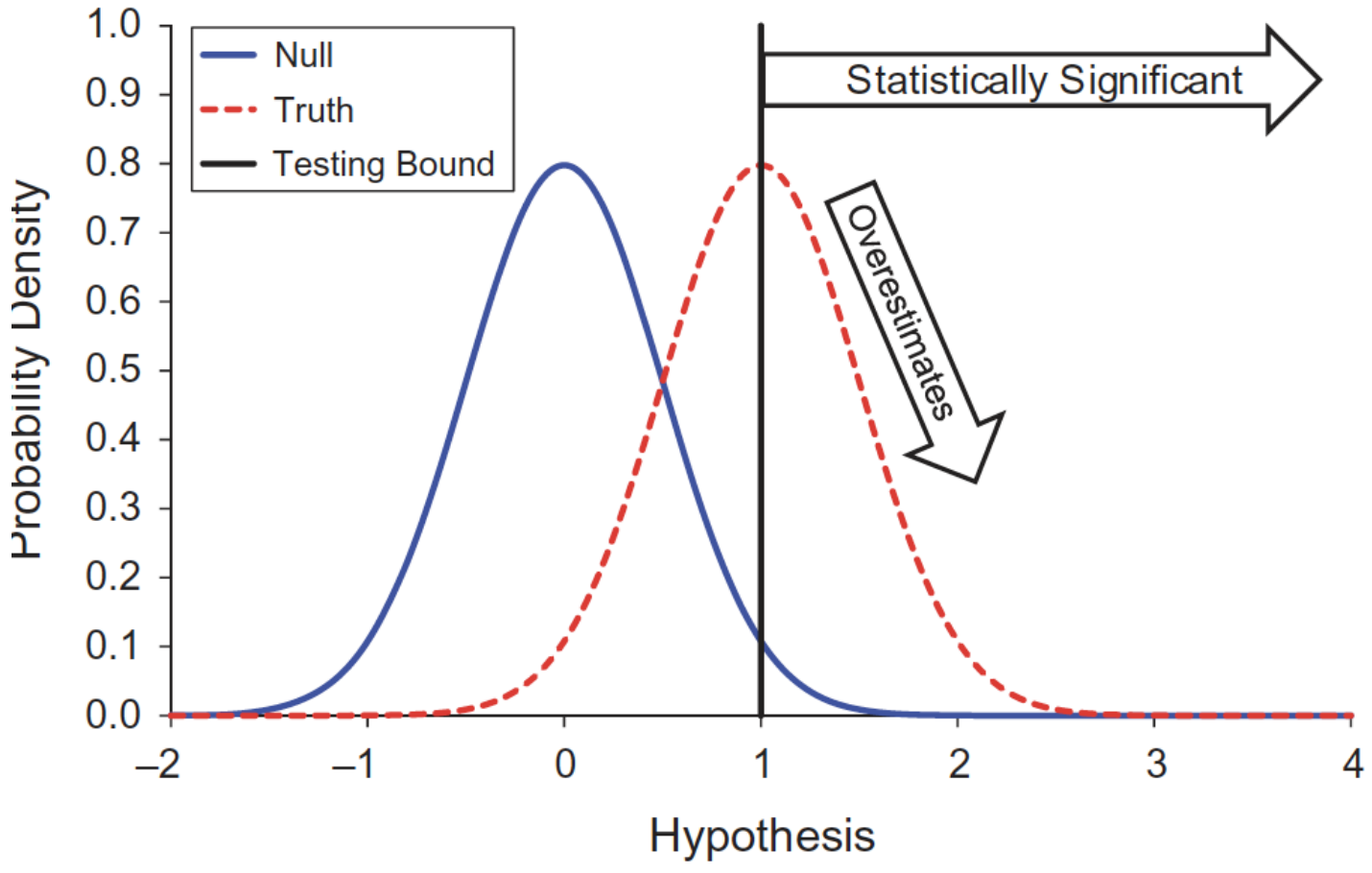
You achieved a p-value of **less than 0.01** and showed that **Democrats** have a **positive** effect on the economy. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

How NHST facilitates non-replication

Study results are sampled from the (---) distribution, but we only see 'statistically significant' ones



Lash (2017)

How do we know there is p-hacking?

(1) Look at what people are doing.

Two estimates:

- HR=0.90, 95%CI: 0.81, 0.99 "Significantly lower"
- HR=0.89, 95%CI: 0.78, 1.00009 "No difference"

Normalization of Testosterone Levels After Testosterone Replacement Therapy Is Associated With Decreased Incidence of Atrial Fibrillation

Rishi Sharma, MD, MHSA; Olurinde A. Oni, MBBS, MPH; Kamal Gupta, MD; Mukut Sharma, PhD; Ram Sharma, PhD; Vikas Singh, MD, MHSA; Deepak Parashara, MD; Surineni Kamalakar, MBBS, MPH; Buddhadeb Dawn, MD; Guoqing Chen, MD, PhD, MPH; John A. Ambrose, MD; Rajat S. Barua, MD, PhD

Background—Atrial fibrillation (AF) is the most common cardiac dysrhythmia associated with significant morbidity and mortality. Several small studies have reported that low serum total testosterone (TT) levels were associated with a higher incidence of AF. In contrast, it is also reported that anabolic steroid use is associated with an increase in the risk of AF. To date, no study has explored the effect of testosterone normalization on new incidence of AF after testosterone replacement therapy (TRT) in patients with low testosterone.

Methods and Results—Using data from the Veterans Administrations Corporate Data Warehouse, we identified a national cohort of 76 639 veterans with low TT levels and divided them into 3 groups. Group 1 had TRT resulting in normalization of TT levels (normalized TRT), group 2 had TRT without normalization of TT levels (nonnormalized TRT), and group 3 did not receive TRT (no TRT). Propensity score–weighted stabilized inverse probability of treatment weighting Cox proportional hazard methods were used for analysis of the data from these groups to determine the association between post-TRT levels of TT and the incidence of AF. Group 1 (40 856 patients, median age 66 years) had significantly lower risk of AF than group 2 (23 939 patients, median age 65 years; hazard ratio 0.90, 95% CI 0.81–0.99, P=0.0255) and group 3 (11 853 patients, median age 67 years; hazard ratio 0.79, 95% CI 0.70–0.89, P=0.0001). There was no statistical difference between groups 2 and 3 (hazard ratio 0.89, 95% CI 0.78–1.0009, P=0.0675) in incidence of AF.

Conclusions—These novel results suggest that normalization of TT levels after TRT is associated with a significant decrease in the incidence of AF. (***J Am Heart Assoc.* 2017;6:e004880. DOI: 10.1161/JAHA.116.004880.**)

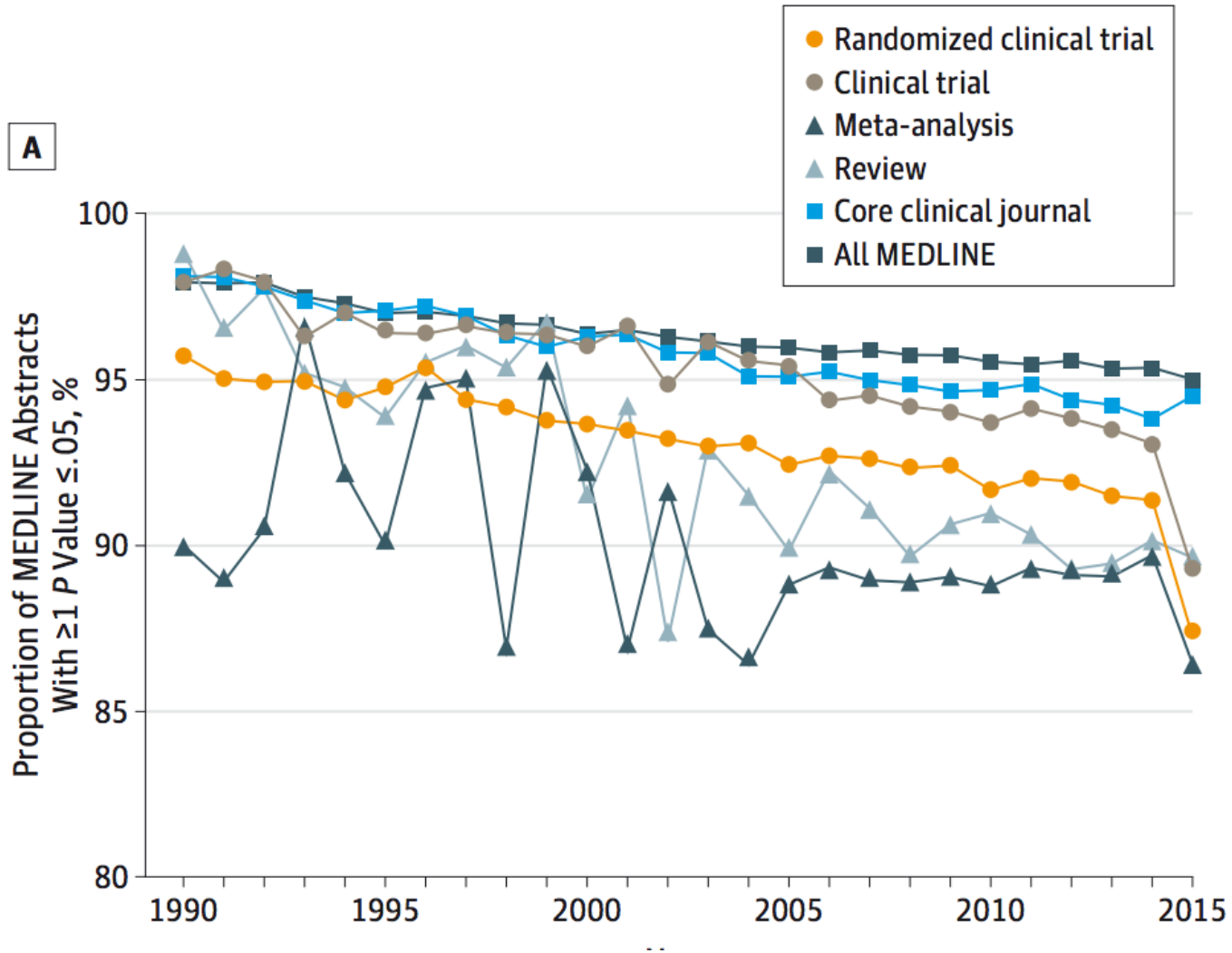
Key Words: atrial fibrillation • testosterone • testosterone replacement therapy

https://www.ahajournals.org/doi/abs/10.1161/jaha.116.004880

How do we know there is p-hacking?

(2) Seriously, everything is significant

P-values in the biomedical literature, 1990-2015



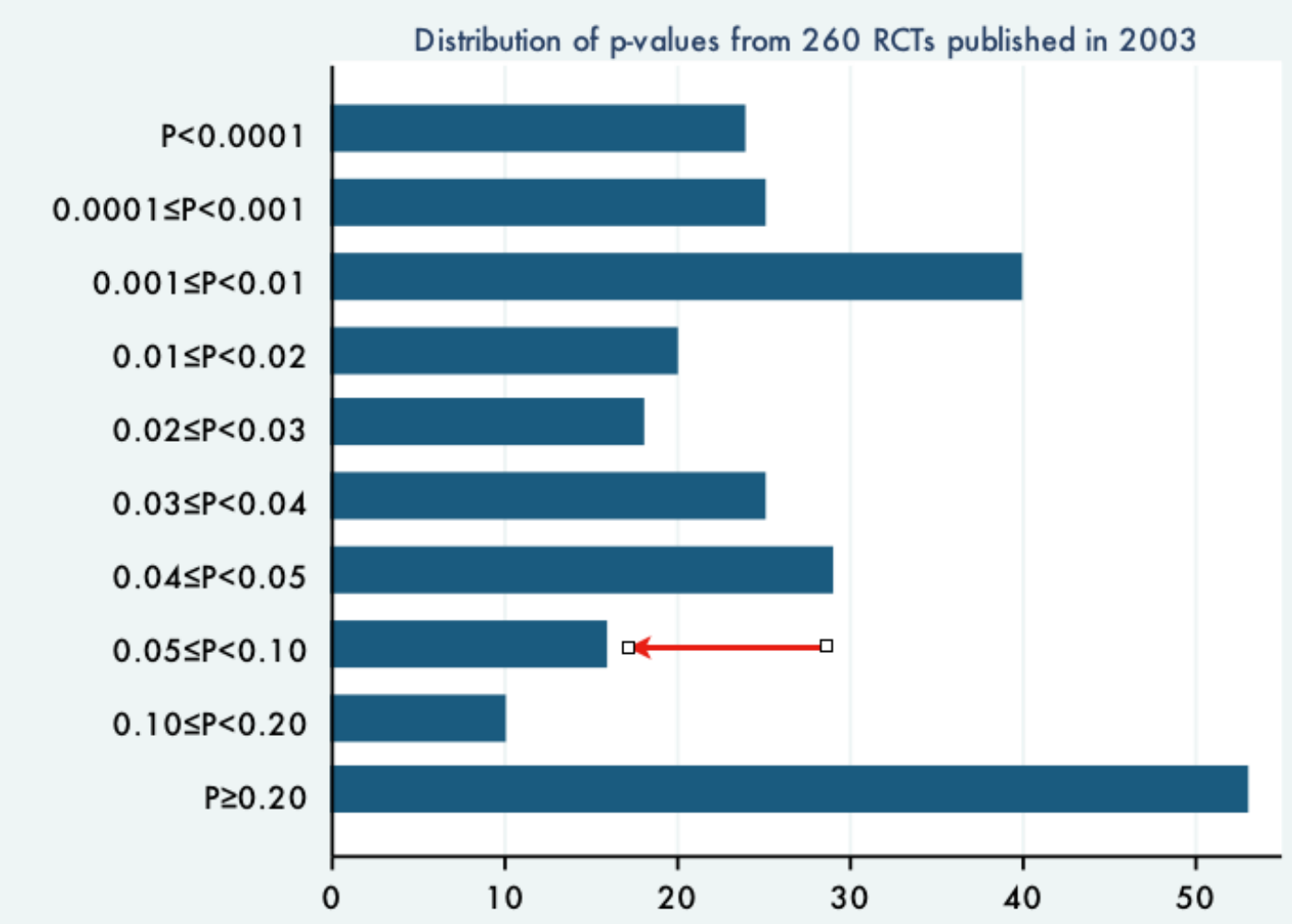
Chavalarais et al. (2013)

How do we know there is p-hacking?

(3) Maldistribution of published p-values

True for medicine, economics, psychology, political science, many other disciplines.

P-values from 260 RCTs

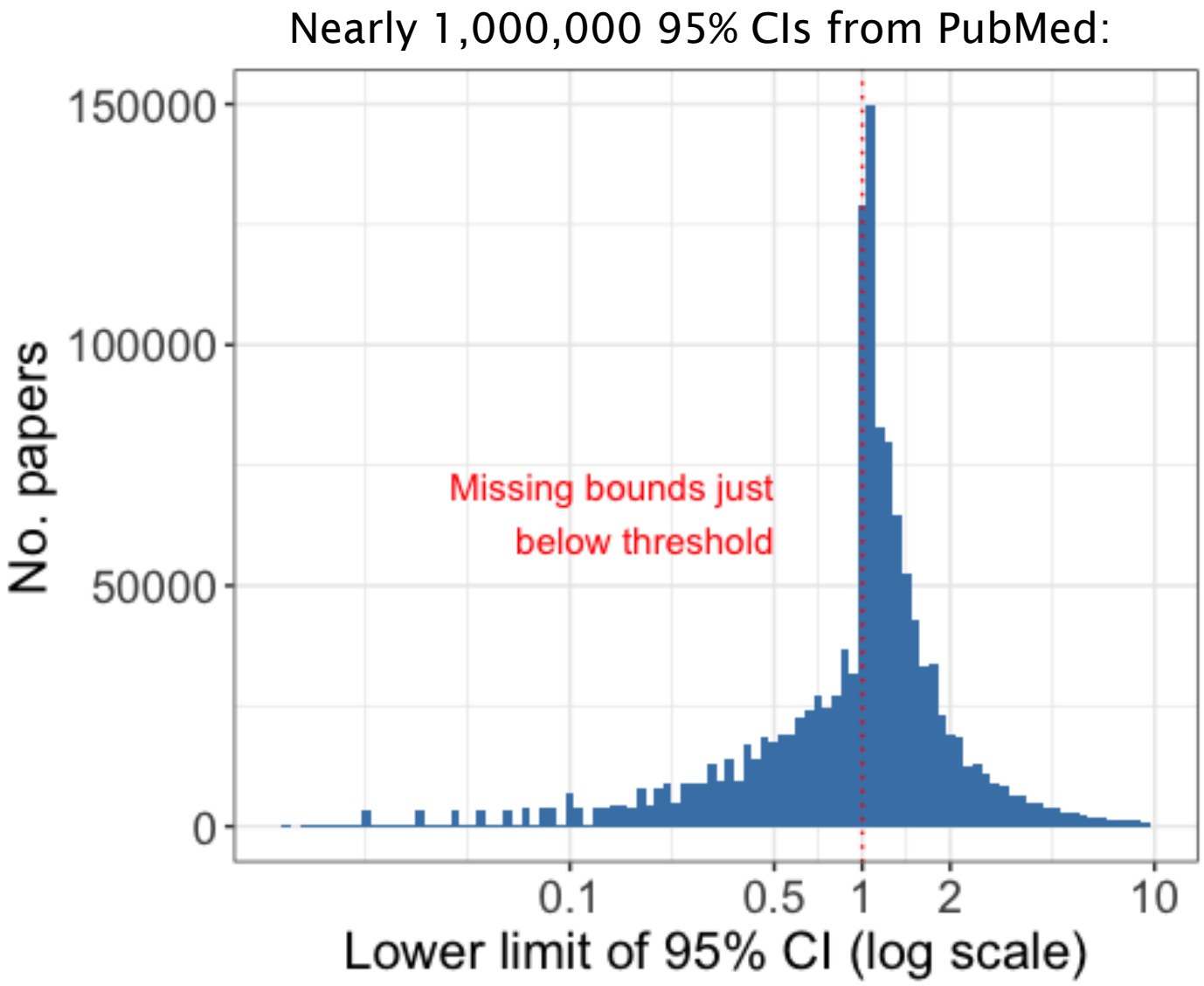


Gotzsche (2006)

Won't 95% confidence intervals help?

No.

Researchers still dichotomize them.



data from Barnett and Wren (2019)

NHST also leads to missing evidence and publication bias

Missing evidence

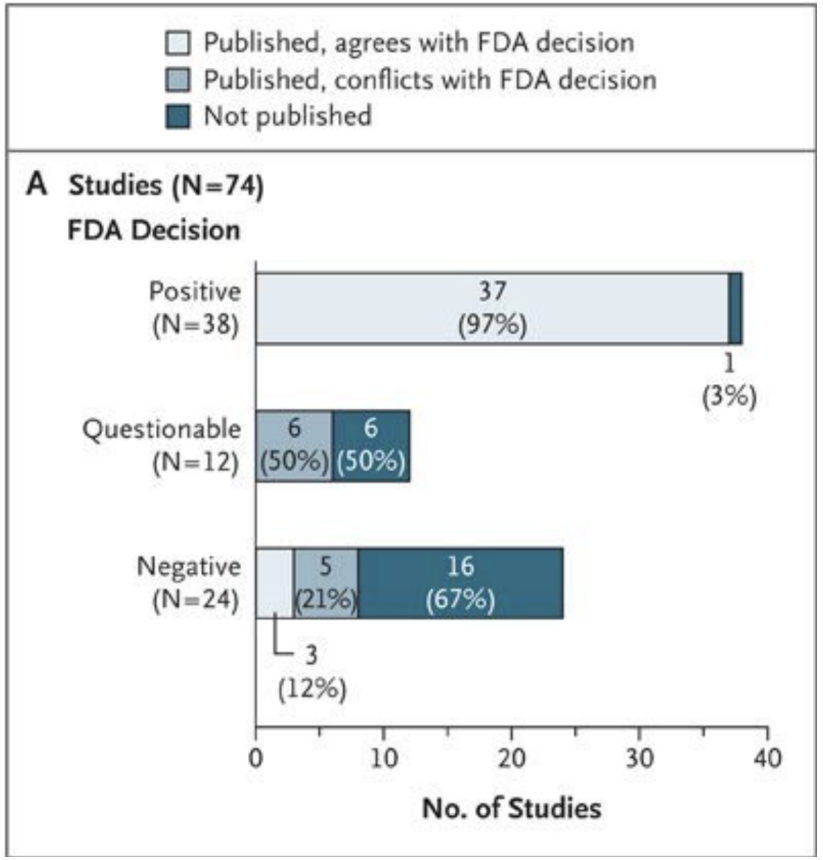
Negative studies of antidepressents less likely to be published.

Impacts regulatory decisions.

SPECIAL ARTICLE

Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy

Erick H. Turner, M.D., Annette M. Matthews, M.D., Eftihia Linardatos, B.S., Robert A. Tell, L.C.S.W., and Robert Rosenthal, Ph.D.



Turner et al. *NEJM* (2008)

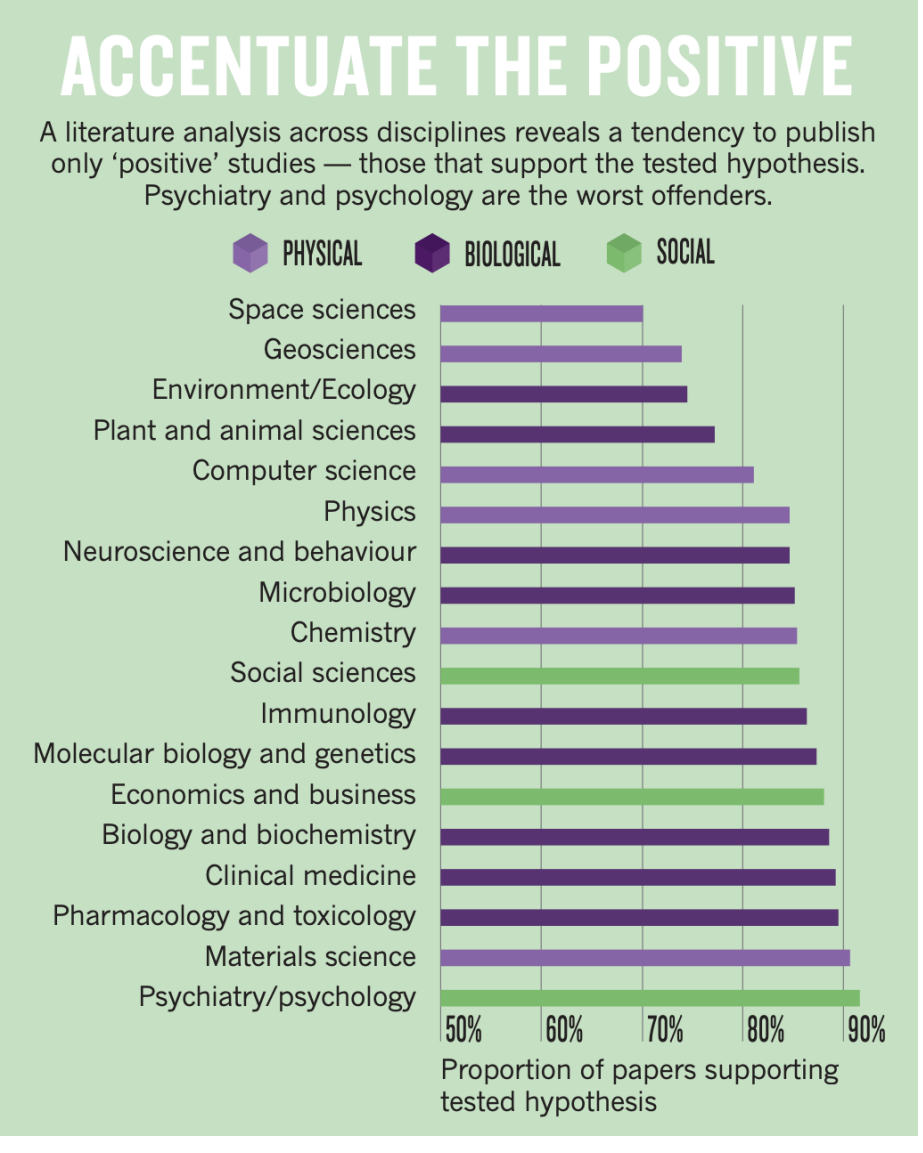
Publication bias affects nearly all disciplines

Statistically significant results are more likely to be published, across virtually all disciplines.

May be worse in "softer" sciences.

Much of the bias is likely self-imposed.

Fanelli *PLoS ONE* (2010), Yong *Nature* (2012)



Self-imposed by many researchers

221 survey experiments funded by US NSF.

All peer reviewed, required to be deposited in a registry.

All studies had results.

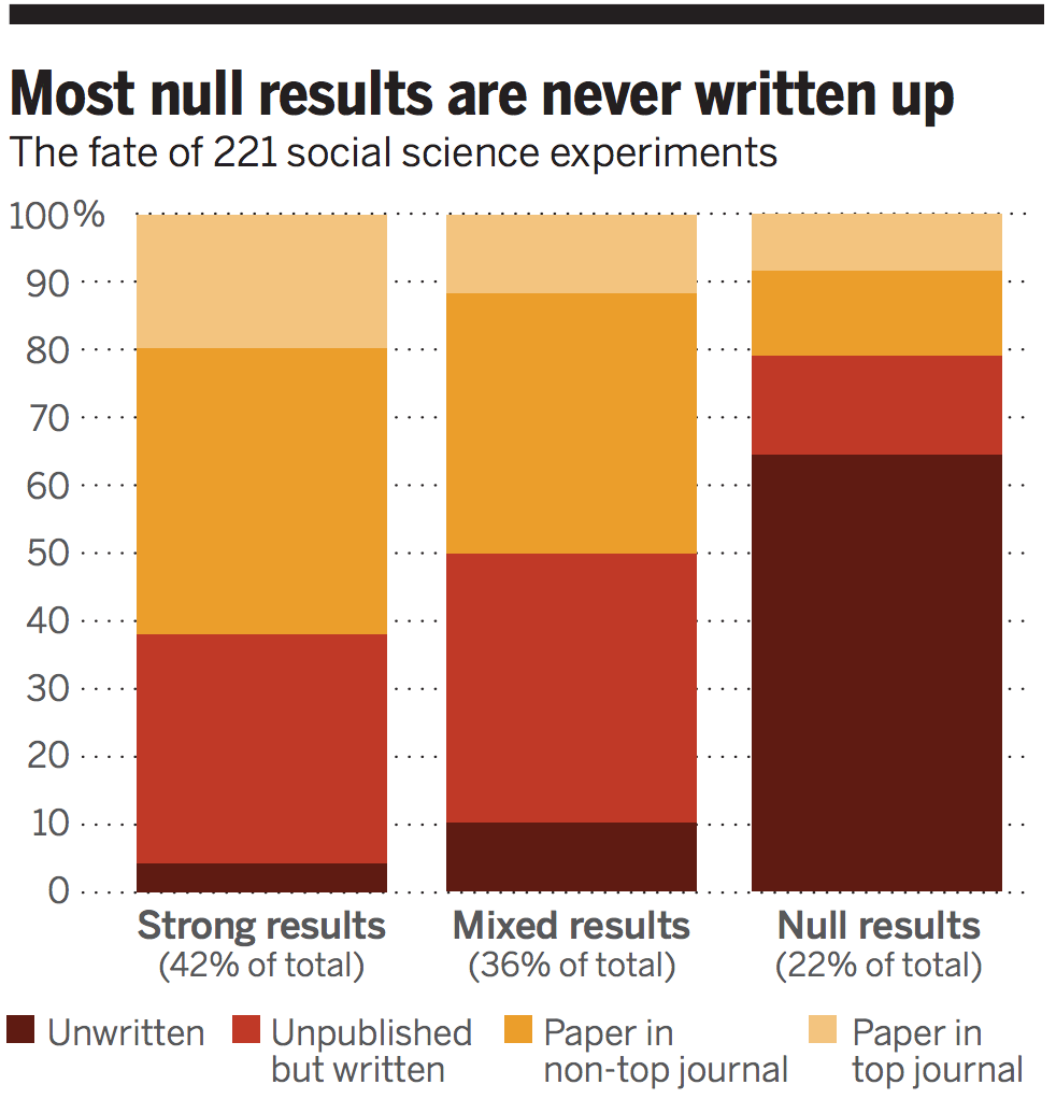


Figure from Mervis in Science 29 Aug 2014;345:992

1. Scientific Integrity Problems

1.1 Mertonian norms

1.2 Significance testing

1.3 Non-replication

1.4 Incentive structure

Distinctions between commonly used terms

Replication

Using using independent investigators, methods, data, equipment, and protocols, we arrive at the same conclusions and/or the same estimate of the effect.

There can be good reasons why findings do not replicate.

Reproducibility

If we start from the *same* data gathered by the scientist we can reproduce the same results, p-values, confidence intervals, tables and figures as in the original report.

There are fewer reasons for non-reproducibility.

National Academy of Sciences (2019)

Large scale efforts to replicate studies are not reassuring

In Psychology

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. 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; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

In Economics

ECONOMICS

Evaluating replicability of laboratory experiments in economics

Colin F. Camerer,[1*†] Anna Dreber,[2†] Eskil Forsell,[2†] Teck-Hua Ho,[3,4†] Jürgen Huber,[5†] Magnus Johannesson,[2†] Michael Kirchler,[5,6†] Johan Almenberg,[7] Adam Altmejd,[2] Taizan Chan,[8] Emma Heikensten,[2] Felix Holzmeister,[5] Taisuke Imai,[1] Siri Isaksson,[2] Gideon Nave,[1] Thomas Pfeiffer,[9,10] Michael Razen,[5] Hang Wu[4]

The replicability of some scientific findings has recently been called into question. To contribute data about replicability in economics, we replicated 18 studies published in the *American Economic Review* and the *Quarterly Journal of Economics* between 2011 and 2014. All of these replications followed predefined analysis plans that were made publicly available beforehand, and they all have a statistical power of at least 90% to detect the original effect size at the 5% significance level. We found a significant effect in the same direction as in the original study for 11 replications (61%); on average, the replicated effect size is 66% of the original. The replicability rate varies between 67% and 78% for four additional replicability indicators, including a prediction market measure of peer beliefs.

Nosek et al. (2017), Camerer et al. (2016)

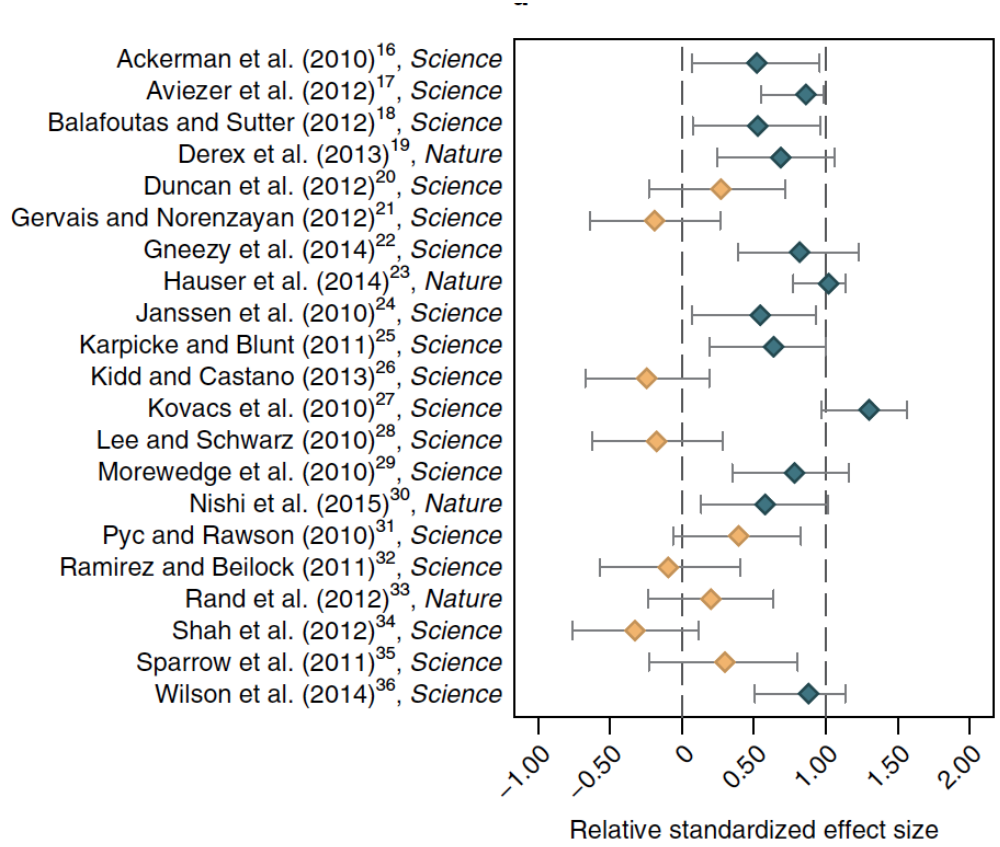
Effect sizes are much lower in replication studies.

Nosek et al. (2017)

Surely the "top" journals are better, right?

"We find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size"

"The relative effect size of true positives is estimated to be 71%, suggesting that both false positives and inflated effect sizes of true positives contribute to imperfect reproducibility."



Camerer et al. (2018)

What about peer review?

Peer review is:

- Slow, inefficient, and expensive.
- Reviewers agreement no better than chance.
- Does not detect errors.

Reviewiers are biased against:

- Less prestigious institutions.
- Against new or original ideas.

Smith (2010), editor at *BMJ* for many years.

If we wanted to reproduce, often the materials aren't there

No raw data, no science: another possible source of the reproducibility crisis

Tsuyoshi Miyakawa

Abstract

A reproducibility crisis is a situation where many scientific studies cannot be reproduced. Inappropriate practices of science, such as HARKing, p-hacking, and selective reporting of positive results, have been suggested as causes of irreproducibility. In this editorial, I propose that a lack of raw data or data fabrication is another possible cause of irreproducibility.

As an Editor-in-Chief of *Molecular Brain*, I have handled 180 manuscripts since early 2017 and have made 41 editorial decisions categorized as "Revise before review," requesting that the authors provide raw data. Surprisingly, among those 41 manuscripts, 21 were withdrawn without providing raw data, indicating that requiring raw data drove away more than half of the manuscripts. I rejected 19 out of the remaining 20 manuscripts because of insufficient raw data. Thus, more than 97% of the 41 manuscripts did not present the raw data supporting their results when requested by an editor, suggesting a possibility that the raw data did not exist from the beginning, at least in some portions of these cases.

Considering that any scientific study should be based on raw data, and that data storage space should no longer be a challenge, journals, in principle, should try to have their authors publicize raw data in a public database or journal site upon the publication of the paper to increase reproducibility of the published results and to increase public trust in science.

Keywords: Raw data, Data fabrication, Open data, Open science, Misconduct, Reproducibility

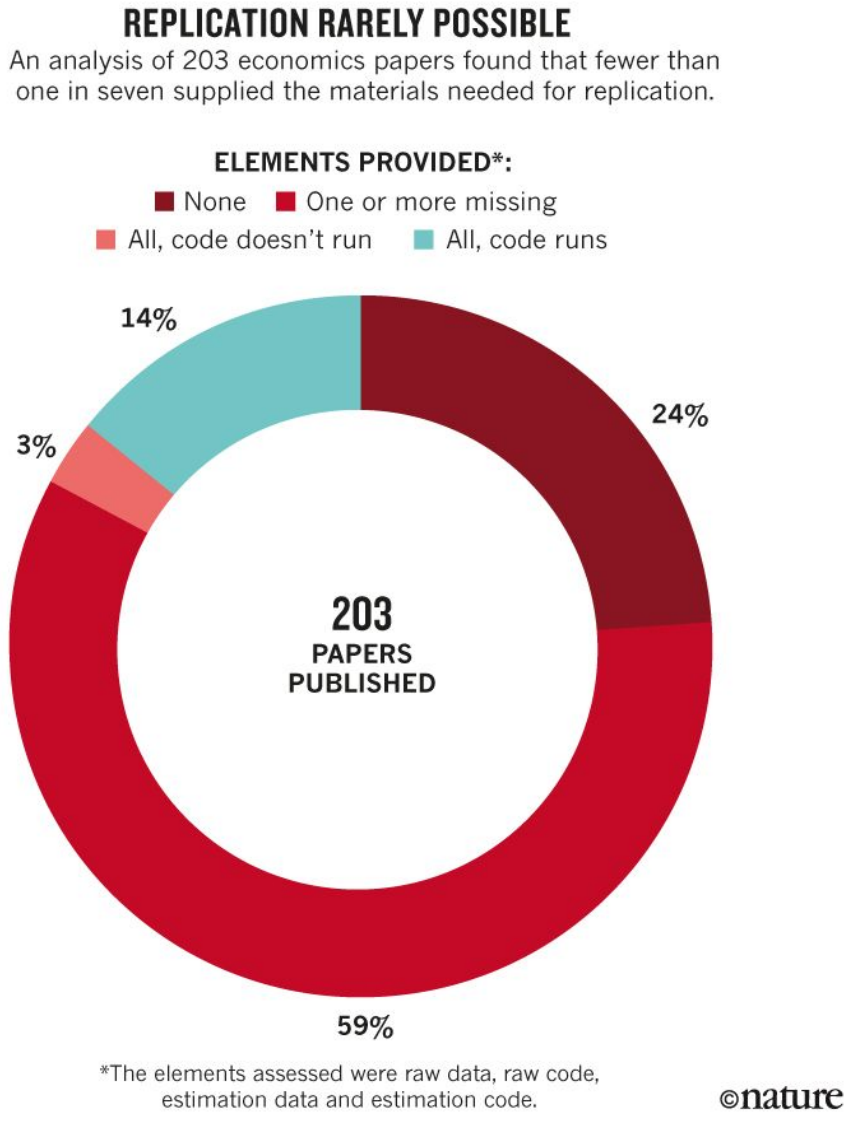
Miyakawa *Molecular Brain* (2020) 13:24

Even with data, efforts to reproduce are rarely successful

Gertler et al. gathered replication materials from published papers in econ.

Most authors only included estimation code.

Estimation code only ran in 40% of cases.



Gertler et al. 2018

1. Scientific Integrity Problems

1.1 Mertonian norms

1.2 Significance testing

1.3 Non-replication

1.4 Incentive structure

Incentive problems

Reward structure

Papers, grants, media, "novel" and "significant" results.

Incentives

Gift authorship, CV padding, salami-slicing

Overstating claims, ignoring "non-significant" results, p-hacking

Hoarding data, non-transparent materials and methods

Incentive problems

Remember Brian Wansink?

After encouraging his postdoc to "find" specific results, fish for interactions, change the dependent variable, and eliminate outliers, he concluded:

> This is really important to try and find as many things here as possible *before* you come. First, it will make a good impression on people and helps you stand out a bit. Second, it would be the highest likelihood of you getting something publishable out of your visit.

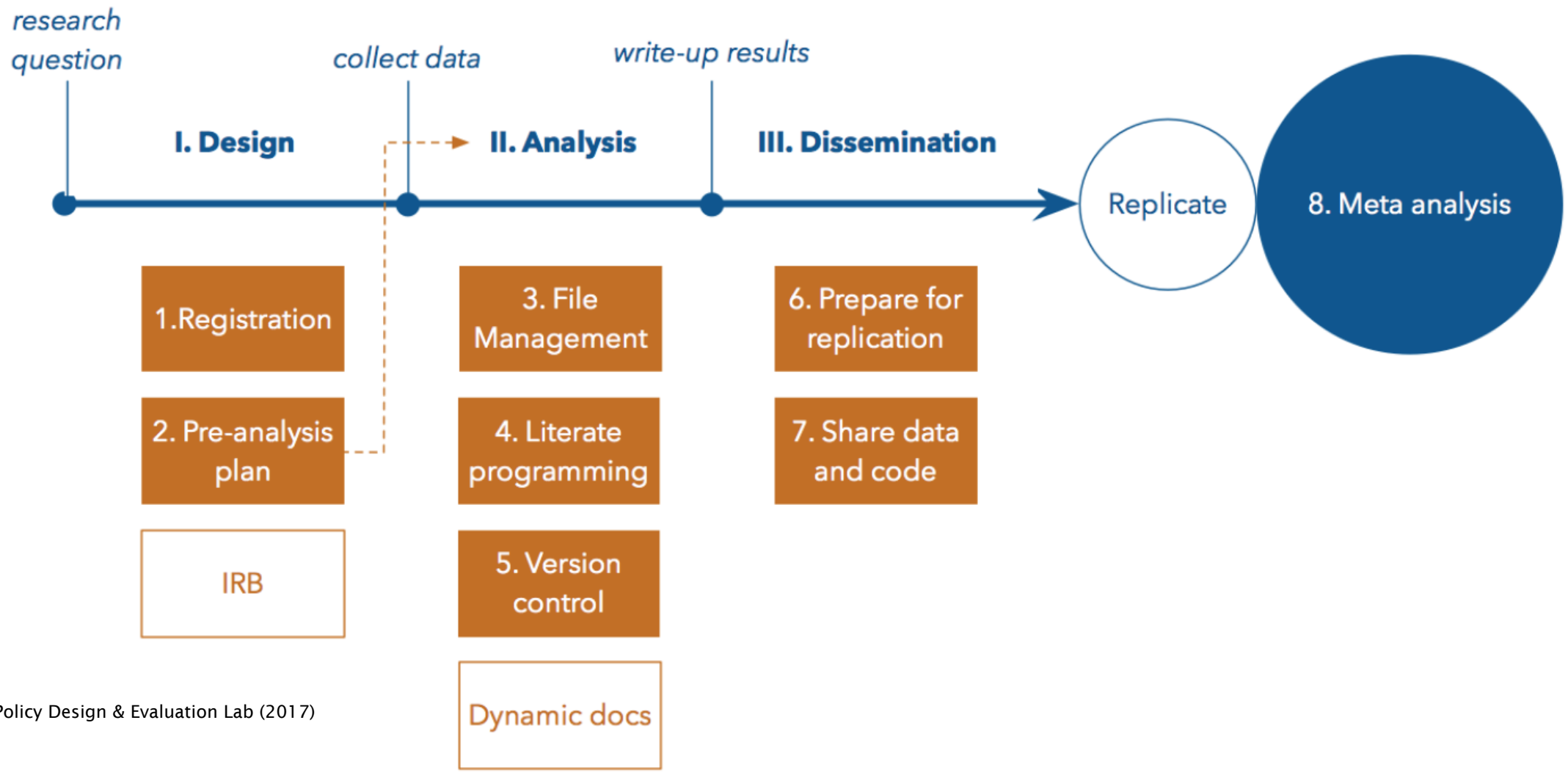
Summary points

Science is conducted by humans.

Many counternorms exist that undermine scientific integrity.

What can we do about it?

A reproducible path forward: Reminaging the research lifecycle?



Policy Design & Evaluation Lab (2017)

Break! ☕

10:00