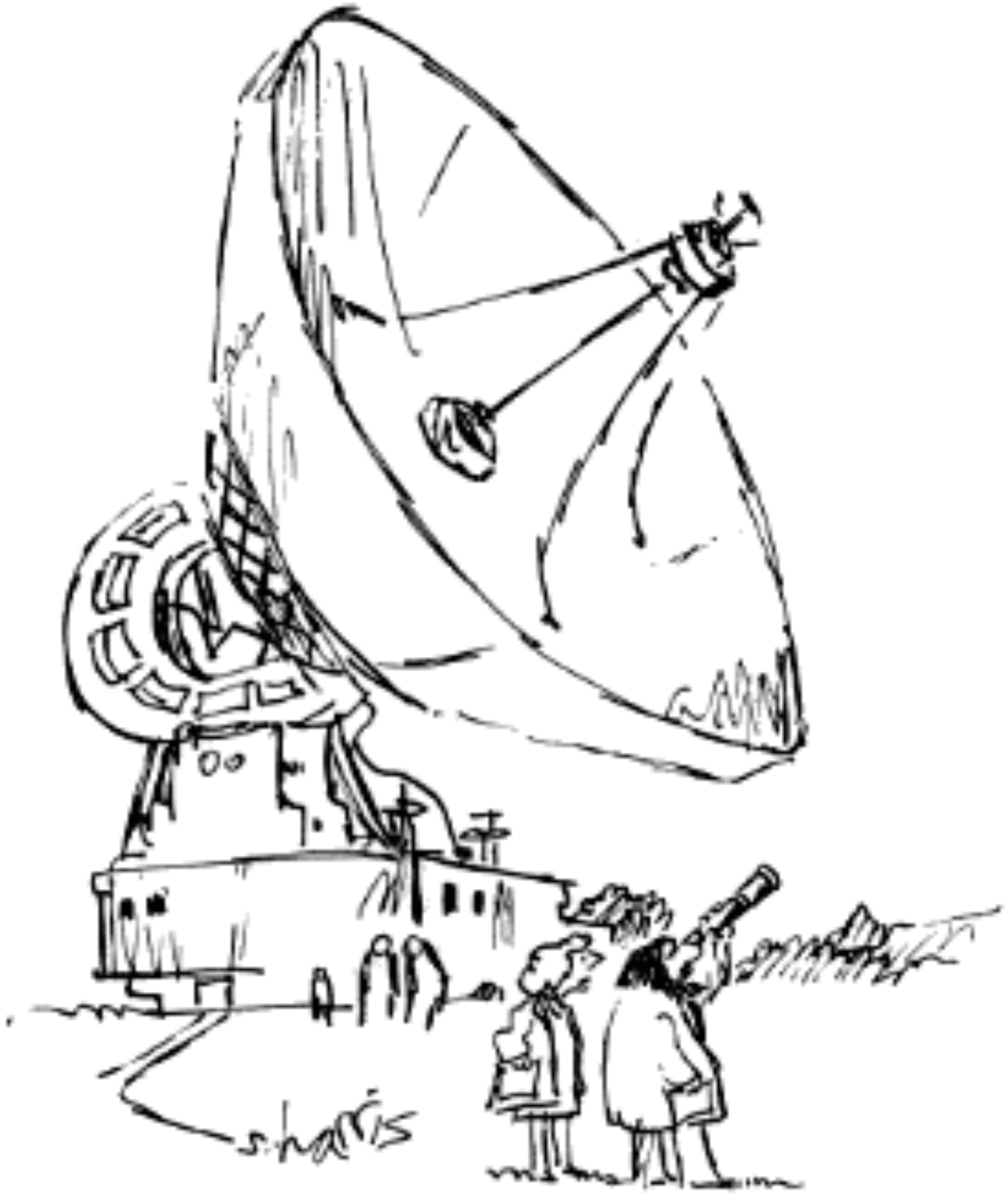


"Just checking."

# Introduction to the Reproducibility Crisis

**David Jensen**
College of Information & Computer Sciences
University of Massachusetts Amherst

*UMass Medical School & UMass Boston*
*8 September 2021*

# What is "The Reproducibility Crisis"?

"…an ongoing methodological crisis
in which it has been found that many scientific studies
are difficult or impossible to replicate or reproduce.
The replication crisis most severely affects
the social sciences and medicine,
while survey data strongly indicates that
all of the natural sciences
are probably implicated as well."

*Also called the "replication crisis", "replicability crisis", or "decline effect"*

*Source: Wikipedia, "Replication crisis"*

# Reproducibility is a persistent concern within all sciences…

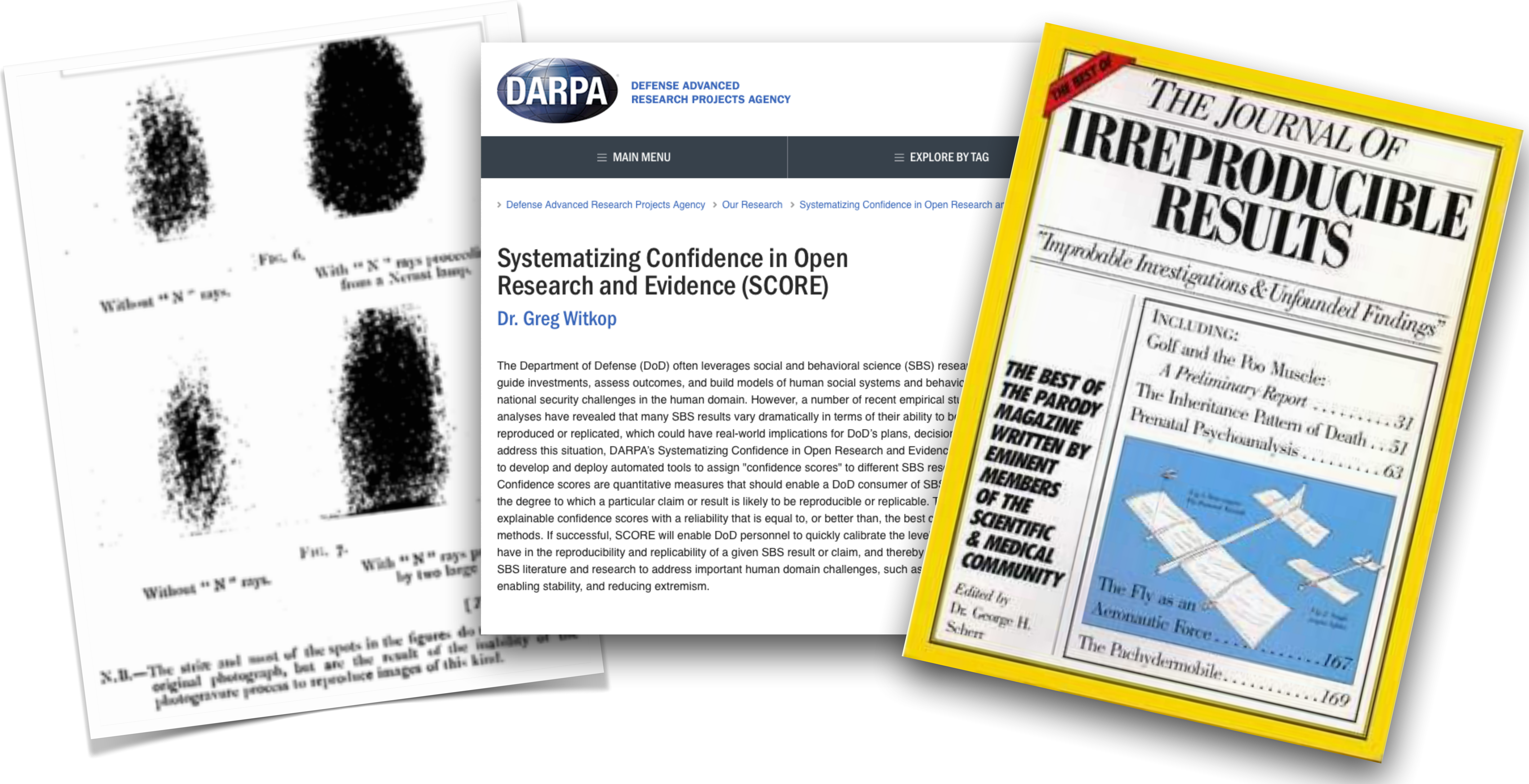


FIG. 6.

Without "N" rays.

With "N" rays proceeding from a Nernst lamp.

FIG. 7.

Without "N" rays.

With "N" rays p
by two large

N.B.—The striæ and most of the spots in the figures do
original photograph, but are the result of the inability of the
photogravure process to reproduce images of this kind.

DARPA DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

≡ MAIN MENU
≡ EXPLORE BY TAG
› Defense Advanced Research Projects Agency › Our Research › Systematizing Confidence in Open Research ar

Systematizing Confidence in Open Research and Evidence (SCORE)

Dr. Greg Witkop

The Department of Defense (DoD) often leverages social and behavioral science (SBS) resea
guide investments, assess outcomes, and build models of human social systems and behavio
national security challenges in the human domain. However, a number of recent empirical st
analyses have revealed that many SBS results vary dramatically in terms of their ability to b
reproduced or replicated, which could have real-world implications for DoD's plans, decisio
address this situation, DARPA's Systematizing Confidence in Open Research and Evidenc
to develop and deploy automated tools to assign "confidence scores" to different SBS res
Confidence scores are quantitative measures that should enable a DoD consumer of SBS
the degree to which a particular claim or result is likely to be reproducible or replicable. T
explainable confidence scores with a reliability that is equal to, or better than, the best c
methods. If successful, SCORE will enable DoD personnel to quickly calibrate the leve
have in the reproducibility and replicability of a given SBS result or claim, and thereby
SBS literature and research to address important human domain challenges, such as
enabling stability, and reducing extremism.

THE BEST OF

THE JOURNAL OF IRREPRODUCIBLE RESULTS

"Improbable Investigations & Unfounded Findings"

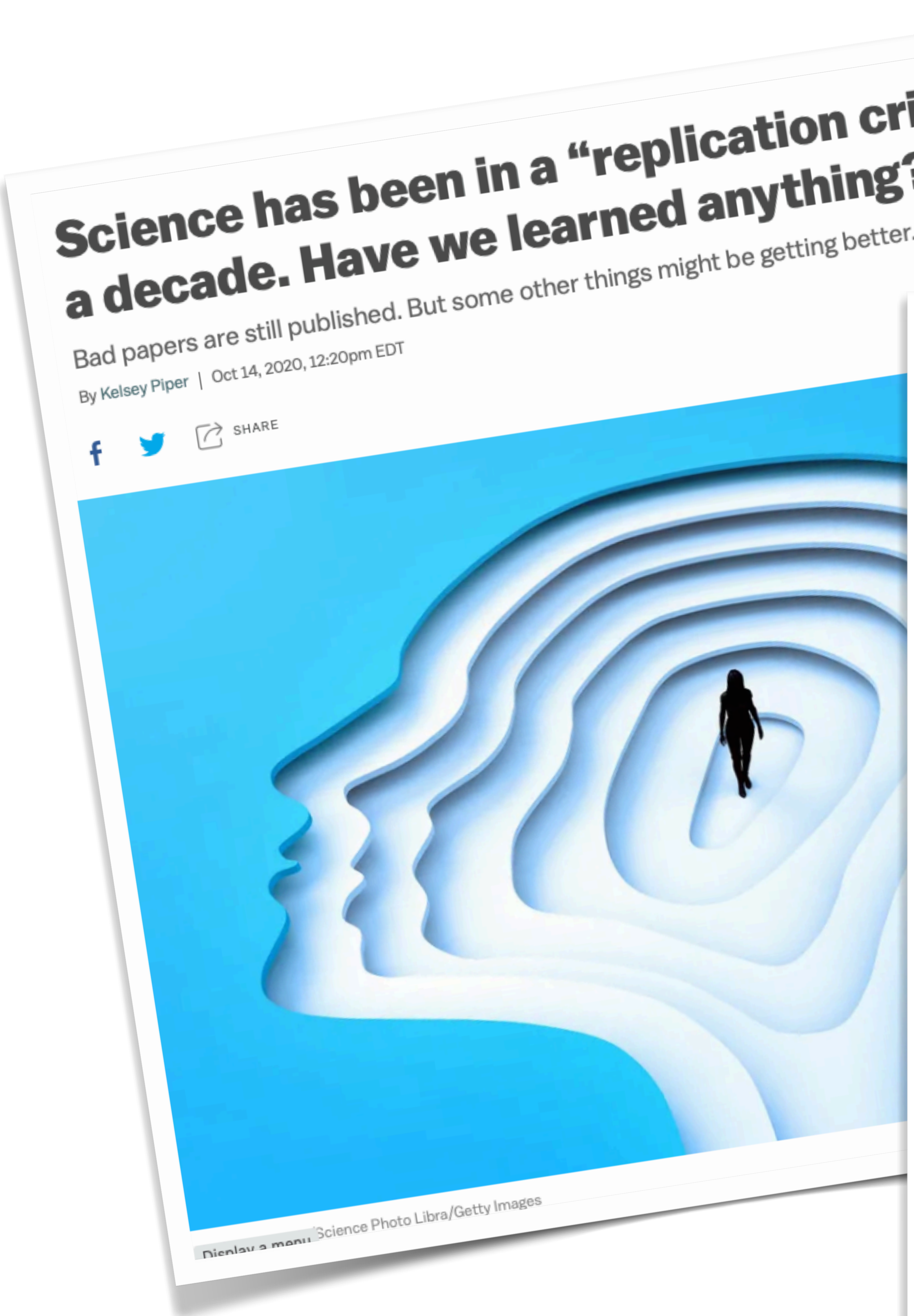
INCLUDING:
Golf and the Poo Muscle: A Preliminary Report ..........31
The Inheritance Pattern of Death ..51
Prenatal Psychoanalysis ..........63
The Fly as an Aeronautic Force ..........167
The Pachydermobile ..........169

THE BEST OF THE PARODY MAGAZINE WRITTEN BY EMINENT MEMBERS OF THE SCIENTIFIC & MEDICAL COMMUNITY

Edited by Dr. George H. Scherr

*Source: Wikipedia, "N-Ray"; JoIR*

# …but it has also become a highly public issue for science



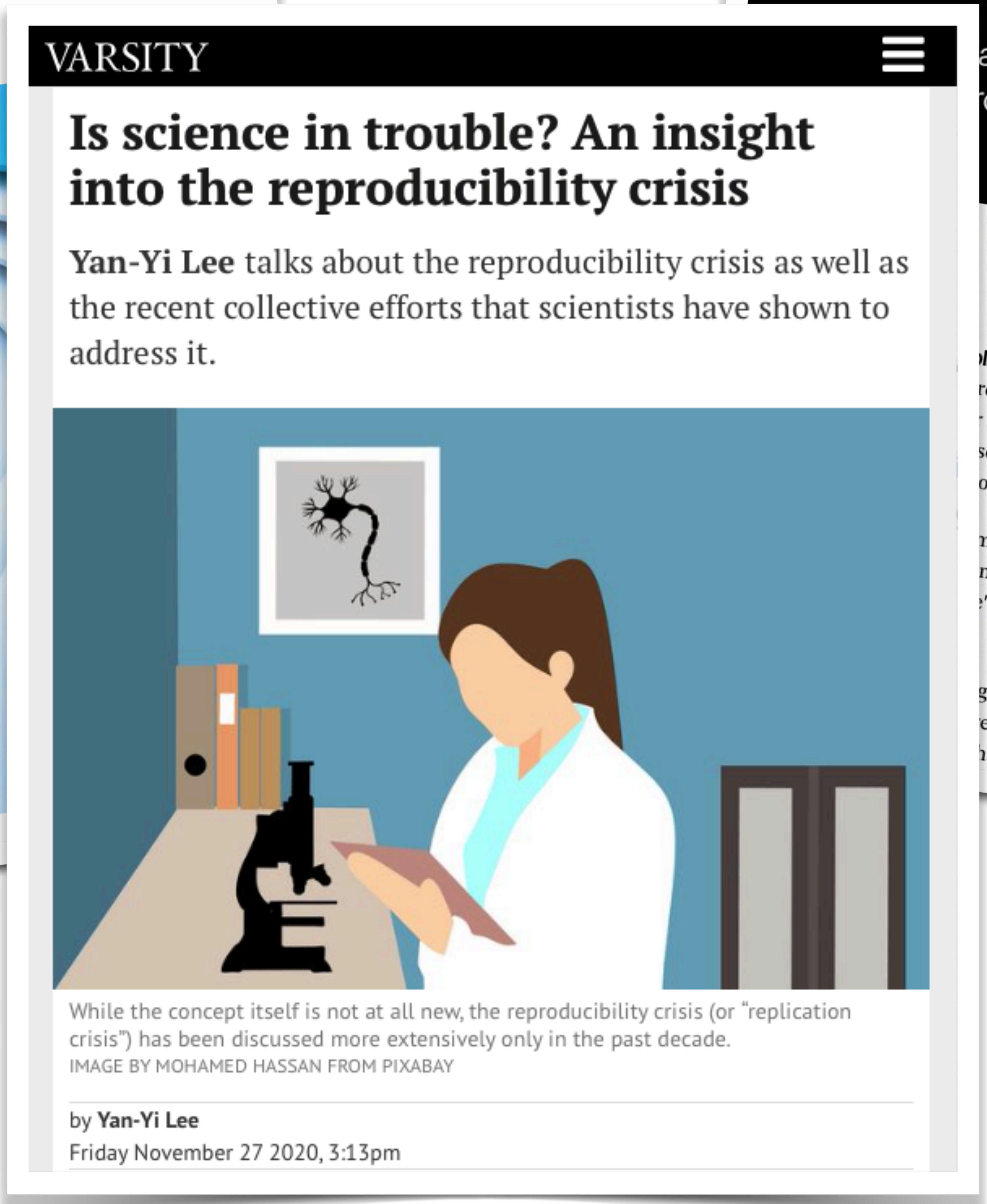
Science has been in a "replication crisis" for a decade. Have we learned anything?

Bad papers are still published. But some other things might be getting better.

By Kelsey Piper | Oct 14, 2020, 12:20pm EDT

SHARE

Display a menu Science Photo Libra/Getty Images



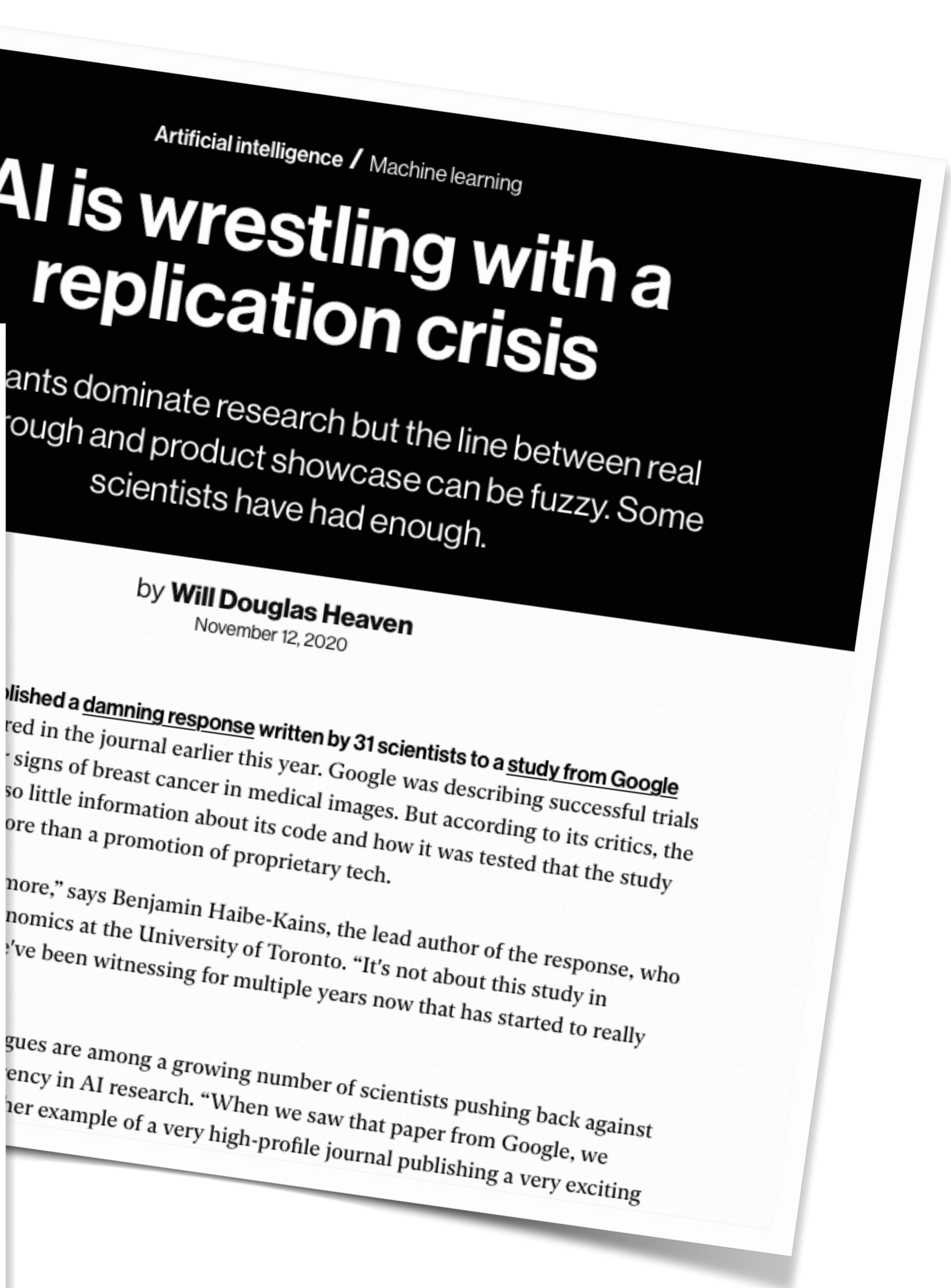
VARSITY

## Is science in trouble? An insight into the reproducibility crisis

**Yan-Yi Lee** talks about the reproducibility crisis as well as the recent collective efforts that scientists have shown to address it.

While the concept itself is not at all new, the reproducibility crisis (or "replication crisis") has been discussed more extensively only in the past decade.
IMAGE BY MOHAMED HASSAN FROM PIXABAY

by **Yan-Yi Lee**
Friday November 27 2020, 3:13pm



Artificial intelligence / Machine learning

## AI is wrestling with a replication crisis

ants dominate research but the line between real
ough and product showcase can be fuzzy. Some
scientists have had enough.

by **Will Douglas Heaven**
November 12, 2020

lished a damning response written by 31 scientists to a study from Google
red in the journal earlier this year. Google was describing successful trials
signs of breast cancer in medical images. But according to its critics, the
so little information about its code and how it was tested that the study
ore than a promotion of proprietary tech.

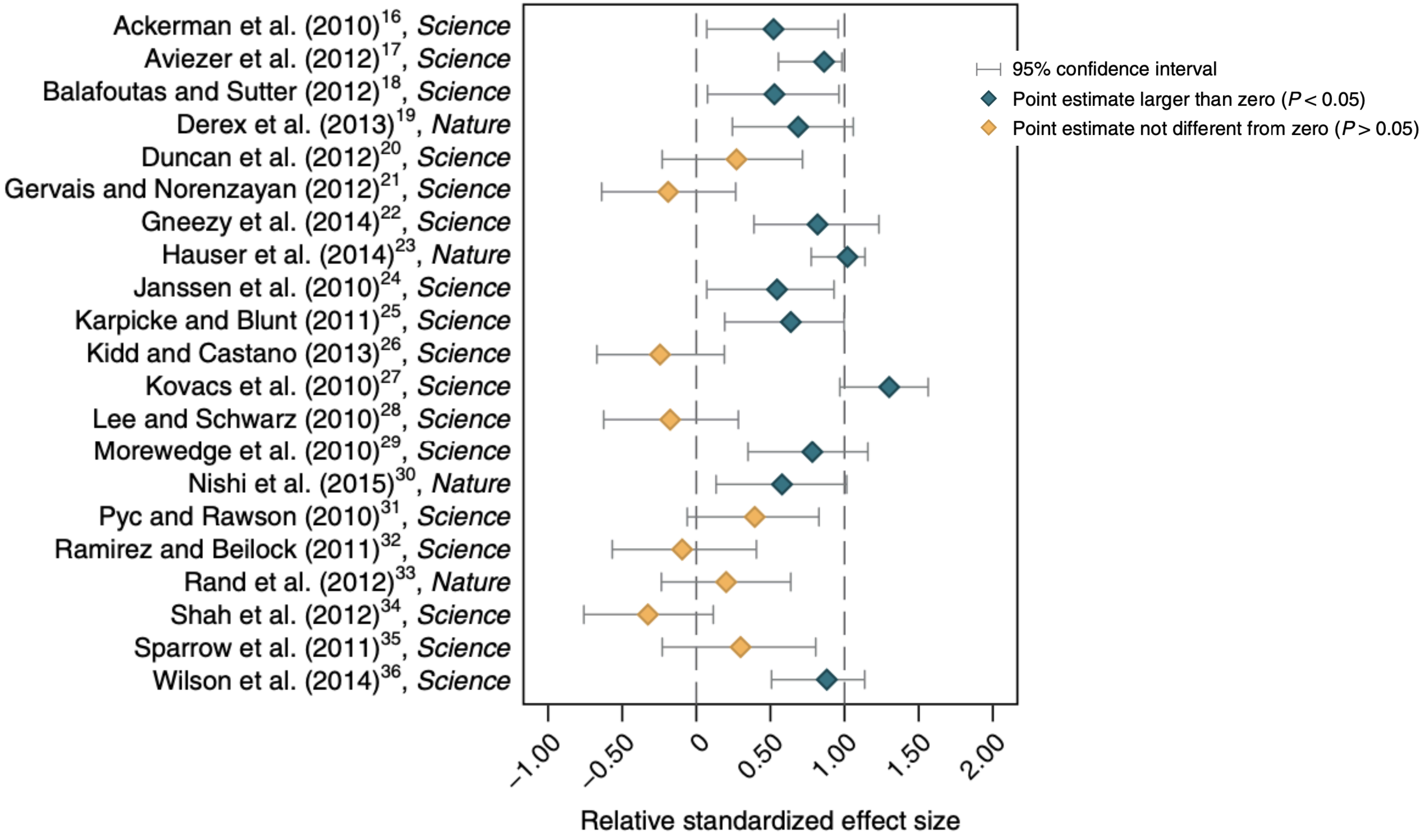
nore," says Benjamin Haibe-Kains, the lead author of the response, who
nomics at the University of Toronto. "It's not about this study in
've been witnessing for multiple years now that has started to really

gues are among a growing number of scientists pushing back against
ency in AI research. "When we saw that paper from Google, we
her example of a very high-profile journal publishing a very exciting

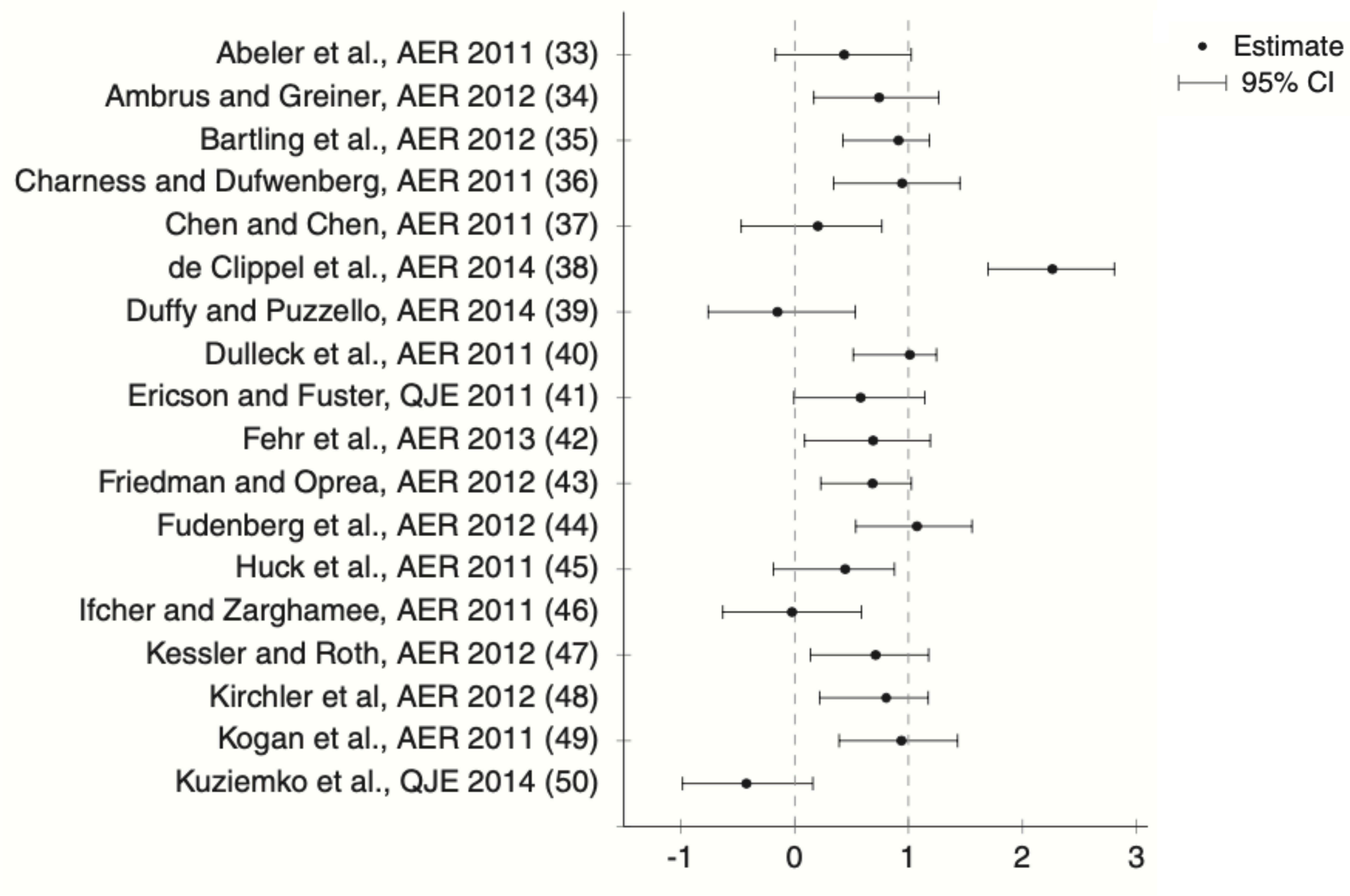
# One question appears relatively simple…

***Do a substantial proportion of published studies fail to replicate?***

- There is good evidence that the answer is “yes”.
- “Out of 49 medical studies from 1990–2003 with more than 1000 citations, 45 claimed that the studied therapy was effective. Out of these studies, 16% were contradicted by subsequent studies, 16% had found stronger effects than did subsequent studies, 44% were replicated, and 24% remained largely unchallenged.” (Ioannidis 2005)
- Only 67% of social science studies in *Nature* and *Science* between 2010 and 2015 replicated (Camerer et al. 2018)
- Only 61% of a set of studies published in the *American Economic Review* and the *Quarterly Journal of Economics* replicated (Camerer et al. 2016).



*Source: Camerer et al. (2018)*



*Source: Camerer et al. (2016)*

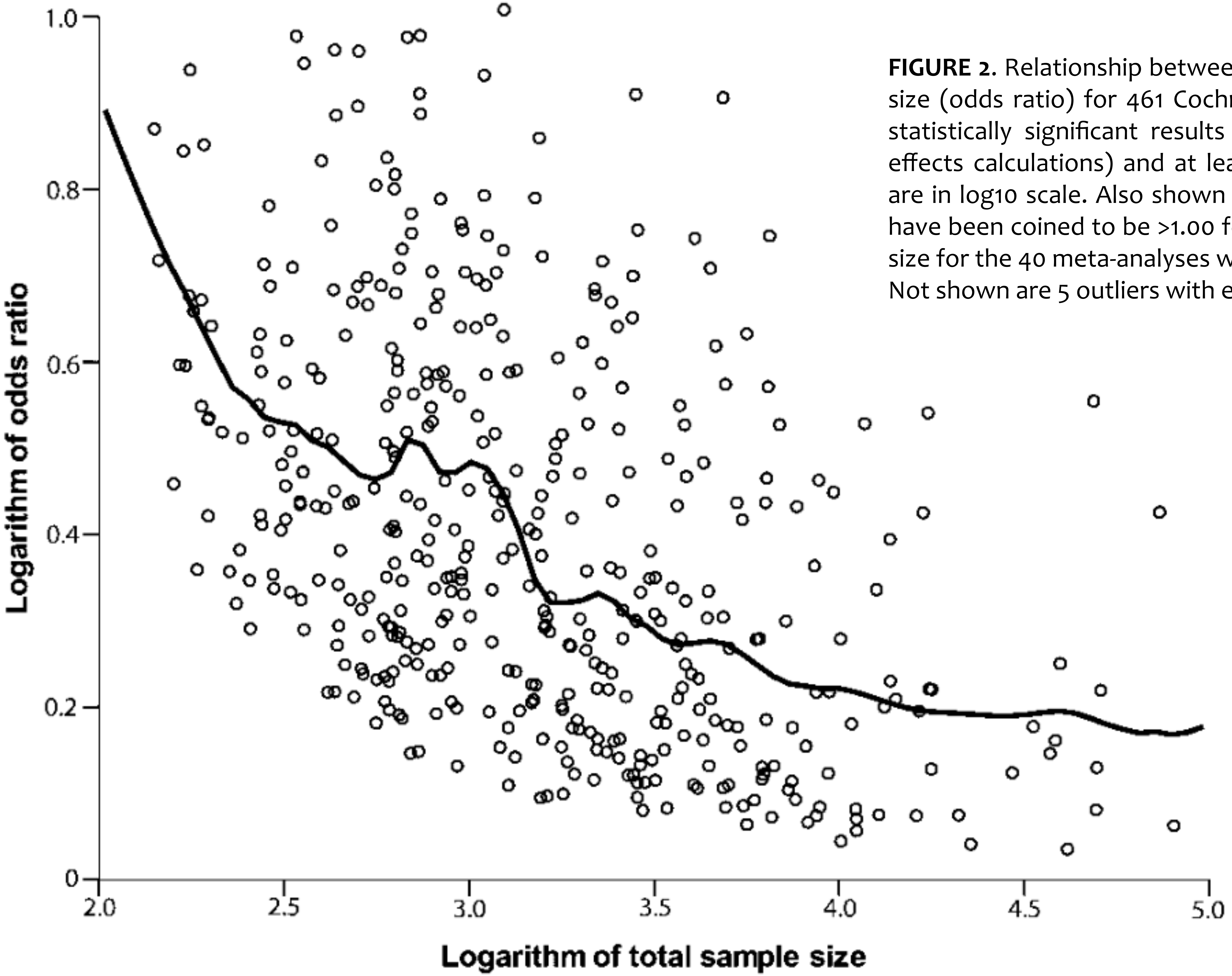


**FIGURE 2.** Relationship between total sample size and the effect size (odds ratio) for 461 Cochrane meta-analyses with formally statistically significant results (P < 0.05 according to random effects calculations) and at least 4 included studies. Both axes are in log10 scale. Also shown is a fit LOESS line. All odds ratios have been coined to be >1.00 for consistency. The median effect size for the 40 meta-analyses with at least 10,000 subjects is 1.53. Not shown are 5 outliers with extreme sample size or effect size.

*Source: Ioannidis (2008)*

# A more interesting question would be…

***Has the proportion of studies that fail to replicate increased or decreased in the past several decades?***

- To my knowledge, we don't have good evidence on this question.
- The interpretation of the answer would also depend on whether we believe that research questions have become easier or more difficult and whether the underlying technologies for research have improved.
- This is being called a "crisis", which implies urgency and recency, but we don't appear to have evidence for this.

# One interesting piece of evidence: Power is not increasing

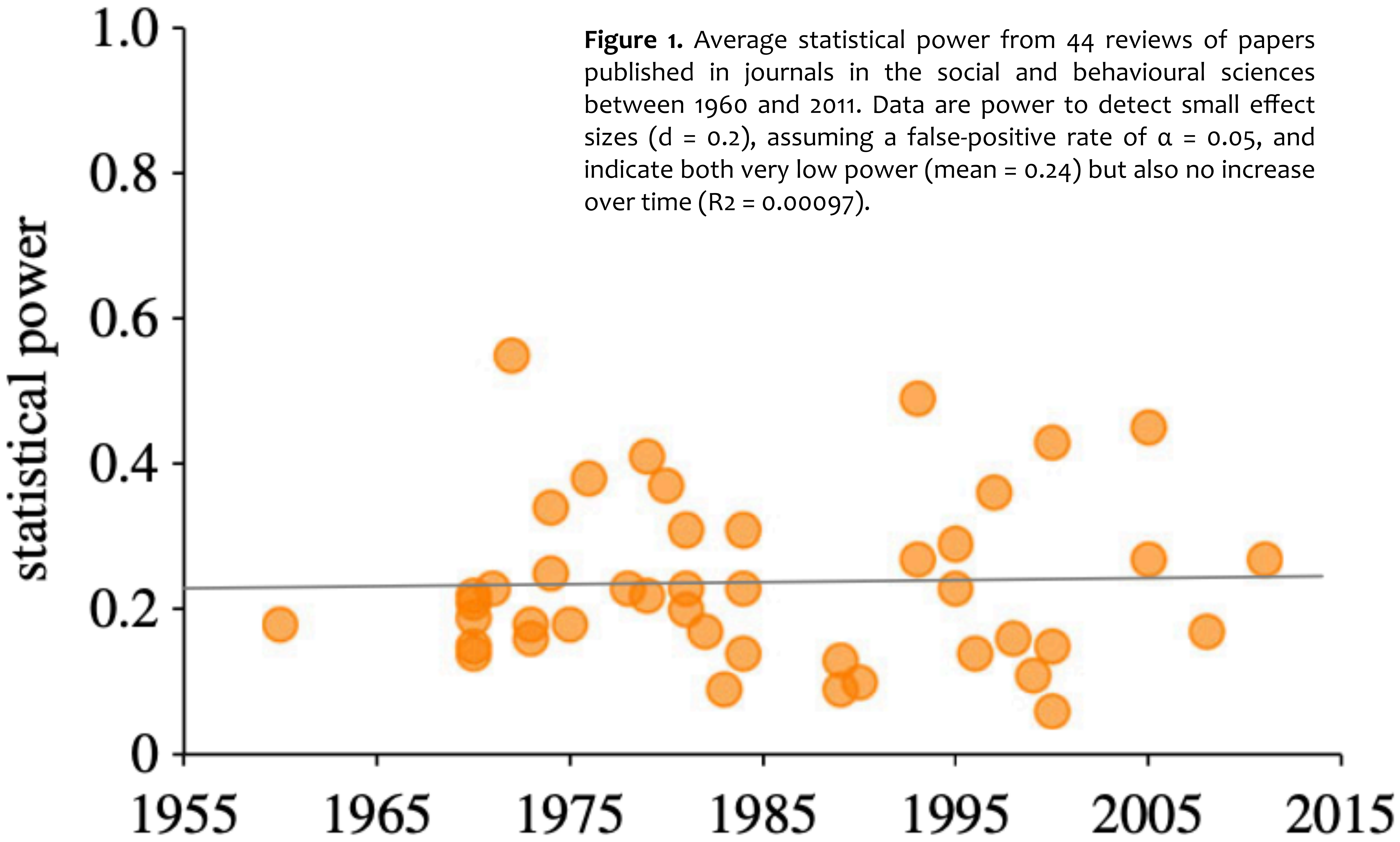


**Figure 1.** Average statistical power from 44 reviews of papers published in journals in the social and behavioural sciences between 1960 and 2011. Data are power to detect small effect sizes (d = 0.2), assuming a false-positive rate of α = 0.05, and indicate both very low power (mean = 0.24) but also no increase over time (R2 = 0.00097).

*Source: Smaldino and McElreath (2016)*

# What is causing the *perception* of a reproducibility crisis?

What is causing the perceived increase
in the number and frequency of cases
in which published results fail to replicate?

Specifically, has something changed
about the ***quality*** of individual published results,
or has something changed about the ***context***
in which those studies are published and reported?

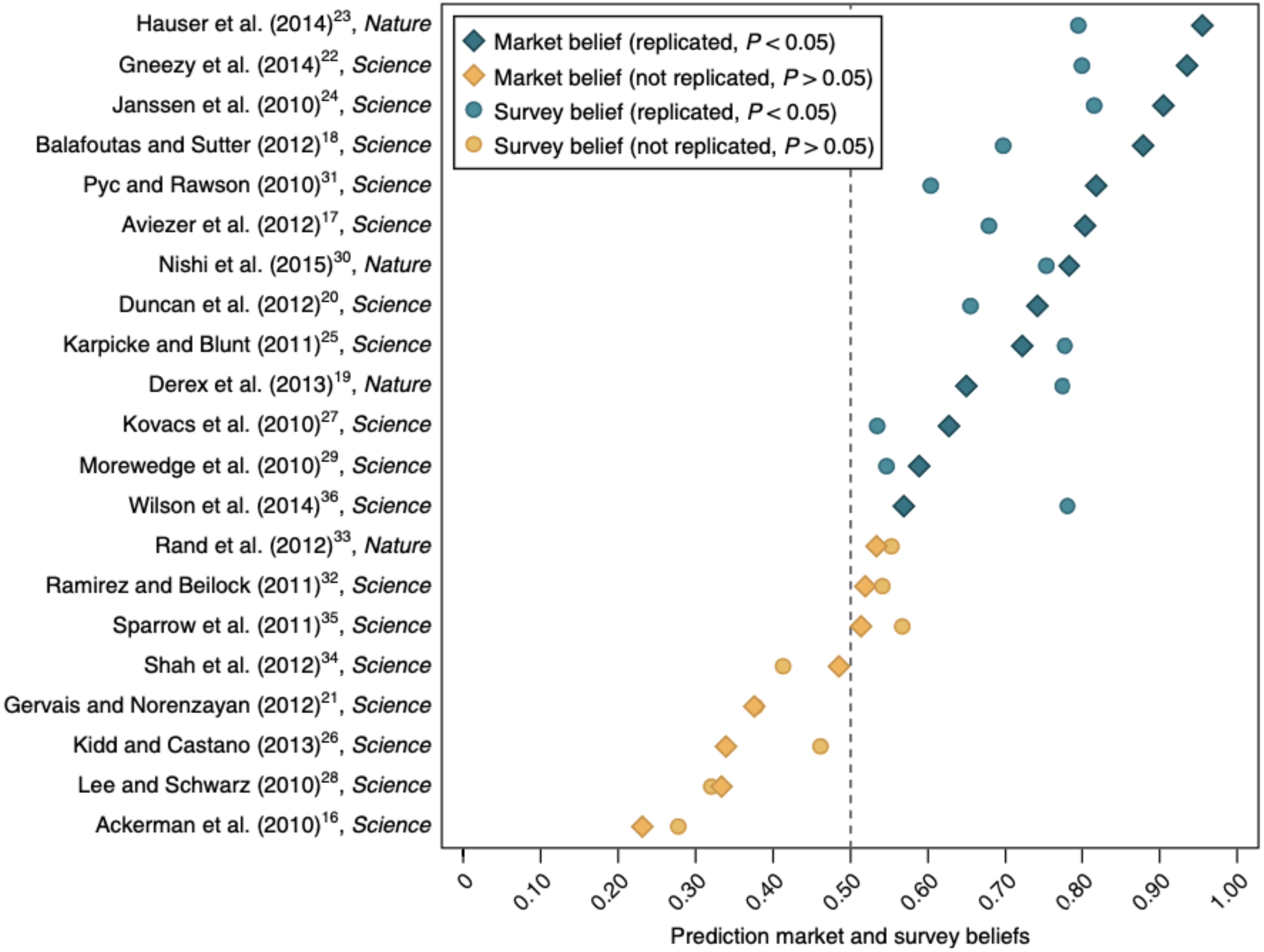
# Several contemporary trends have raised concerns about *quality*

- Greater awareness about questionable research habits
  - “HARKing” — Hypothesizing after the results are known
  - *p*-hacking
  - “Garden of forking paths” (Gelman & Loken 2013) or “researcher degrees of freedom” (Simmons et al. 2011)
- Highly publicized instances of fraud

- Greater awareness of career pressures on young researchers
  - Paper counts
  - Citation counts and h-index
- Greater focus on media profile
  - “Science in the age of Selfies” (Geman & Gelman 2016)
  - Popular science news
  - Greater institutional focus on media and social media.

# However, there are also reasonable responses

- First, we shouldn’t expect most research to be of high quality.
  - To quote the philosopher Daniel Dennett (paraphrasing science fiction author Theodore Sturgeon): “90% of everything is crap. That is true, whether you are talking about physics, chemistry, evolutionary psychology, sociology, medicine —you name it—rock music, country western. 90% of everything is crap."
- At some level, “failure to replicate” is an inevitable part of research.
  We will never remove such failures entirely (and we wouldn’t want to).
- Finally, researchers within a field can often predict the extent to which results will replicate and which won’t…



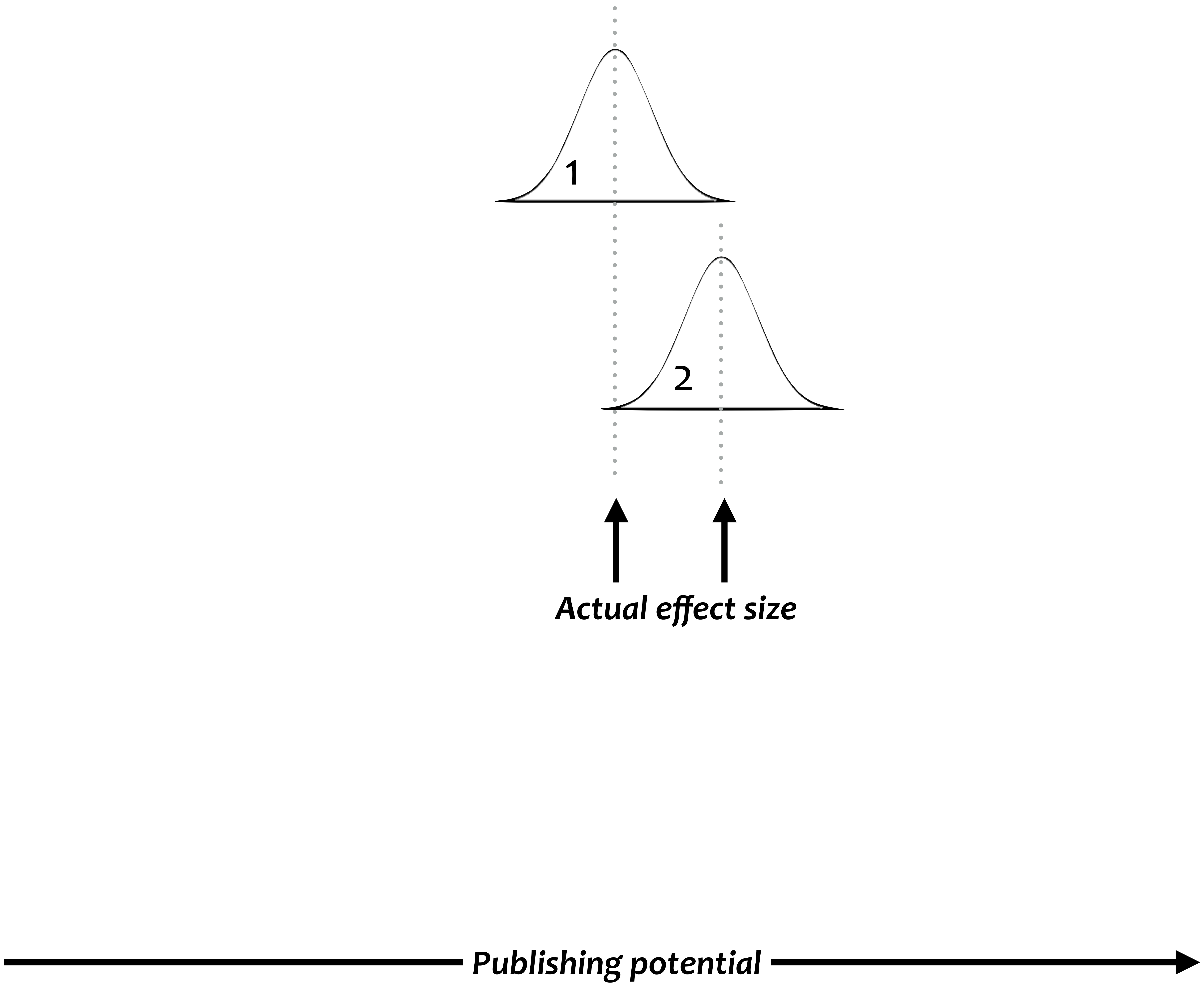
*Source: Camerer et al. (2018)*

Even if we do everything right,
some results aren’t going to replicate,
so we should structure the scientific system
so that high-quality research is recognized.
Unfortunately, many forces currently work
*against* that process.

# The structure of the scientific enterprise produces bias

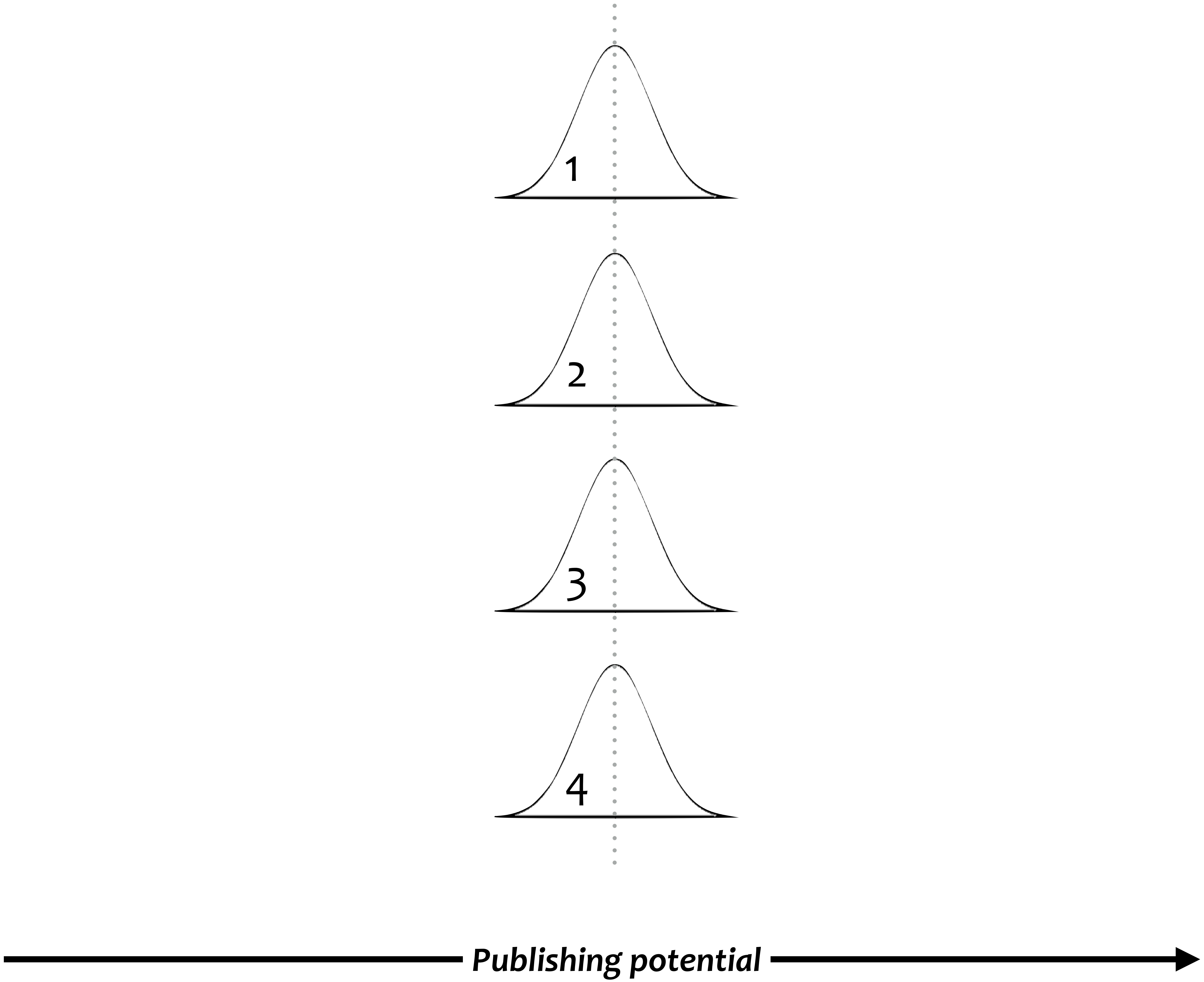
- Any system that…
  - Produces a large number of items
    *(e.g., large numbers of potential findings)*
  - Scores each item with some variance, and
    *(e.g., estimates of effect size)*
  - Selects the item with the maximum score
    *(e.g., publishes the most significant findings)*
- … will produce items with biased scores
  *(e.g., publish findings with inflated estimates of effect size)*

Jensen, D. and Cohen, P. (2000). Multiple comparisons in induction algorithms. *Machine Learning* 38(3):309-338.

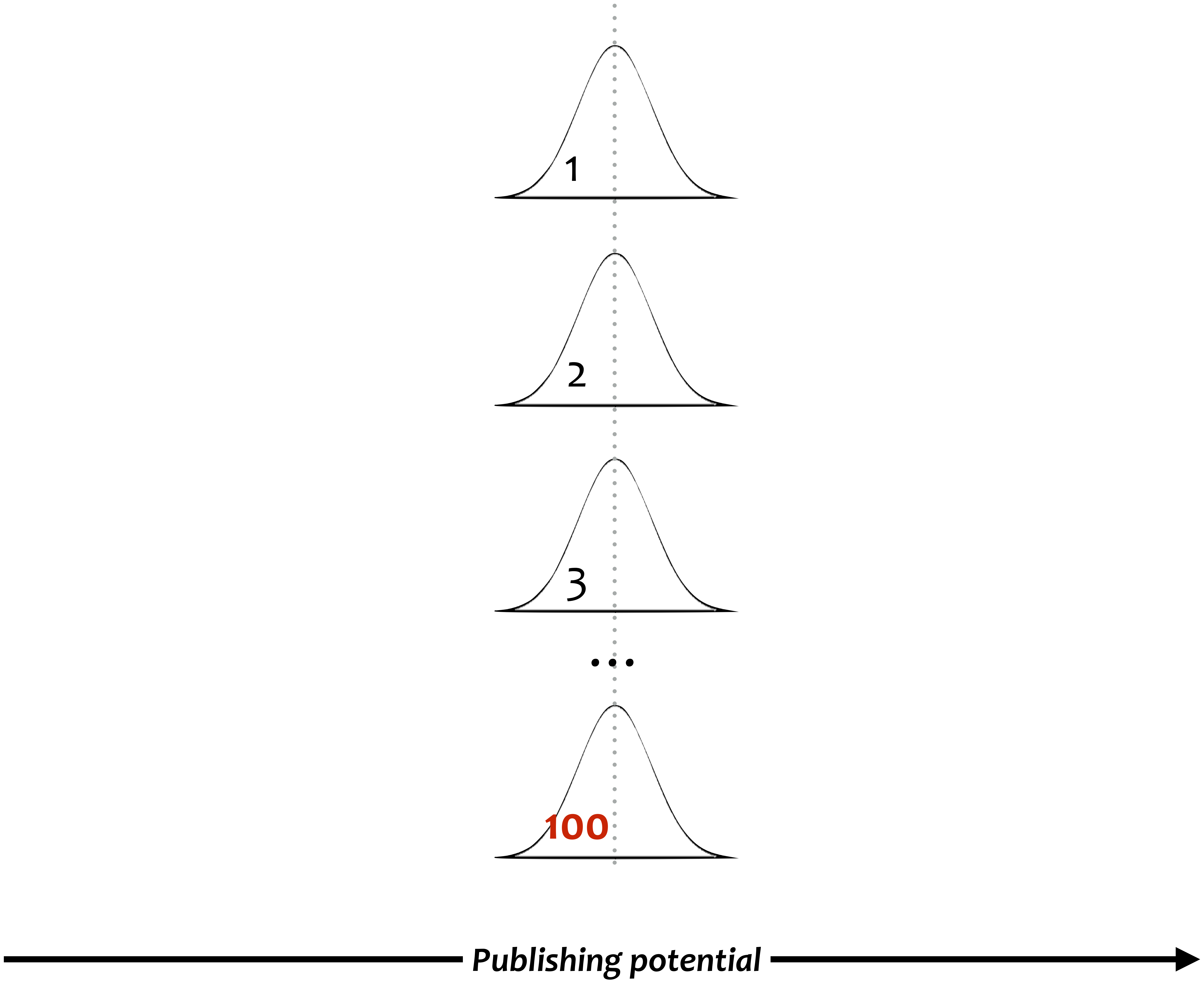


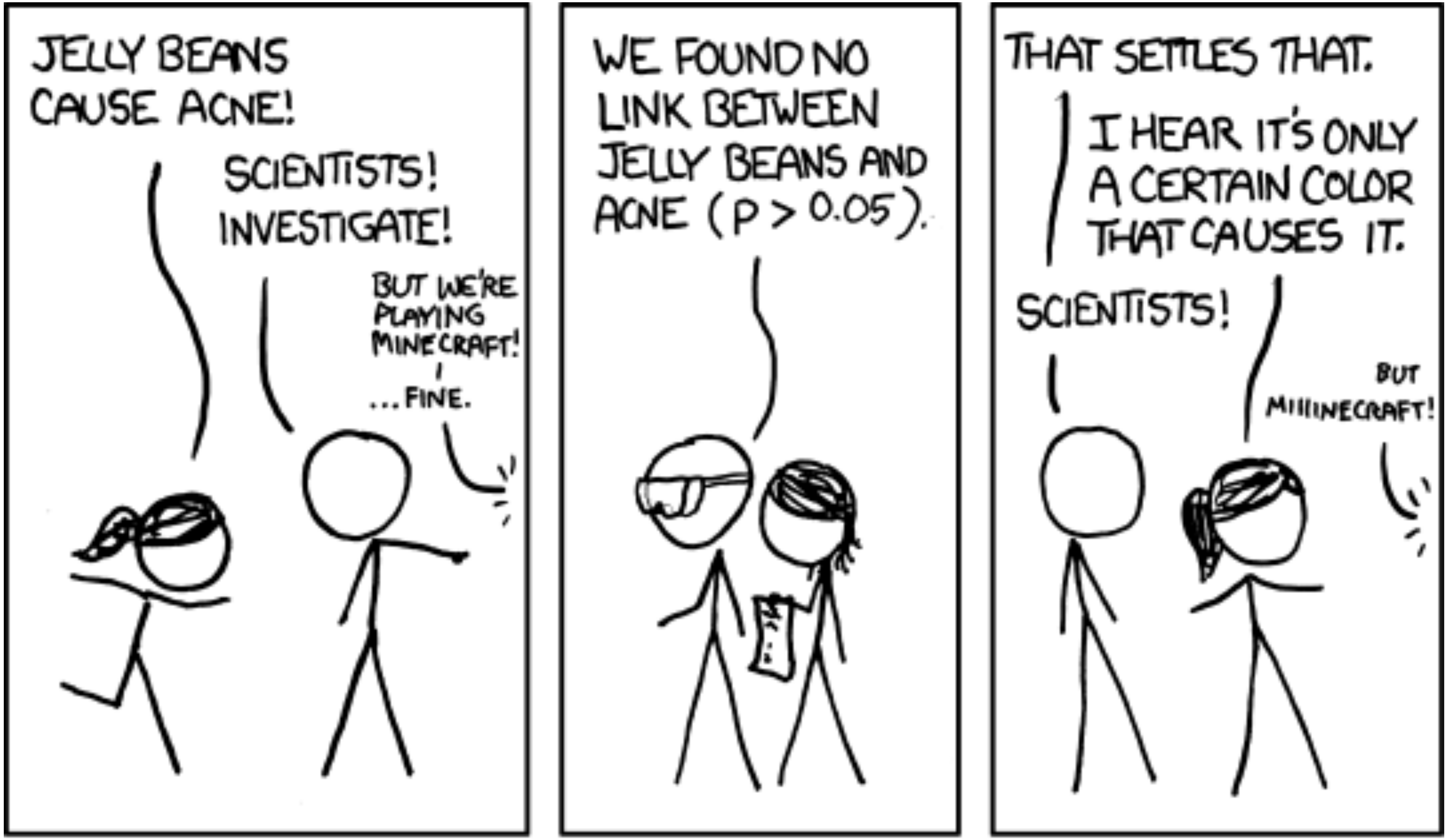
***Actual effect size***

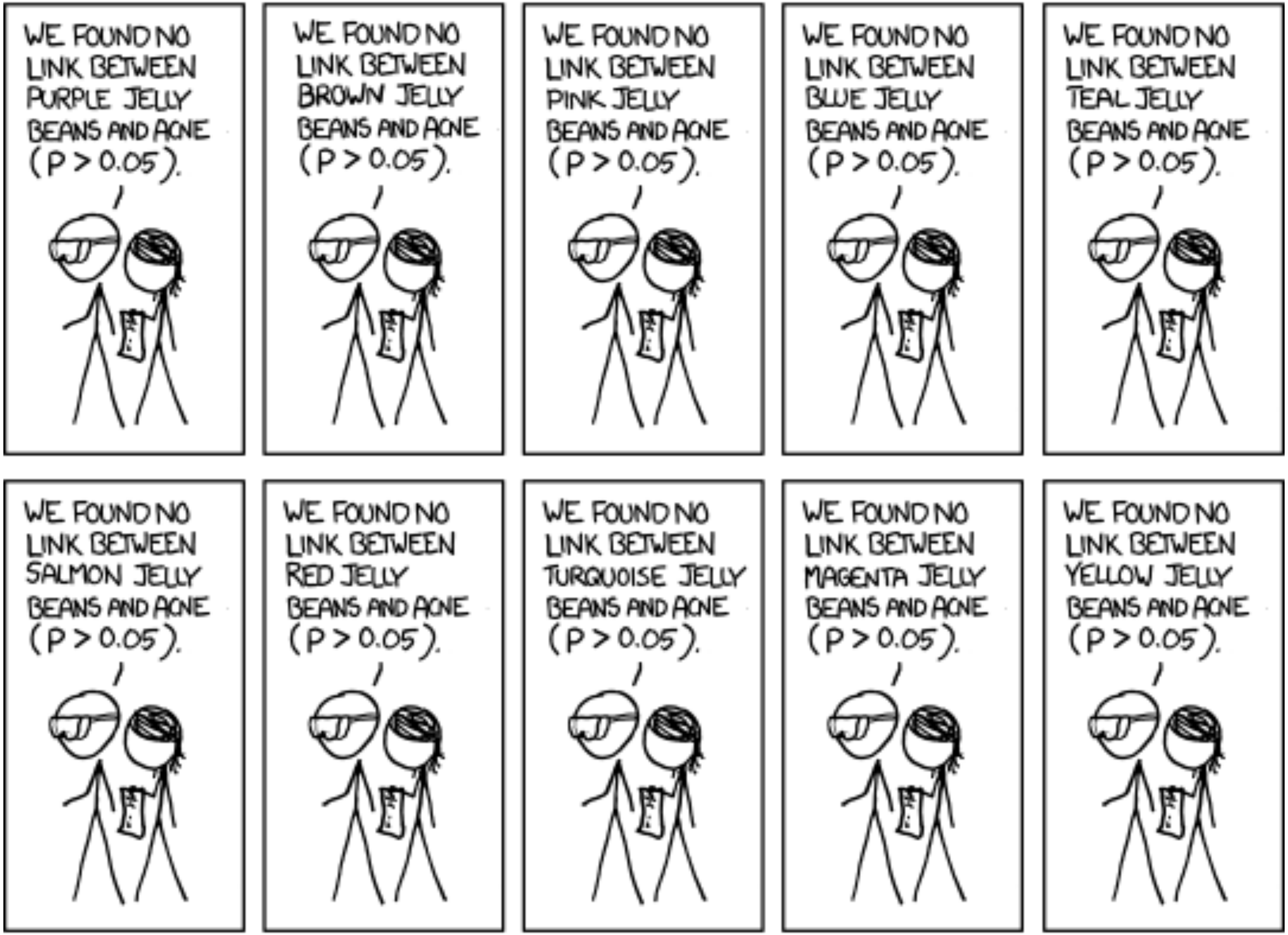
***Publishing potential***

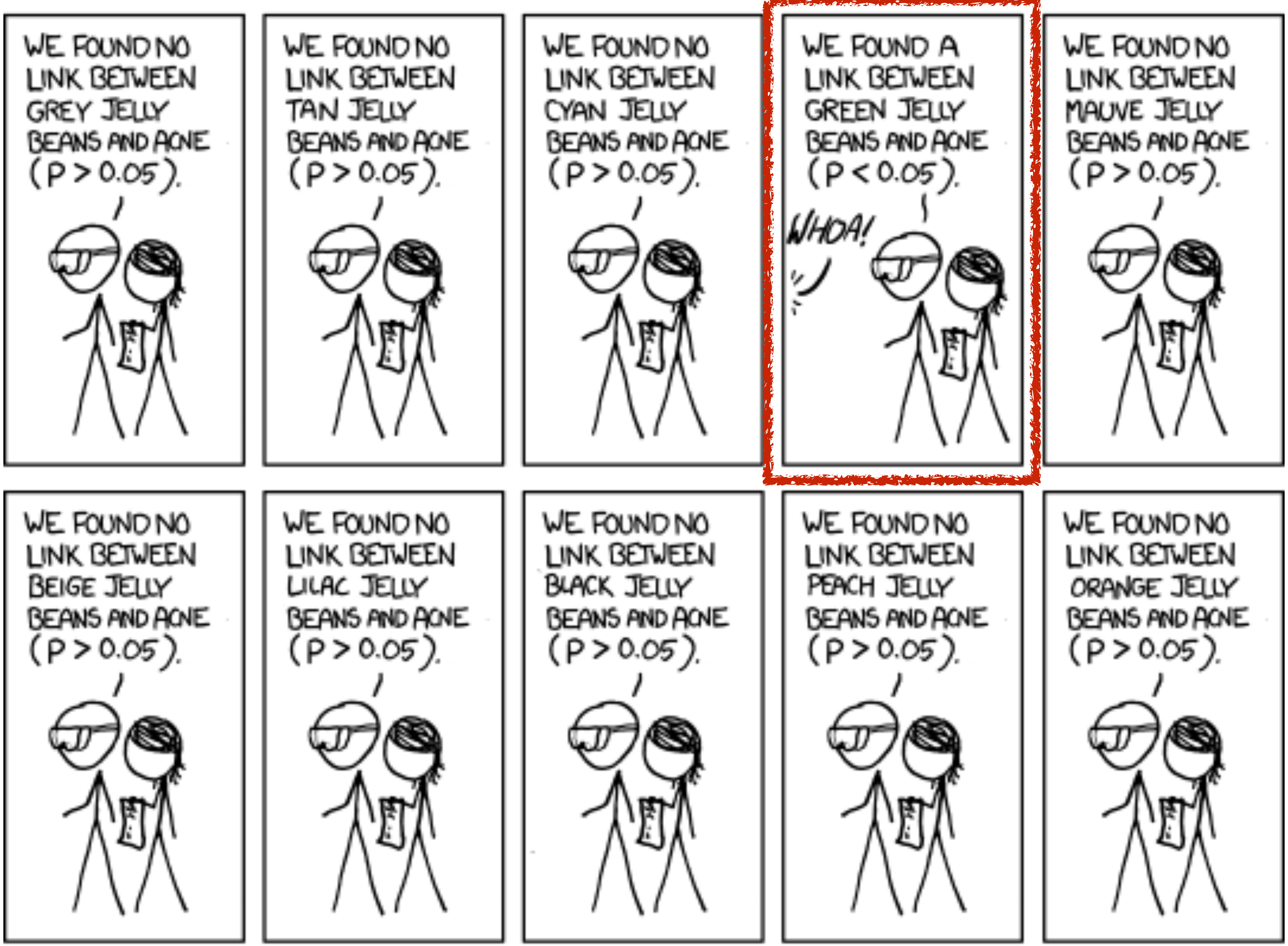


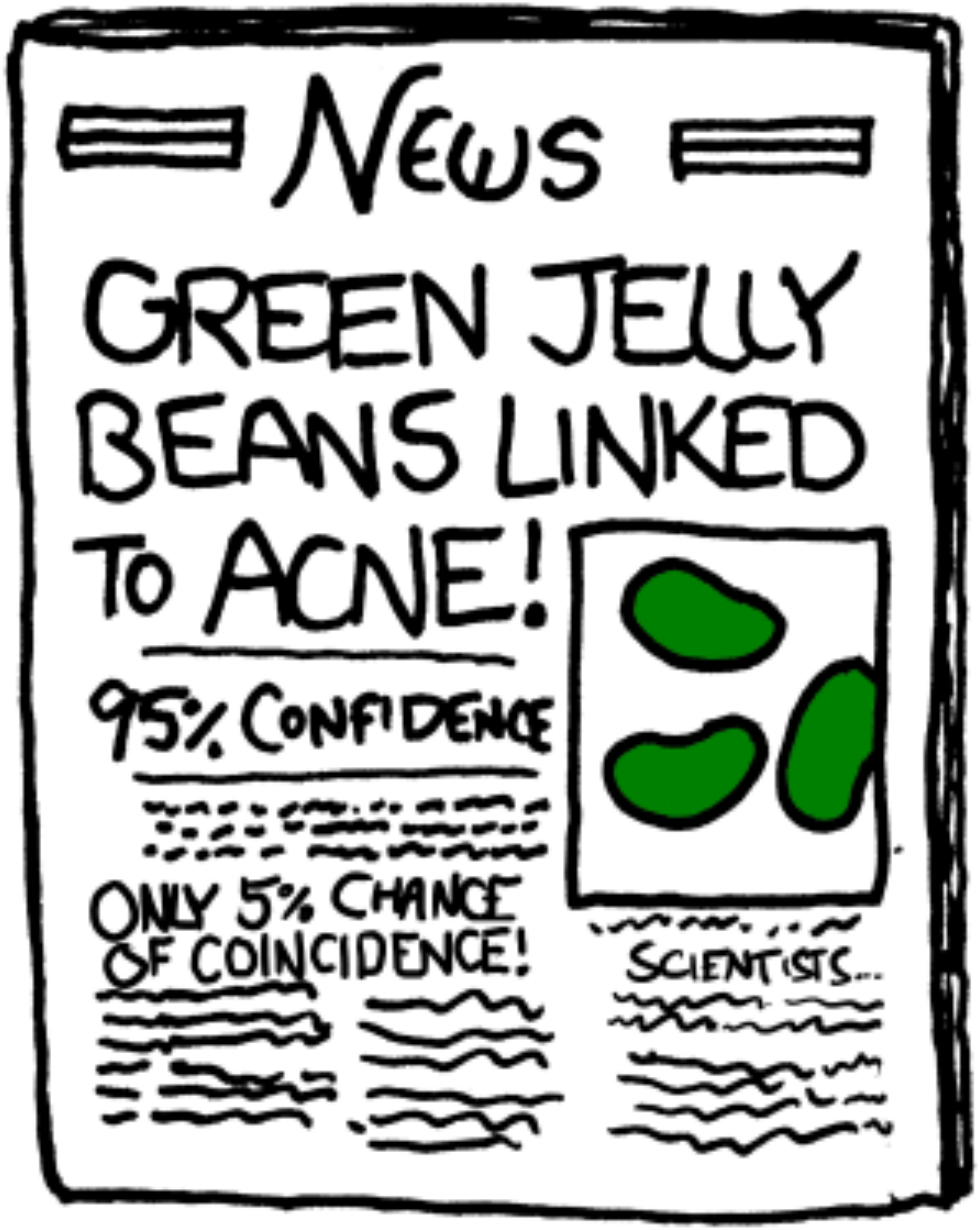
*Publishing potential*

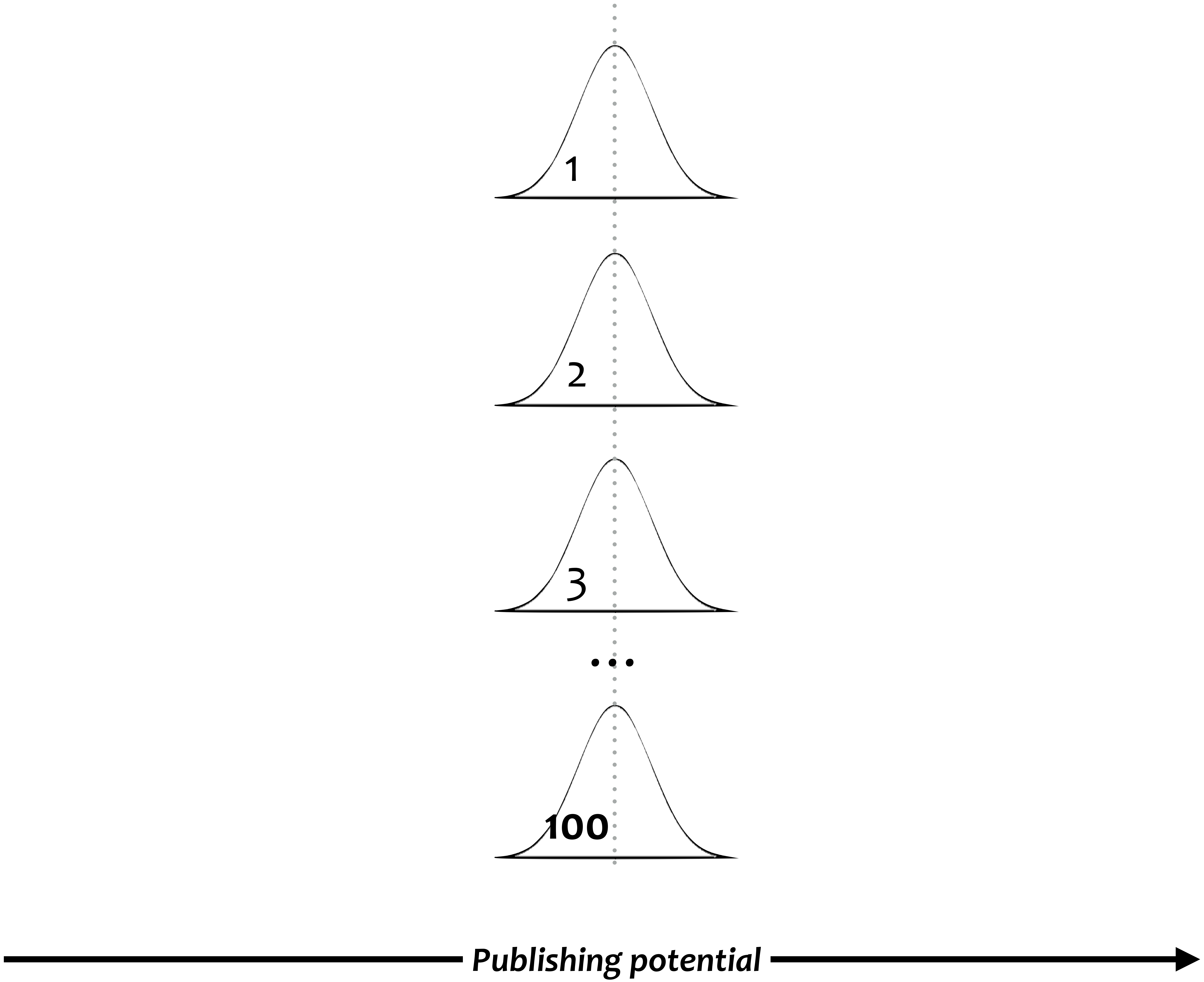

1
2
3
...
100
Publishing potential

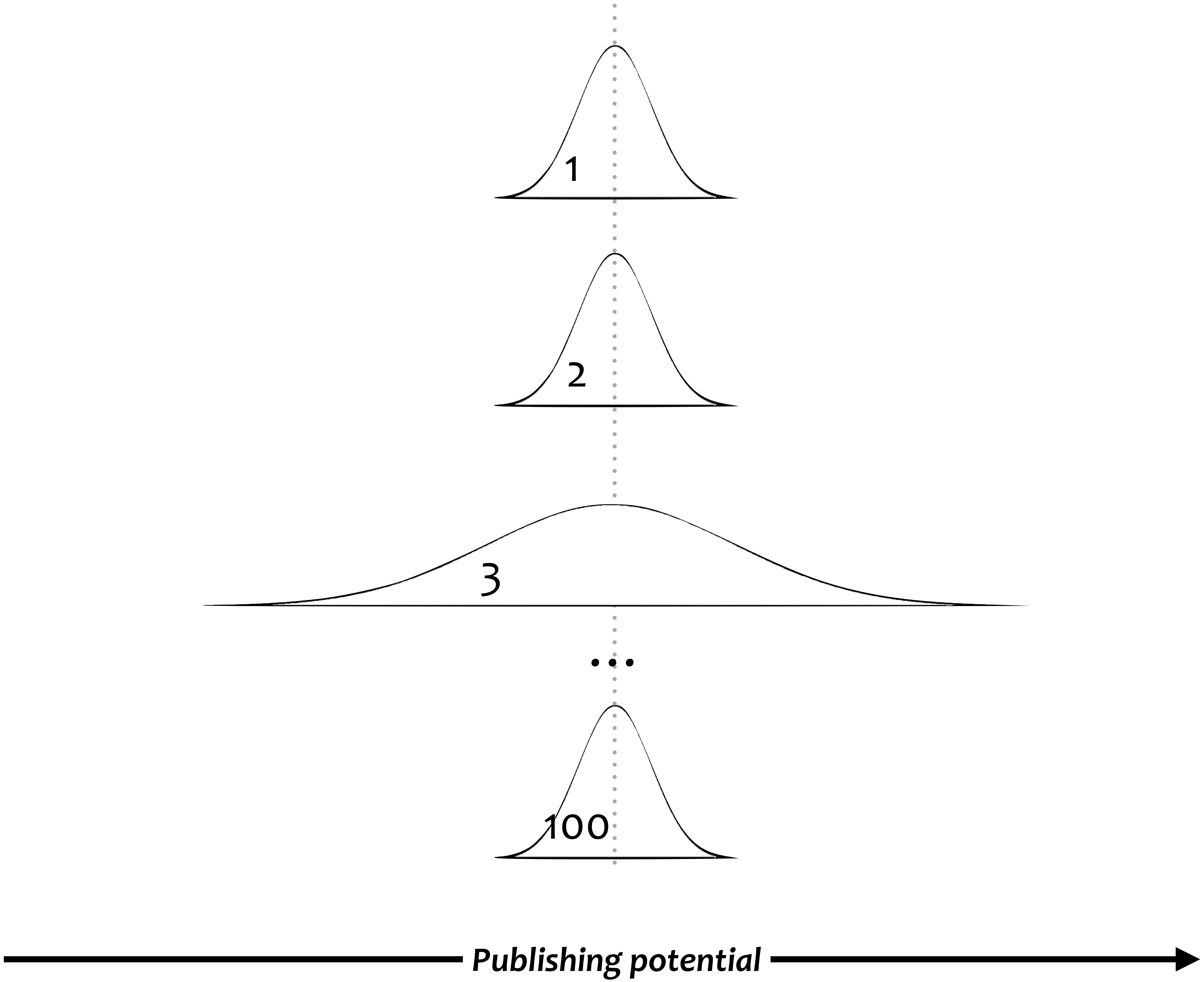

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

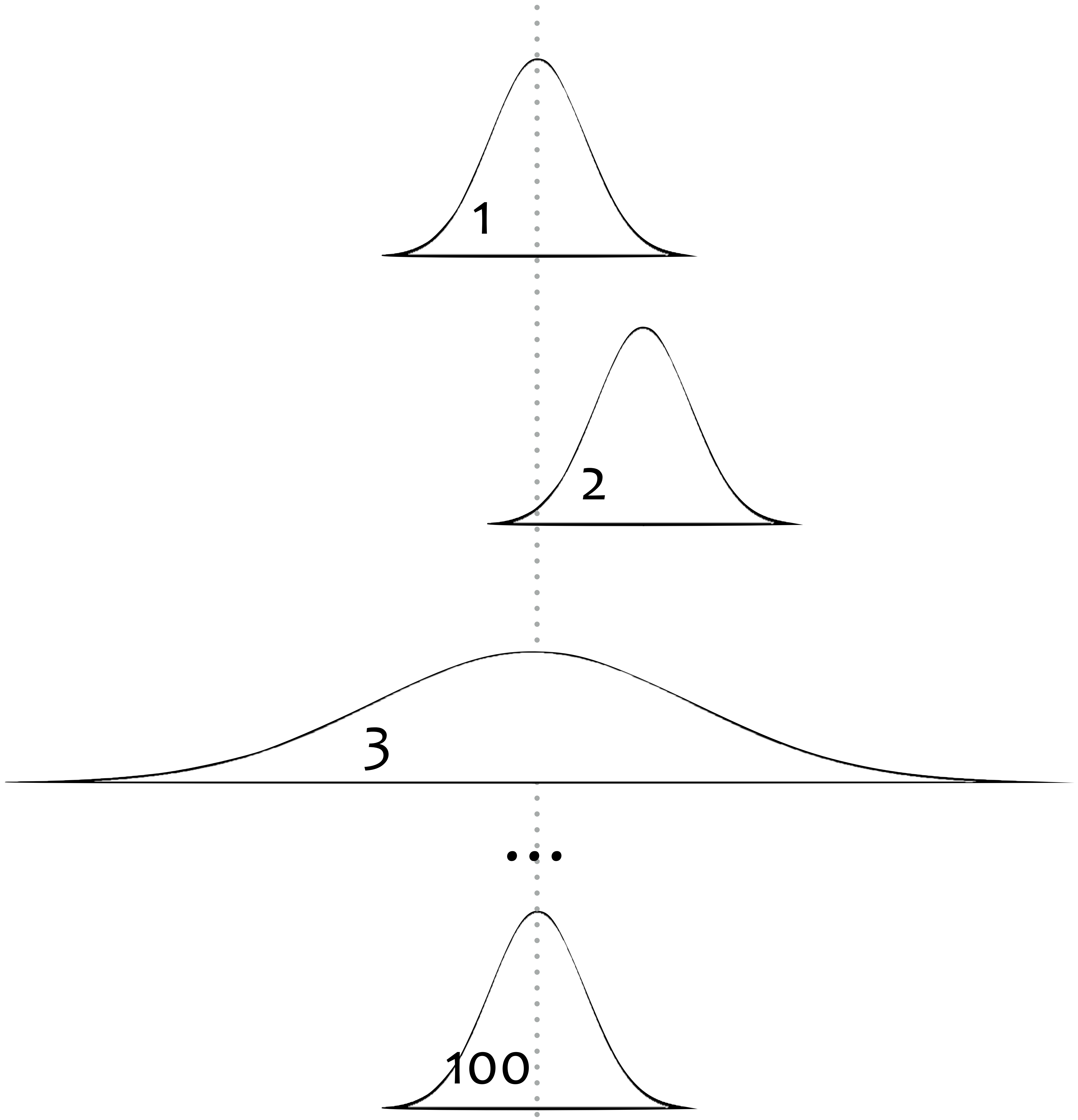

WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

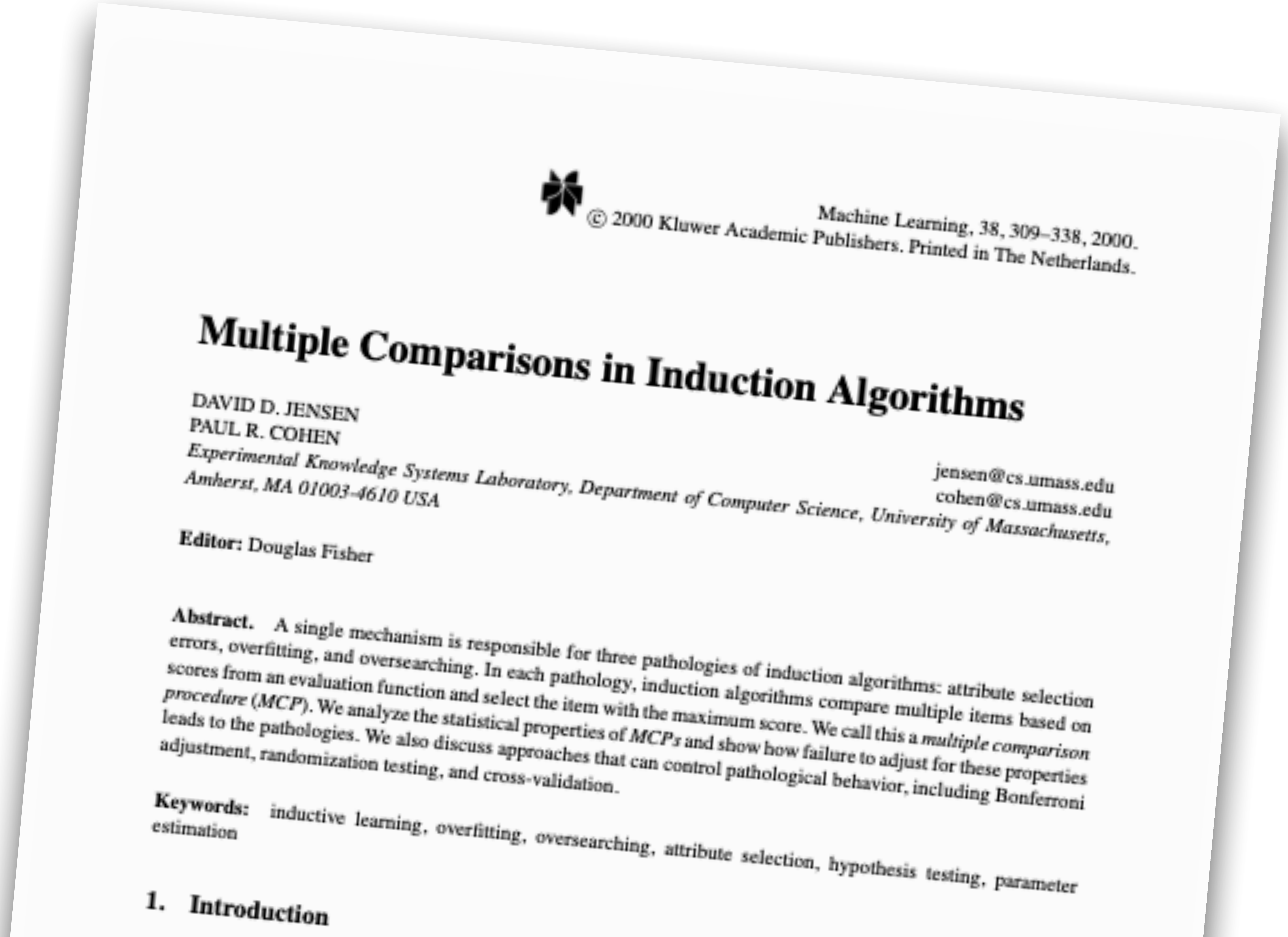

NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...


1
2
3
…
100
Publishing potential


1
2
3
...
100
Publishing potential


1
2
3
...
100

Publishing potential



Machine Learning, 38, 309–338, 2000.
© 2000 Kluwer Academic Publishers. Printed in The Netherlands.

# Multiple Comparisons in Induction Algorithms

DAVID D. JENSEN jensen@cs.umass.edu
PAUL R. COHEN cohen@cs.umass.edu
*Experimental Knowledge Systems Laboratory, Department of Computer Science, University of Massachusetts, Amherst, MA 01003-4610 USA*

**Editor:** Douglas Fisher

**Abstract.** A single mechanism is responsible for three pathologies of induction algorithms: attribute selection errors, overfitting, and oversearching. In each pathology, induction algorithms compare multiple items based on scores from an evaluation function and select the item with the maximum score. We call this a *multiple comparison procedure* (*MCP*). We analyze the statistical properties of *MCPs* and show how failure to adjust for these properties leads to the pathologies. We also discuss approaches that can control pathological behavior, including Bonferroni adjustment, randomization testing, and cross-validation.

**Keywords:** inductive learning, overfitting, oversearching, attribute selection, hypothesis testing, parameter estimation

## 1. Introduction

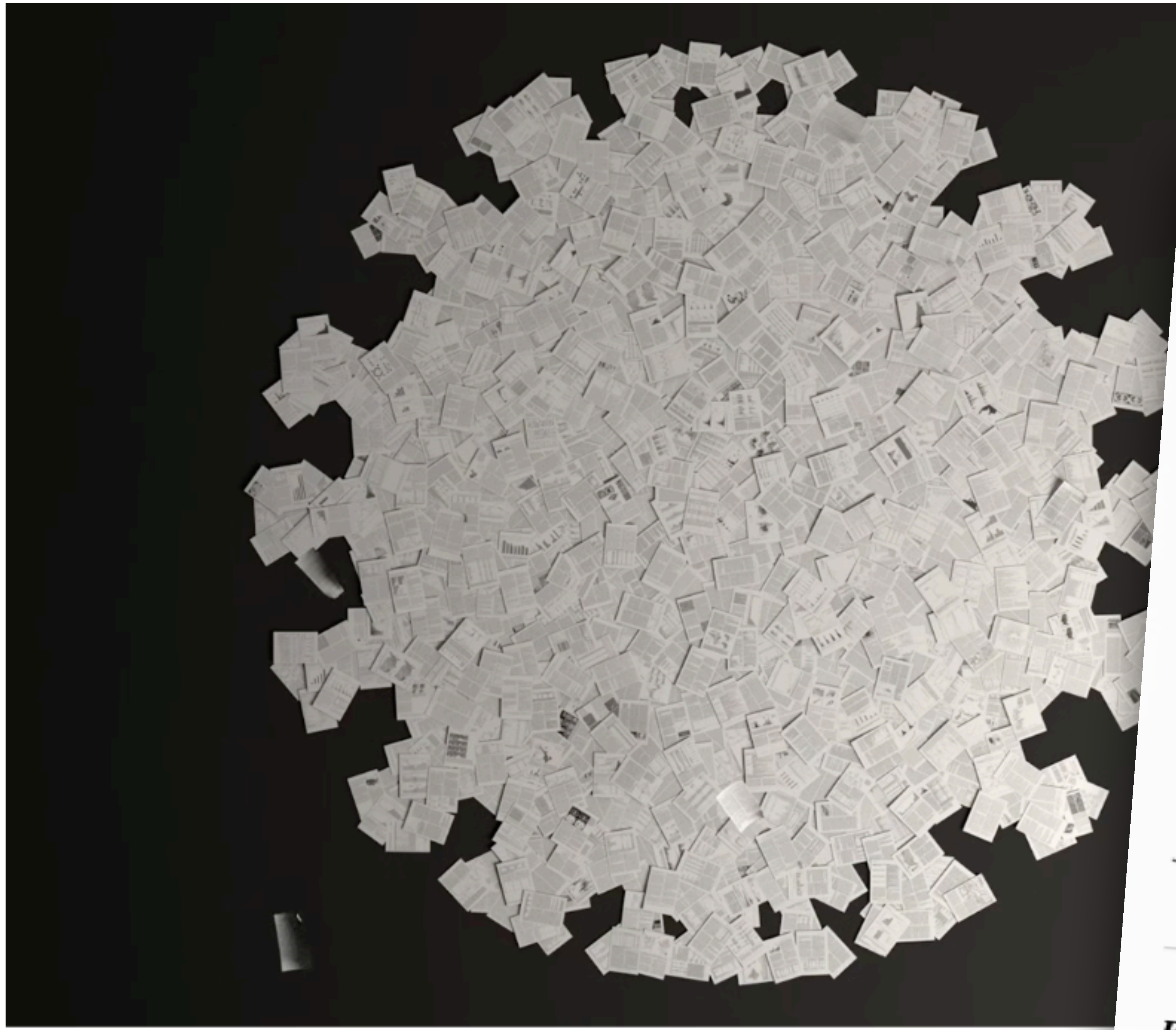
# This bias crops up in multiple places

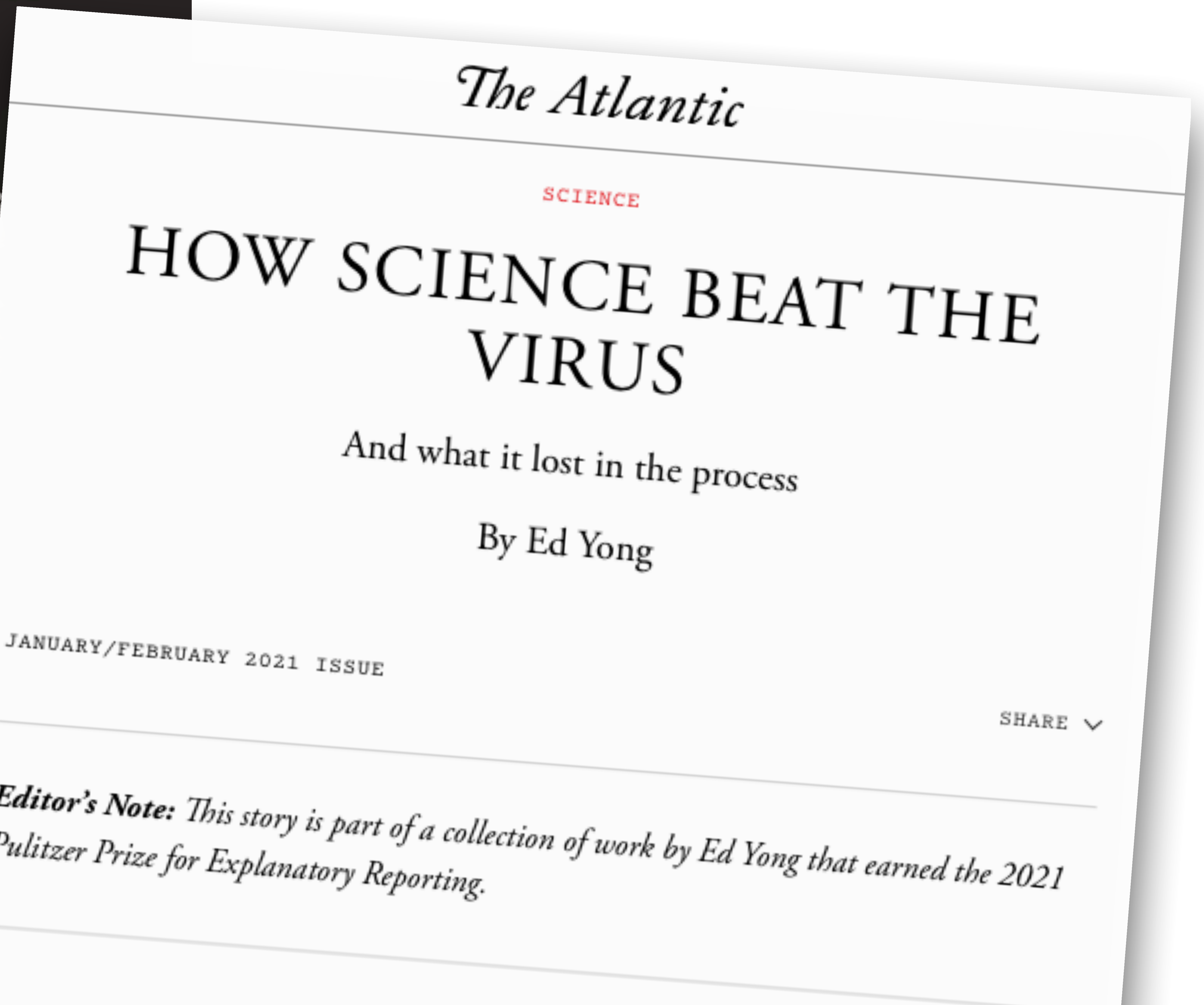
- **Individual studies** — Many design choices, each vary in their effect size they produce, and large effects are favored.
- **Publishing** — Many submitted papers, each with different effect sizes, and large effects are favored.
- **Publicity** — Many published papers, each with different effect sizes, and large effects are favored.
- **Replication** — Many publicized and cited papers, each have different effect sizes, and large effects are favored.
- **Publishing and publicity about replications** — Many replicated studies, each with different effect sizes, and small effects are favored

# Current trends in science make this even more challenging

- **More researchers** — From 1960 to 2010, the number of biological or medical researchers in the U.S. increased sevenfold, from just 30,000 to more than 220,000.
- **More papers** — The number of research papers published in 2014 was more than triple the amount published in 1990, and more than 100 times the amount published in 1950.
- **More access to papers** — ArXiv, BioArXiv, and many others.

# Case Study: COVID-19 Scholarship





The Atlantic

SCIENCE

HOW SCIENCE BEAT THE VIRUS

And what it lost in the process

By Ed Yong

JANUARY/FEBRUARY 2021 ISSUE

SHARE

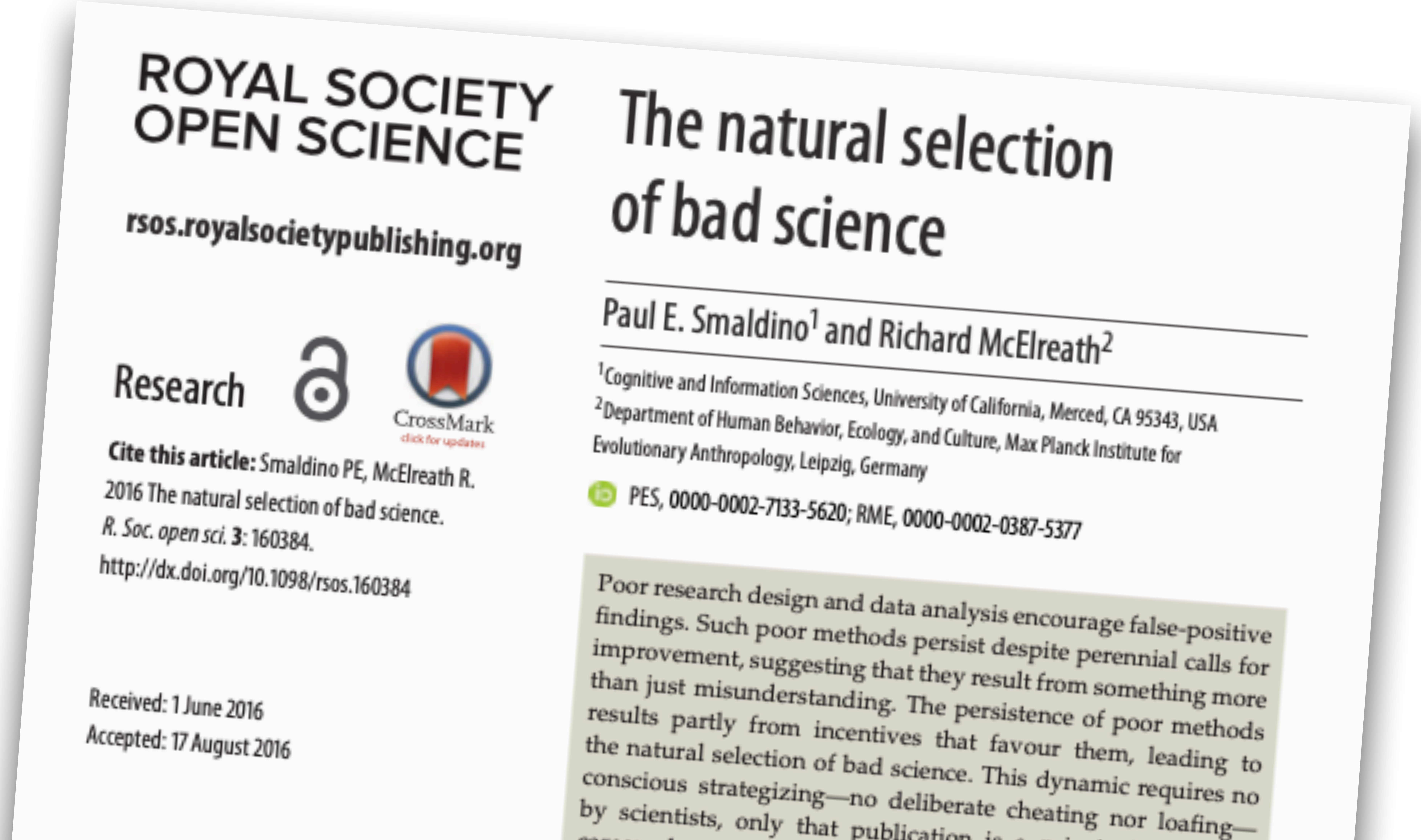
***Editor's Note:*** *This story is part of a collection of work by Ed Yong that earned the 2021 Pulitzer Prize for Explanatory Reporting.*

# Case Study: COVID-19 Scholarship

- “While the most qualified experts became quickly immersed in the pandemic response, others were stuck at home looking for ways to contribute.”
- “Using the same systems that made science faster, they could download data from free databases, run quick analyses with intuitive tools, publish their work on preprint servers, and publicize it on Twitter. Often, they made things worse by swerving out of their scholarly lanes and plowing into unfamiliar territory.”
- “The tsunami of rushed but dubious work made life harder for actual experts, who struggled to sift the signal from the noise. They also felt obliged to debunk spurious research in long Twitter threads… And they were overwhelmed by requests to peer-review new papers.”

*Source: Yong 2020*

# An additional “systems” analysis



ROYAL SOCIETY
OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

CrossMark
click for updates

**Cite this article:** Smaldino PE, McElreath R. 2016 The natural selection of bad science. *R. Soc. open sci.* **3**: 160384.
http://dx.doi.org/10.1098/rsos.160384

Received: 1 June 2016
Accepted: 17 August 2016

## The natural selection of bad science

Paul E. Smaldino[1] and Richard McElreath[2]

[1]Cognitive and Information Sciences, University of California, Merced, CA 95343, USA
[2]Department of Human Behavior, Ecology, and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

PES, 0000-0002-7133-5620; RME, 0000-0002-0387-5377

Poor research design and data analysis encourage false-positive findings. Such poor methods persist despite perennial calls for improvement, suggesting that they result from something more than just misunderstanding. The persistence of poor methods results partly from incentives that favour them, leading to the natural selection of bad science. This dynamic requires no conscious strategizing—no deliberate cheating nor loafing—by scientists, only that publication is

If multiple comparison procedures are one important cause of the perceived replication crisis, what can we do about it?

# How can we do better?

- Through the lens of multiple comparison procedures, there are at least four things we could do:
  - ***Reduce the number of items (researchers, papers, publications, etc.)*** — This seems ill advised and unlikely to succeed.
  - ***Reduce the variability of individual items*** — This seems possible (stay tuned).
  - ***Don't select the items with the maximum score*** — That seems ill advised and unlikely to succeed.
  - ***Retest on new data*** — Re-estimate the score in a way that resamples from the distribution. This also seems possible (stay tuned).

# How can we do better?

- Improve individual behavior (reduce variability)
  - *Education* — Encourage better methodology
  - *Practice* — Encourage more care in research conduct, including pre-registration
  - *Reviewing* — Encourage higher standards for evidence in reviewing.
  - *Hiring* — Hire based on the "best few" papers rather than on the total number of papers.
  - *Publicity* — Emphasize results that have been reviewed and confirmed, rather than those just released.
- However, the highest variance groups will still publish more often if other aspects of the system doesn't change, because…

# Current systems implicitly reward bias

- **Journals** — Looking for “the next big thing”, particularly those with highest profile (e.g., Science, Nature, NEJM)
- **Funding agencies** — Invest in “hot” areas and reward rapid, translational research “nuggets”
- **Press** — Report only the latest surprising findings to drive subscriptions and page-views
- **Business** — Boost short-term profits and acquire venture capital from new technology, drugs, etc.
- **Academia** — Reward “impact” (publication in high-profile journals, funding, publicity, and commercial interest) in hiring, tenure, and promotion practices.

# How can we do better?

- Restructure the system to change incentives for individuals (reduce long-term variability)
  - Enable ongoing, rapid, and transparent revision of the scientific literature (far beyond *errata*) to include long-term, ongoing reviews, tie-backs to prior work that is refuted or confirmed, etc.
  - Encourage replicability (e.g., high-profile publication only if easy-to-replicate)
  - Strongly reward long-standing, replicated results (e.g., “test of time” awards, Cochrane Reviews)
  - Clearly separate normal revision process from fraud and misconduct

jensen@cs.umass.edu

# References

- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637-644.
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436.
- Dennett, D. C. (2013). *Intuition Pumps and Other Tools for Thinking*. WW Norton & Company.
- Gelman, A., & Loken, E. (2014). The statistical crisis in science: data-dependent analysis--a" garden of forking paths"--explains why many statistically significant comparisons don't hold up. *American Scientist*, 102(6), 460-466.
- Geman, D., & Geman, S. (2016). Opinion: Science in the age of selfies. *Proceedings of the National Academy of Sciences*, 113(34), 9384-9387.
- Ioannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, 294(2), 218-228.
- Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, 640-648.
- Jensen, D. and Cohen, P. (2000). Multiple comparisons in induction algorithms. *Machine Learning* 38(3):309-338.
- 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.
- Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9), 160384.
- Yong, E. (2021). How Science Beat the Virus, and What It Lost in the Process. *The Atlantic*. January/February.