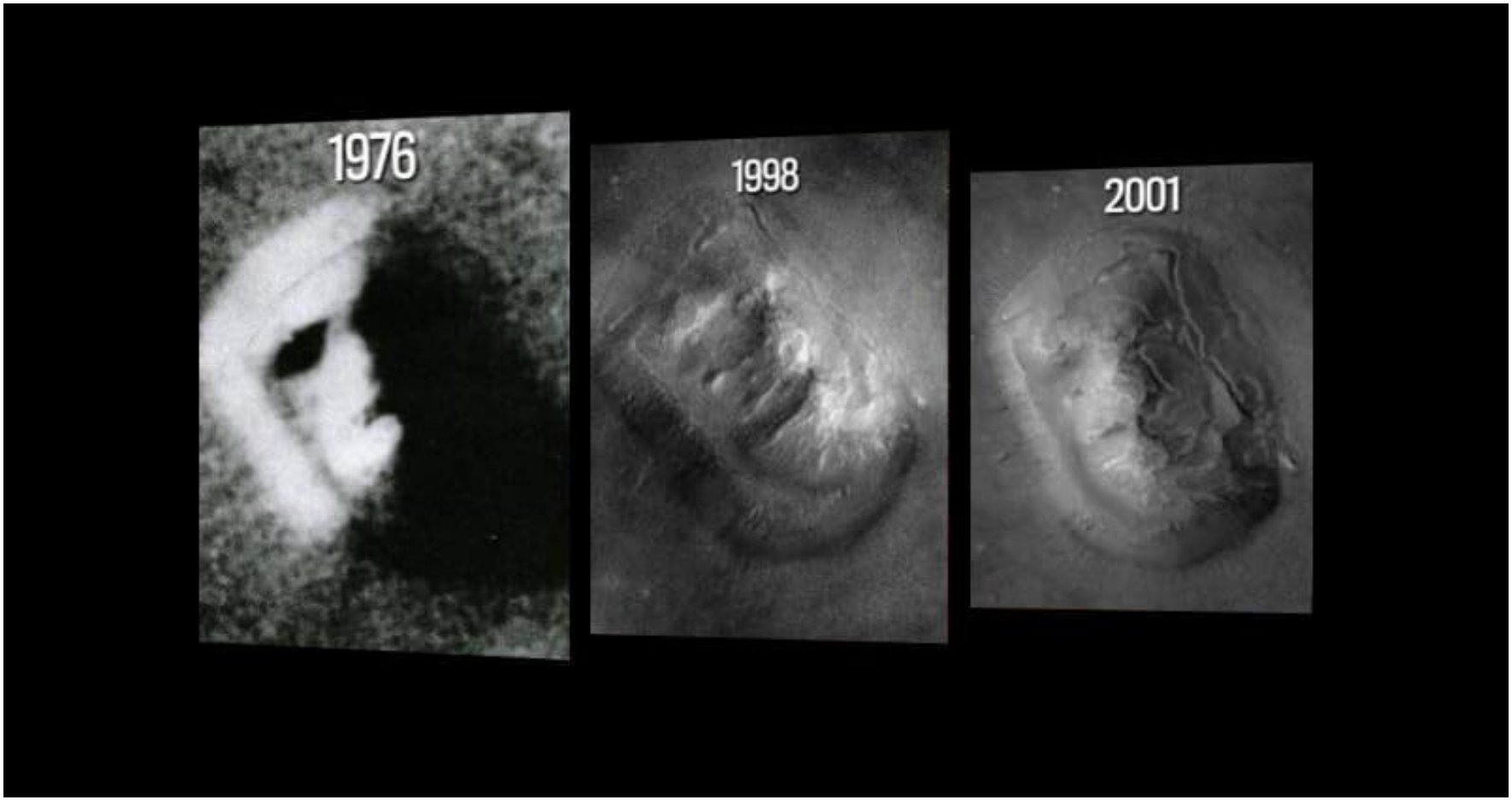
SCIENTIFIC ECOSYSTEMS AND RESEARCH REPRODUCIBILITY

Marcus Munafò

University of BRISTOL

MRC Integrative Epidemiology Unit

Fooled by Randomness



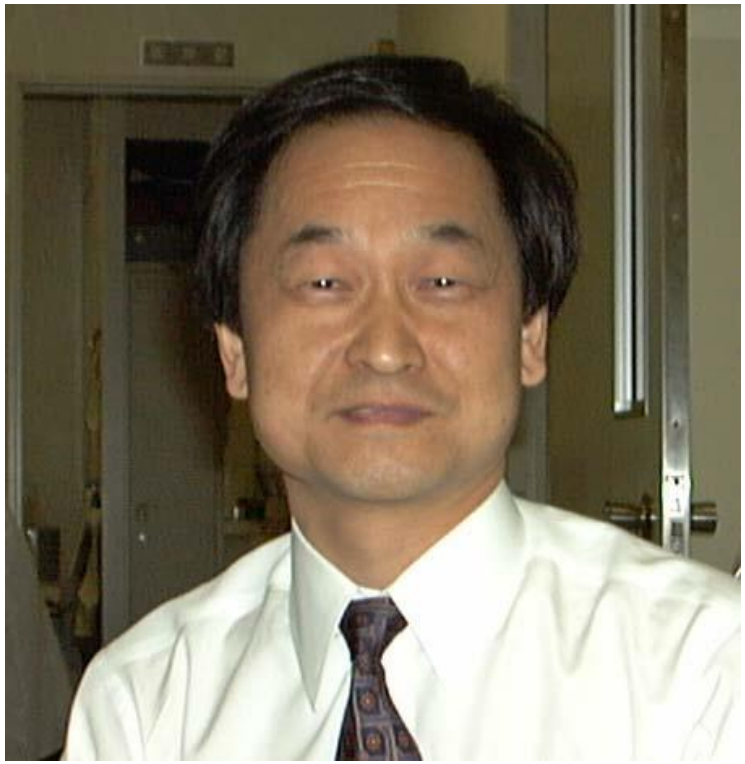
Academic Fraud

"The investigators found 'several dozens of publications in which fictitious data had been used. Fourteen of the 21 PhD theses Stapel supervised are also tainted, the committee concluded".

http://news.sciencemag.org/scienceinsider/2011/10/report-dutch-lord-of-the-data-fo.html

Academic Fraud

"Yoshitaka Fujii is a Japanese researcher in anesthesiology, who in 2012 was found to have fabricated data in at least 172 scientific papers, setting what is believed to be a record for the number of papers by a single author requiring retractions".



http://en.wikipedia.org/wiki/Yoshitaka_Fujii

Academic Fraud

"…the data of most, if not all, of the experiments have been falsified so as to agree closely with Mendel's expectations…"

R.A. Fisher

http://www.nih.gov/about/director/ebiomed/mendel.htm

The Reproducibility Crisis?

OPEN ACCESS Freely available online

PLOS ONE

A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic

Aaron Mobley[1], Suzanne K. Linder[2], Russell Braeuer[1], Lee M. Ellis[1,3]*, Leonard Zwelling[4]*

1 Department of Cancer Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America, **2** Department of General Internal Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America, **3** Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America, **4** Department of Experimental Th[illegible] Houston, Texas, United States of America

CORRESPONDENCE

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

Florian Prinz, Thomas Schlange and Khusru Asadullah

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science

Perspectives on Psychological Science
7(6) 657–660

DOI: 10.1177/1745691612462588
http://pps.sagepub.com
SAGE

Open Science Collaboration[1]

Abstract

Reproducibility is a defining feature of science. However, because of strong incentives for innovation and weak incentives for confirmation, direct replication is rarely practiced or published. The Reproducibility Project is an open, large-scale, collaborative effort to systematically examine the rate and predictors of reproducibility in psychological science. So far, 72 volunteer researchers from 41 institutions have organized to openly and transparently replicate studies published in three prominent psychological journals in 2008. Multiple methods will be used to evaluate the findings, calculate an empirical rate of replication, and investigate factors that predict reproducibility. Whatever the result, a better understanding of reproducibility will ultimately improve confidence in scientific methodology and findings.

University of BRISTOL

MRC Integrative Epidemiology Unit

…or Opportunity?

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

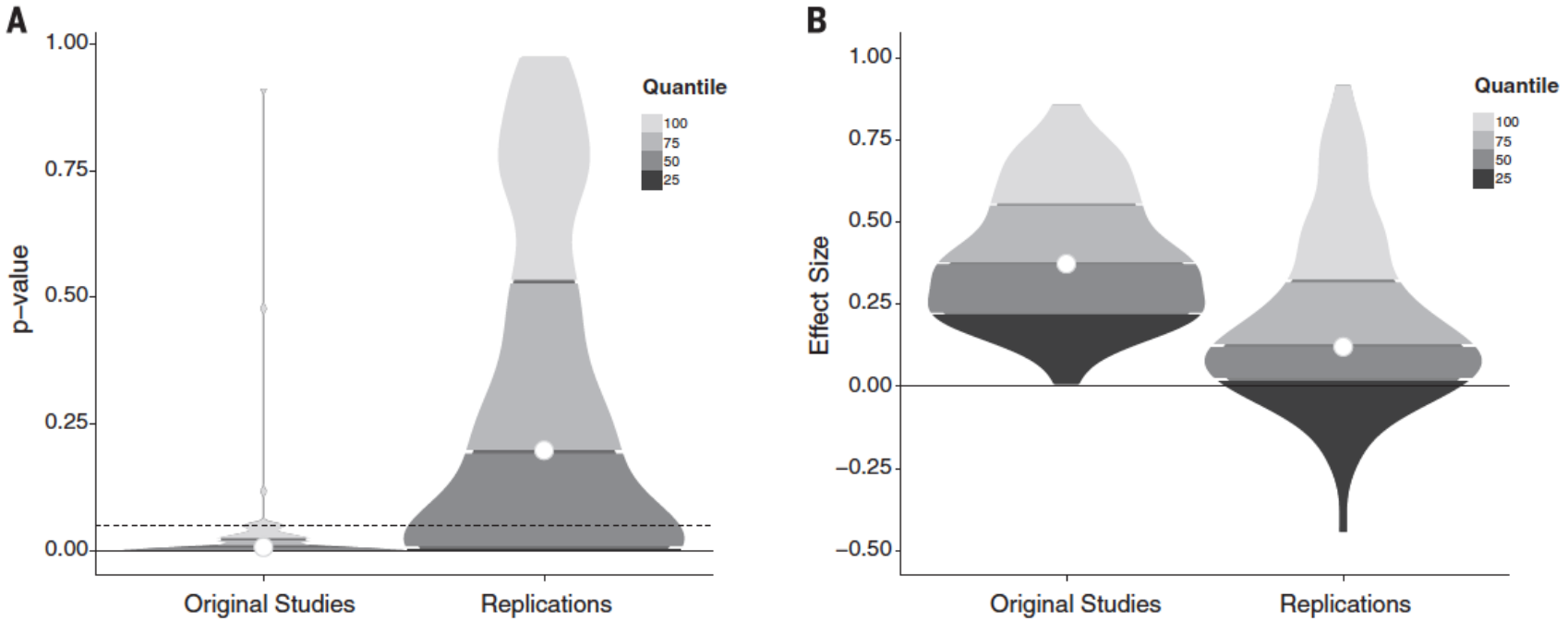


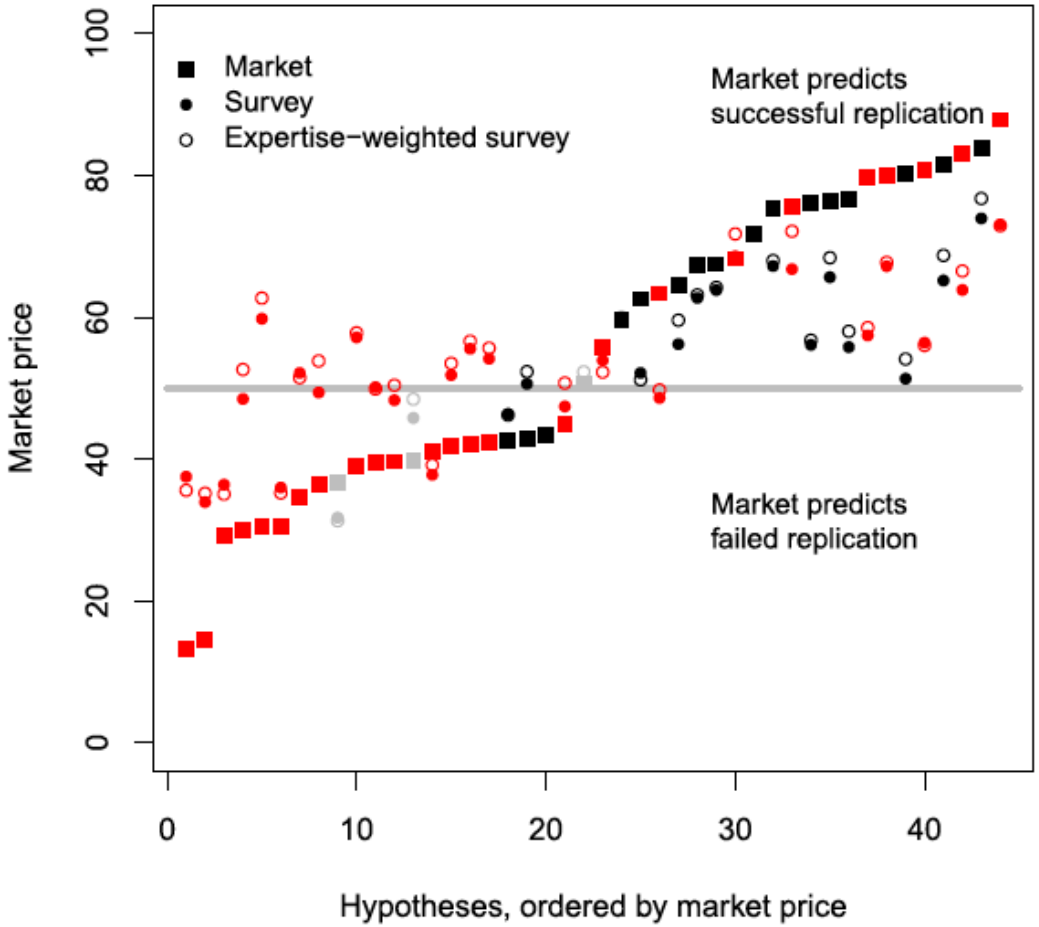
Fig. 1. Density plots of original and replication *P* values and effect sizes. (**A**) *P* values. (**B**) Effect sizes (correlation coefficients). Lowest quantiles for *P* values are not visible because they are clustered near zero.

Open Science Collaboration (2015). Science, 349.

Surprised?

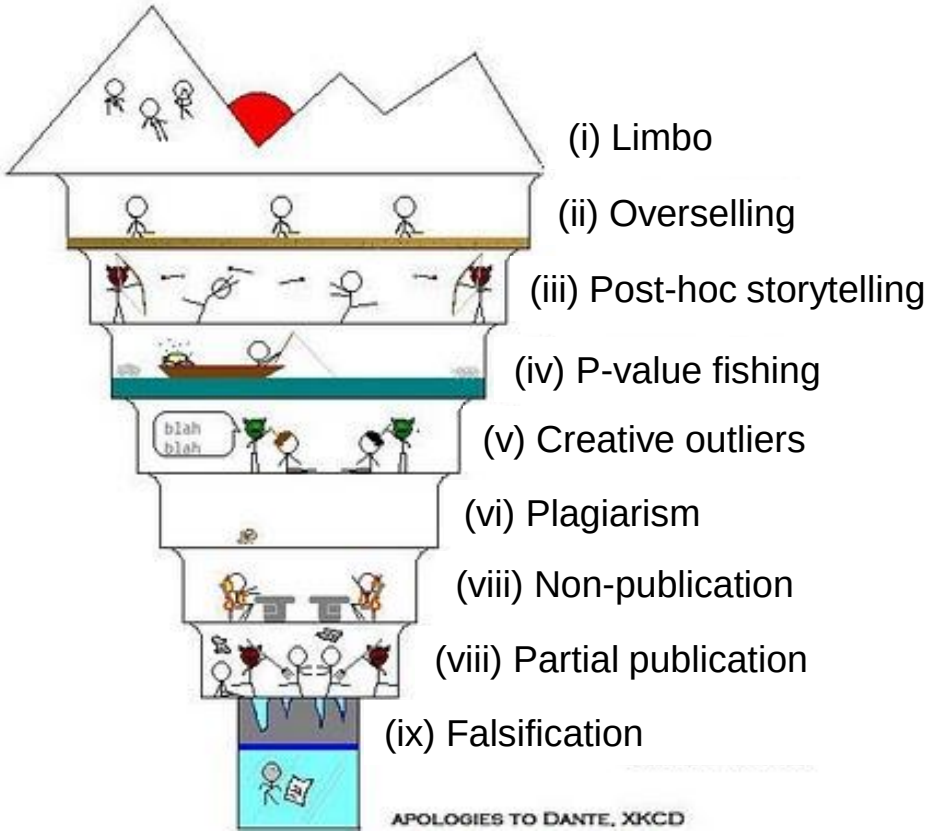
Prediction market on the outcomes of the Reproducibility Project: Psychology

Successful replications are shown in black, unsuccessful replications in red.



Dreber et al. (2015). PNAS, 112, 15343-15347.

Questionable Research Practices



Neuroskeptic (2012). Perspect Psychol Sci, 7, 643-644.

Questionable Practices

General Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
22(11) 1359–1366

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

---

Using the same method as in Study 1, we asked ~~20~~ 34 University of Pennsylvania undergraduates to listen only to either "When I'm Sixty-Four" by The Beatles or "Kalimba" or "Hot Potato" by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. **Then, in an ostensibly unrelated task, they indicated only their birth date (mm/dd/yyyy) and** how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with "computers are complicated machines," **their father's age,** their mother's age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as "the good old days," and their gender. **We used father's age to control for variation in baseline age across participants.**

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted *M* = 20.1 years) rather than to "Kalimba" (adjusted *M* = 21.5 years), *F*(1, 17) = 4.92, *p* = .040. Without controlling for father's age, the age difference was smaller and did not reach significance (*M*s = 20.3 and 21.2, respectively), *F*(1, 18) = 1.01, *p* = .33.

Simmons et al. (2011). Psychol Sci, 22, 1359-1366.

University of BRISTOL

MRC Integrative Epidemiology Unit

Questionable Practices

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

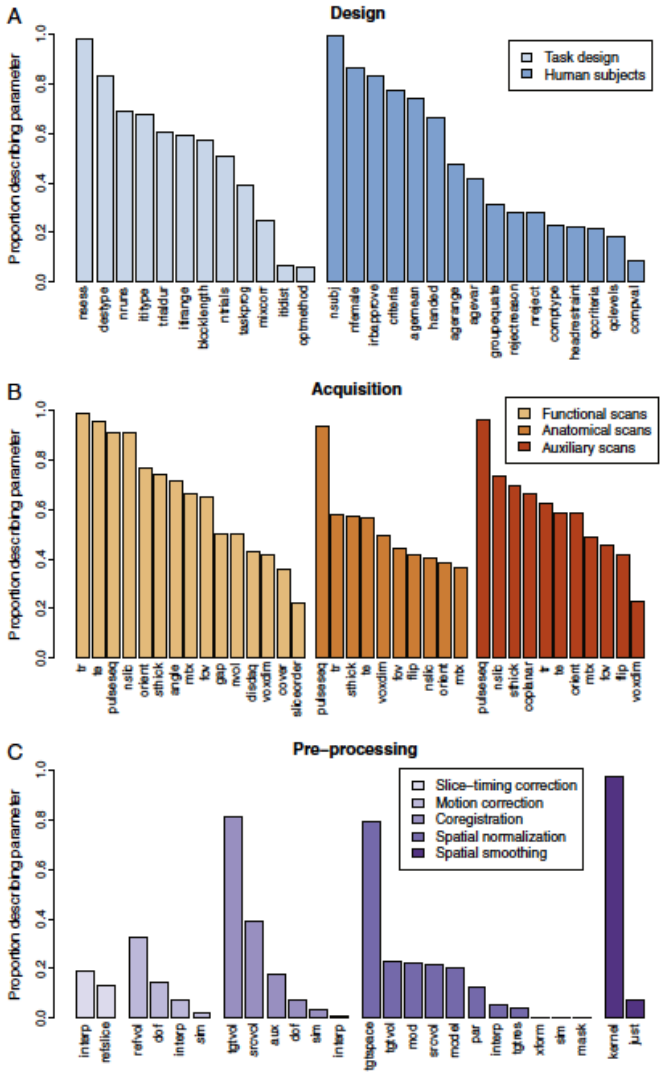
Full Length Articles

The secret lives of experiments: Methods reporting in the fMRI literature

Joshua Carp

University of Michigan, Department of Psychology, 530 Church Street, Ann Arbor, MI, 48109, USA

"…nearly as many unique analysis pipelines as there were studies in the sample…"

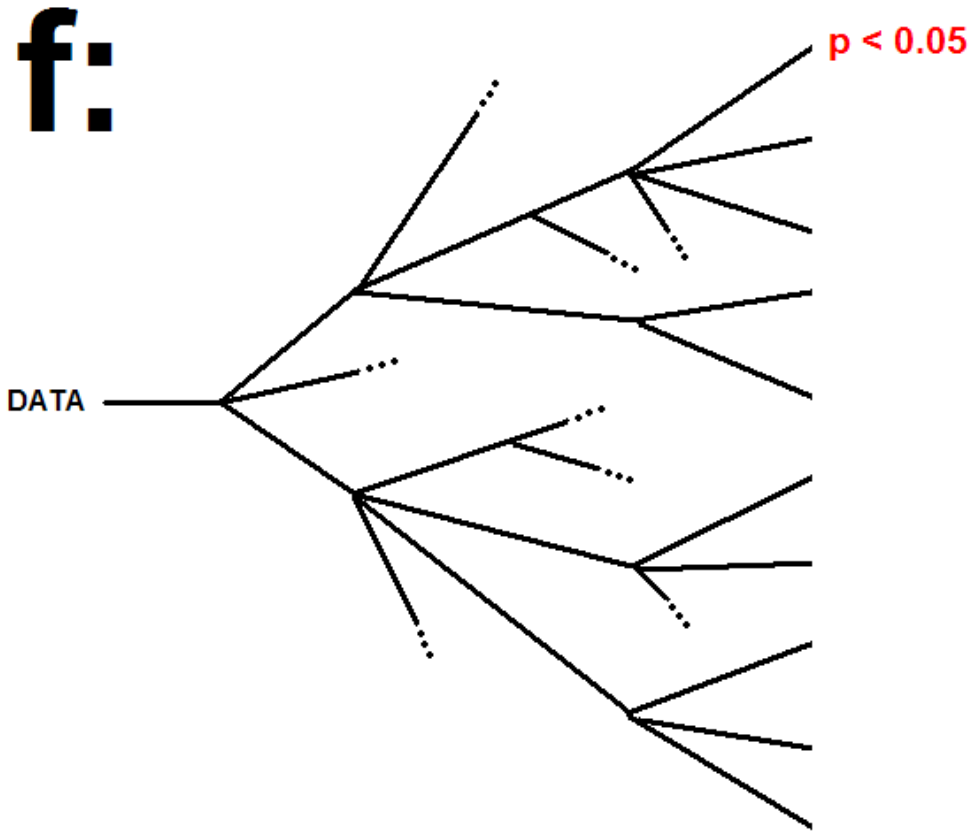


Carp (2012). Neuroimage, 63, 289-300.

University of BRISTOL

MRC Integrative Epidemiology Unit

Questionable Practices



http://blogs.discovermagazine.com/neuroskeptic/2013/10/16/the-f-problem

Incentive Structures

aps
ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(3) 307–309

DOI: 10.1177/1745691612441216
http://pps.sagepub.com
SAGE

We Knew the Future All Along: Scientific Hypothesizing is Much More Accurate Than Other Forms of Precognition—A Satire in One Part

Arina K. Bones
University of Darache, Monte Carlo, Monaco

Bones (2012). Perspect Psychol Sci, 7, 307.

Incentive Structures

US studies may overestimate effect sizes in softer research

Daniele Fanelli[a,1] and John P. A. Ioannidis[b,c,d]

Predictor	Nonbehavioral (k = 40, n = 566)	Behavioral, all (k = 42, n = 608)	Biobehavioral (k = 20, n = 308)	Behavioral (k = 22, n = 300)
(Intercept)	0.42 [0.40, 0.46]	0.55 [0.51, 0.56]	0.51 [0.47, 0.54]	0.57 [0.50, 0.59]
United States vs. rest	−0.02 [−0.06, 0.00]	0.03 [0.02, 0.06]	0.03 [0.00, 0.07]	0.04 [0.01, 0.07]
Study size (SE)	0.43 [0.27, 0.53]	0.11 [0.07, 0.23]	0.20 [0.11, 0.31]	0.06 [0.01, 0.29]
Pub. order	0.02 [0.00, 0.03]	0.00 [−0.01, 0.01]	0.01 [0.00, 0.05]	0.00 [−0.02, 0.01]
USA*SE	−0.21 [−0.47, 0.22]	−0.19 [−0.31, −0.03]	−0.16 [−0.34, 0.12]	−0.22 [−0.46, −0.02]
USA*pub. order	−0.02 [−0.05, 0.01]	0.00 [−0.02, 0.03]	−0.02 [−0.06, 0.01]	0.01 [−0.02, 0.05]

Fanelli & Ioannidis (2013). PNAS, 5, e10271.

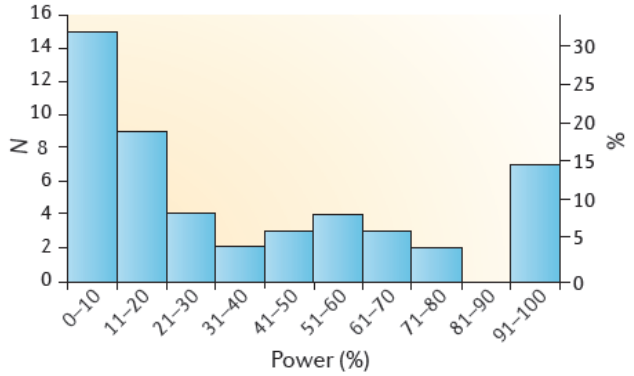
Incentive Structures

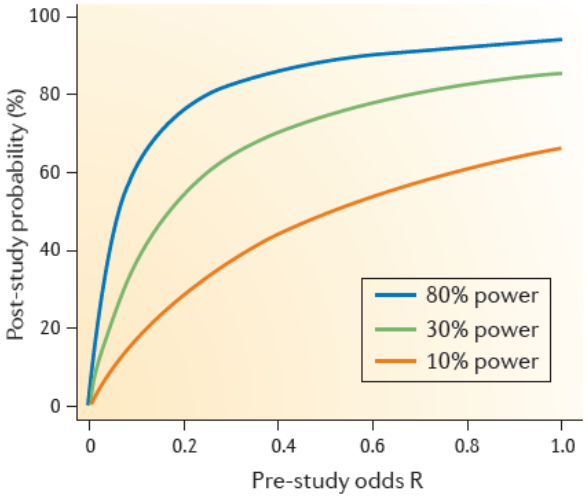
ANALYSIS

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

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored methodological principles.





Button et al. (2013). Nat Rev Neurosci, 14, 365-376.

University of BRISTOL

MRC Integrative Epidemiology Unit

Incentive Structures

Table 1. Reporting of measures to reduce the risk of bias in publications from 2009–2010 that were randomly selected, identified in the context of systematic reviews or from leading UK institutions.

	Randomisation		Blinding		Sample Size Calculation	
	n/N	% (95% CI)	*n/N*	% (95% CI)	*n/N*	% (95% CI)
PubMed	7/14	50 (23–77)	2/14	14 (2–43)	0/14	0 (0–23)
CAMARADES	76/213	36 (29–42)	79/213	37 (30–44)	2/213	1 (0–3)
Institutions	148/1028	14 (12–17)	201/1165	17 (15–20)	16/1168	1 (1–2)

Studies from top-ranked UK institutions perform worse on reporting of measures to reduce the risk of bias than studies selected at random from PubMed…

Macleod et al. (2015). PLOS Biology.

But Science is Self-Correcting!

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(6) 645–654

DOI: 10.1177/1745691612464056
http://pps.sagepub.com
SAGE

Why Science Is Not Necessarily Self-Correcting

John P. A. Ioannidis
Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy, Stanford University School of Medicine, and Department of Statistics, Stanford University School of Humanities and Sciences

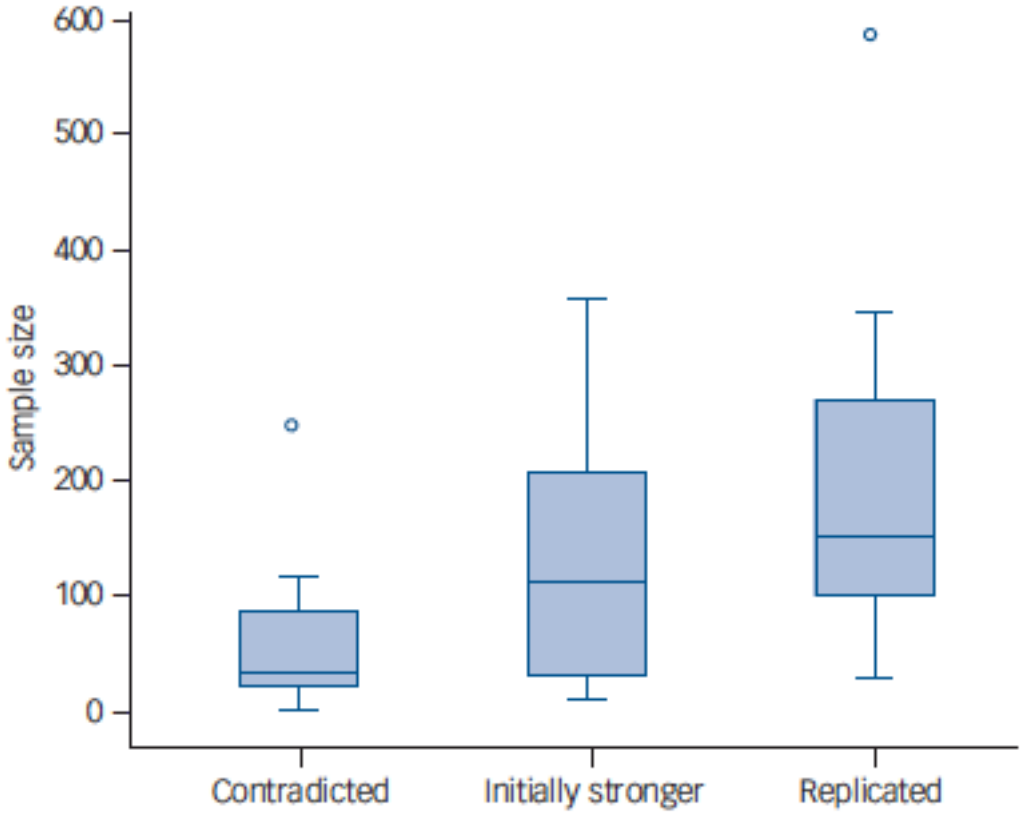
Ioannidis (2012). Perspect Psychol Sci, 7, 645-654.

University of BRISTOL

MRC Integrative Epidemiology Unit

But Science is Self-Correcting!

"Among 83 articles recommending effective interventions, 40 had not been subject to any attempt at replication…"



Tajika et al. (2015). Br J Psychiatry, 207, 357-362.

But Science is Self-Correcting!

Journal of Clinical Epidemiology ■ (2012) ■

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Primary study authors of significant studies are more likely to believe that a strong association exists in a heterogeneous meta-analysis compared with methodologists

Orestis A. Panagiotou[a], John P.A. Ioannidis[b,c,d,e,*]

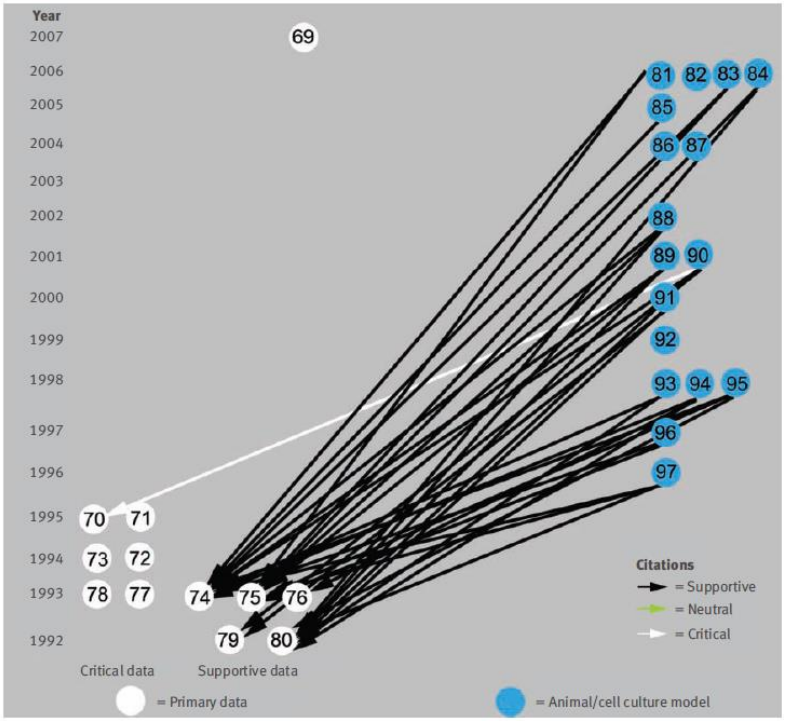
Panagiotou & Ioannidis (2012). J Clin Epidemiol, 65, 740-747.

But Science is Self-Correcting!

How citation distortions create unfounded authority: analysis of a citation network

Steven A Greenberg, associate professor of neurology

Investigated citation network of papers addressing the belief that B amyloid, a protein accumulated in the brain in Alzheimer's disease, is produced by and injures skeletal muscle of patients with inclusion body myositis.



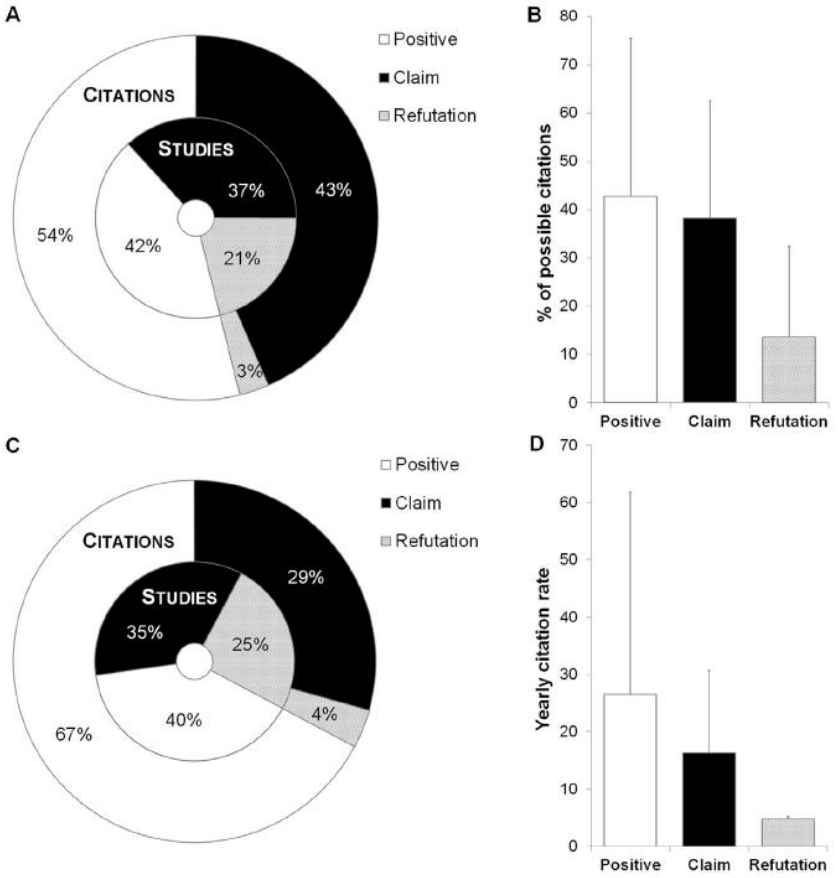
Greenberg (2009). Br Med J, 339, b2680.

But Science is Self-Correcting!

Abstracts often “spin” results to give impression that results are positive when they are not.

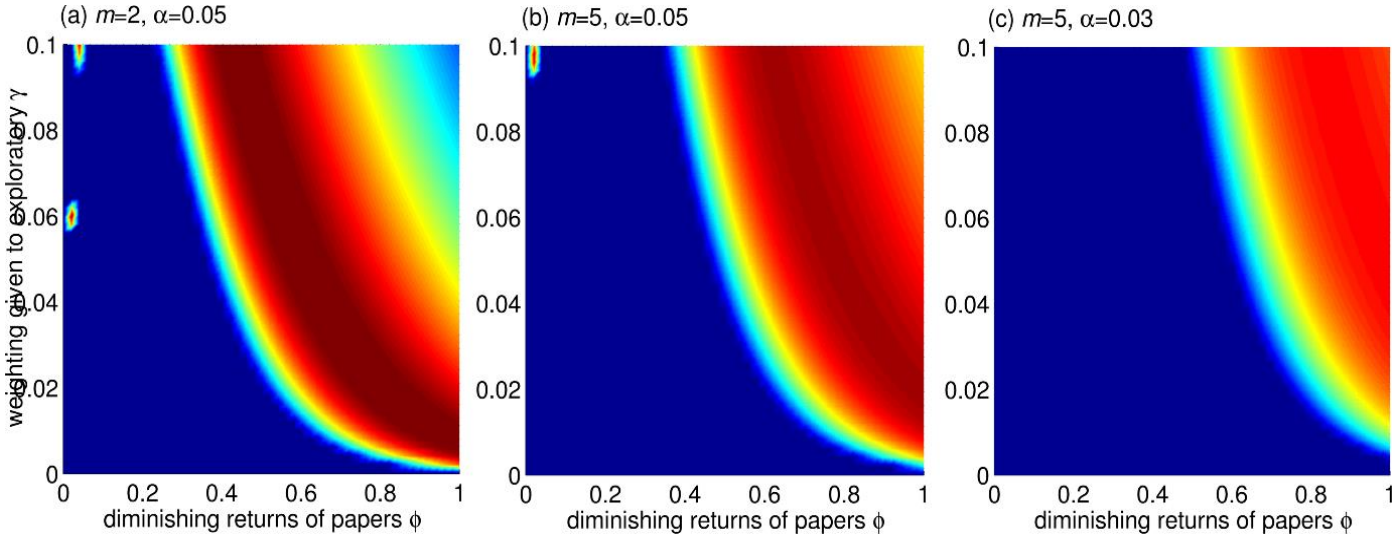
Citation inflation exists for both “positive” studies and “claim” studies in this literature.

True both within this literature (A, B) and in the wider (Web of Science) literature (C, D).



Bastiaansen et al. (2015). Biol Psychiatry.

Scientific Ecosystems



% correct

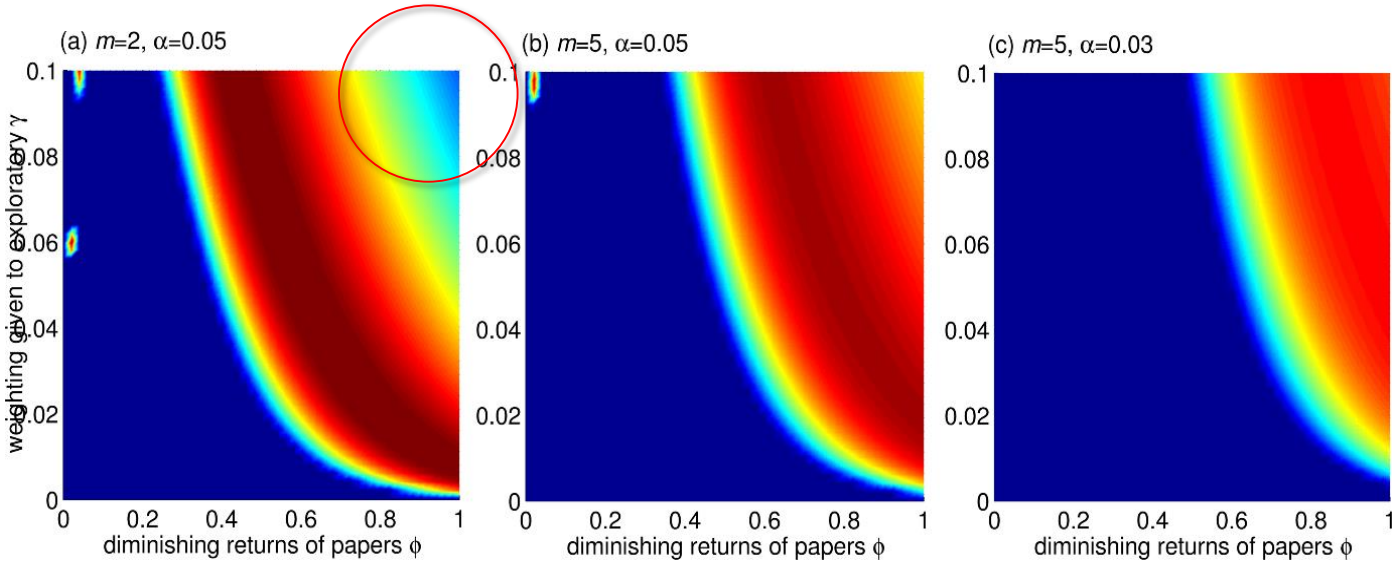
red = high
blue = low

Journals can require higher power (*m*) and/or more stringent P-value (α).

Evaluations can give more weight to confirmatory studies and number of studies.

Higginson & Munafò (in preparation).

Scientific Ecosystems

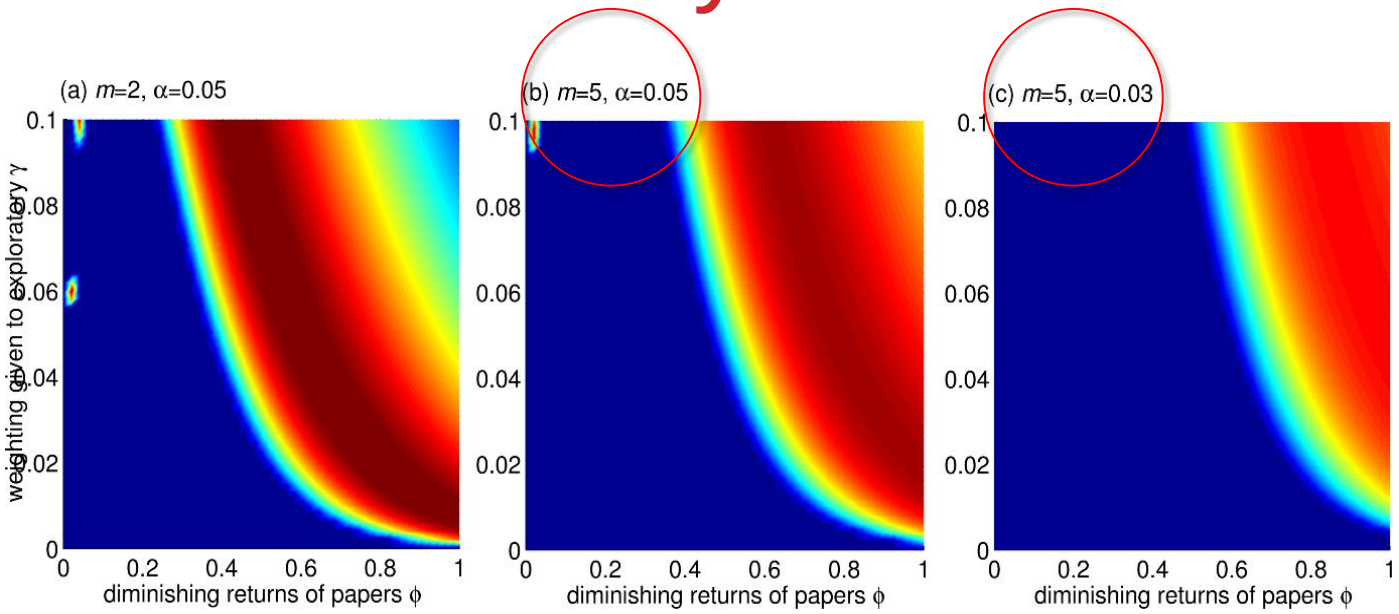


Value of Science

red = high
blue = low

Higginson & Munafò (in preparation).

Scientific Ecosystems



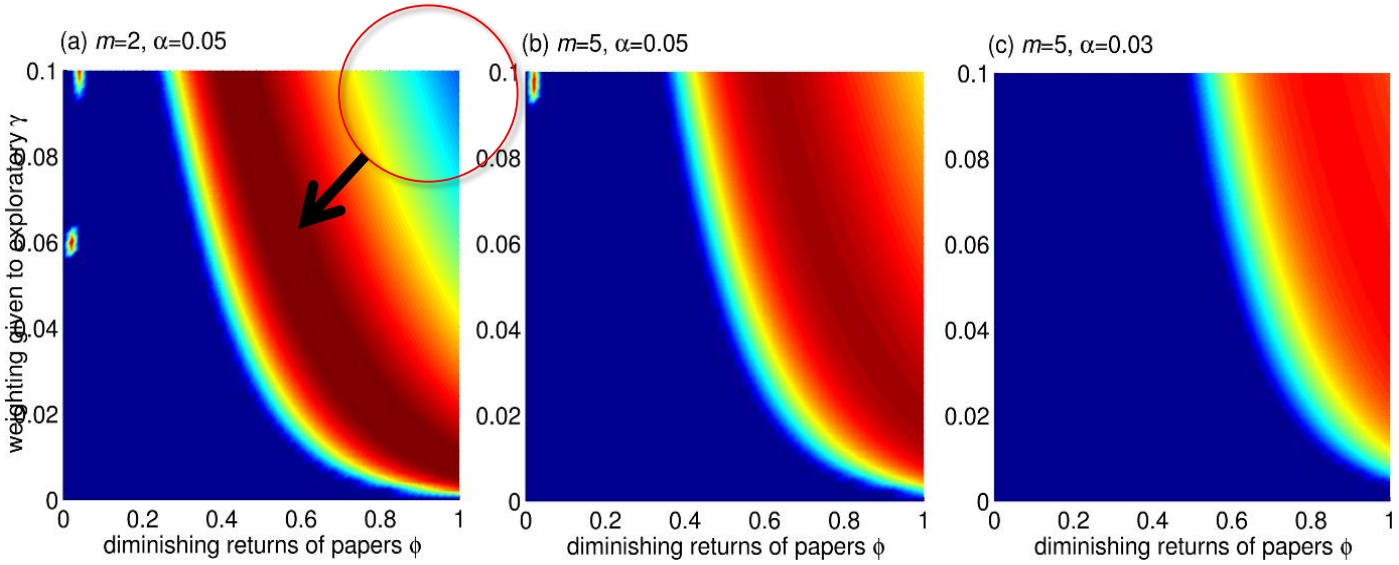
Value of Science

red = high
blue = low

Journals can require higher **power** (*m*) and/or more **stringent P-values** (α).

Higginson & Munafò (in preparation).

Scientific Ecosystems



Value of Science

red = high
blue = low

Journals can require higher **power** (*m*) and/or more **stringent P-values** (α).

Evaluations can give more weight to **confirmatory** studies and **number** of studies.

Higginson & Munafò (in preparation).

Crisis or Opportunity?

Scientific rigor and the art of motorcycle maintenance

Marcus Munafò, Simon Noble, William J Browne, Dani Brunner, Katherine Button, Joaquim Ferreira, Peter Holmans, Douglas Langbehn, Glyn Lewis, Martin Lindquist, Kate Tilling, Eric-Jan Wagenmakers & Robi Blumenstein

The reliability of scientific research is under scrutiny. A recently convened working group proposes cultural adjustments to incentivize better research practices.

Like auto manufacturing in the 1970s, scientific research is producing too many lemons.

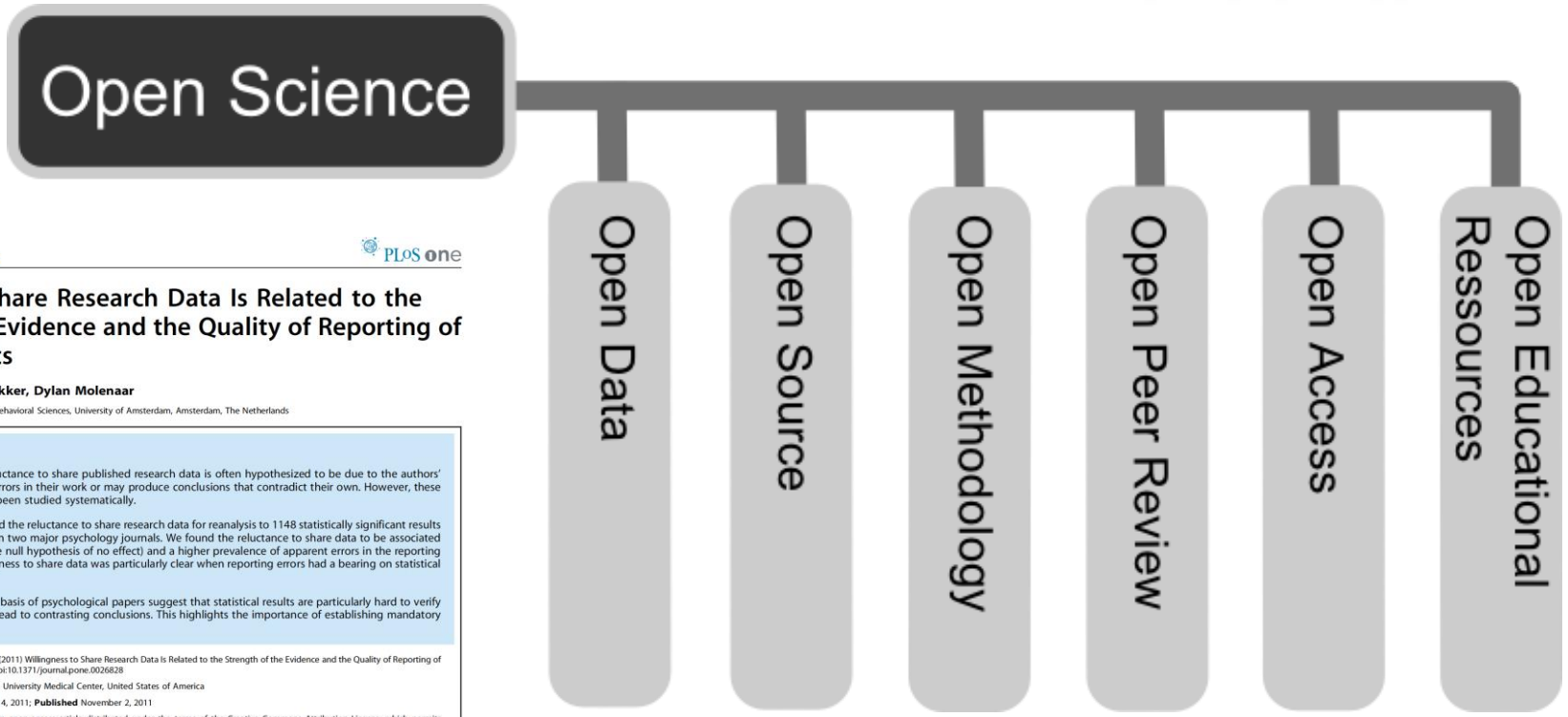
Munafò et al. (2014), Nat Biotech, 32, 871-873.

University of BRISTOL

MRC Integrative Epidemiology Unit

Crisis or Opportunity?

CENTER FOR OPEN SCIENCE



OPEN ACCESS Freely available online

PLoS one

Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results

Jelte M. Wicherts*, Marjan Bakker, Dylan Molenaar

Psychology Department, Faculty of Social and Behavioral Sciences, University of Amsterdam, Amsterdam, The Netherlands

Abstract

Background: The widespread reluctance to share published research data is often hypothesized to be due to the authors' fear that reanalysis may expose errors in their work or may produce conclusions that contradict their own. However, these hypotheses have not previously been studied systematically.

Methods and Findings: We related the reluctance to share research data for reanalysis to 1148 statistically significant results reported in 49 papers published in two major psychology journals. We found the reluctance to share data to be associated with weaker evidence (against the null hypothesis of no effect) and a higher prevalence of apparent errors in the reporting of statistical results. The unwillingness to share data was particularly clear when reporting errors had a bearing on statistical significance.

Conclusions: Our findings on the basis of psychological papers suggest that statistical results are particularly hard to verify when reanalysis is more likely to lead to contrasting conclusions. This highlights the importance of establishing mandatory data archiving policies.

Citation: Wicherts JM, Bakker M, Molenaar D (2011) Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. PLoS ONE 6(11): e26828. doi:10.1371/journal.pone.0026828

Editor: Rochelle E. Tractenberg, Georgetown University Medical Center, United States of America

Received May 20, 2011; **Accepted** October 4, 2011; **Published** November 2, 2011

Funding: The preparation of this article was supported by three grants (021-001-124, 451-07-016, and 400-08-214) from the Netherlands Organization for Scientific Research (NWO). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

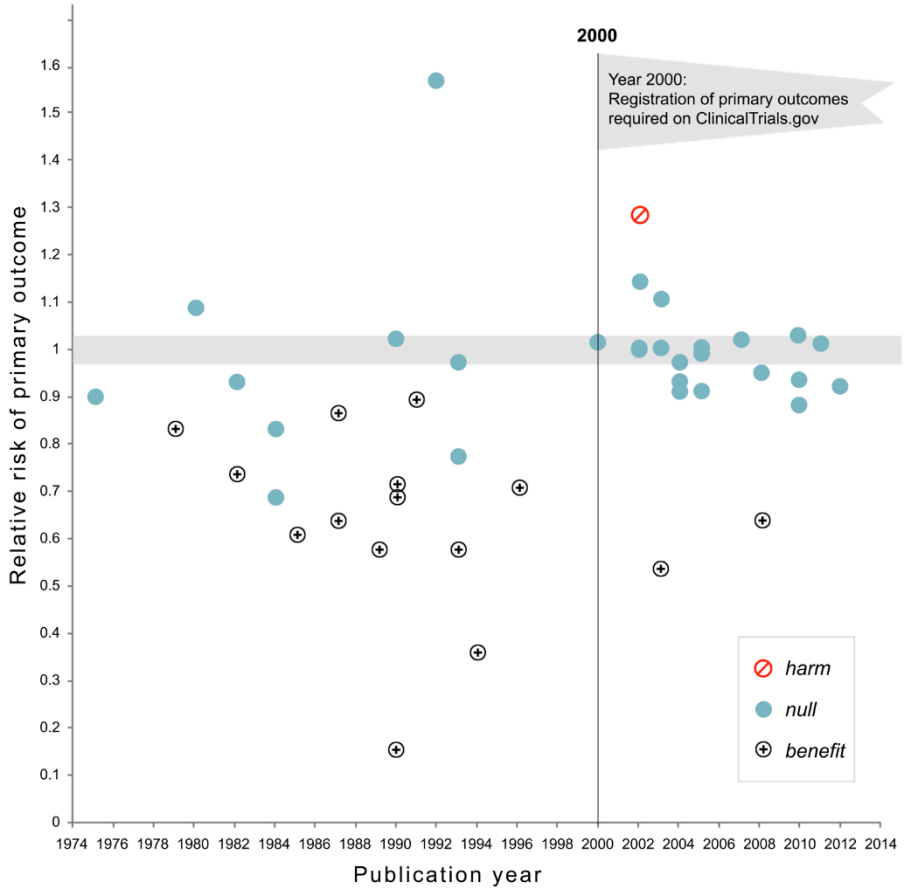
* E-mail: j.m.wicherts@uva.nl

University of BRISTOL

MRC Integrative Epidemiology Unit

Crisis or Opportunity?

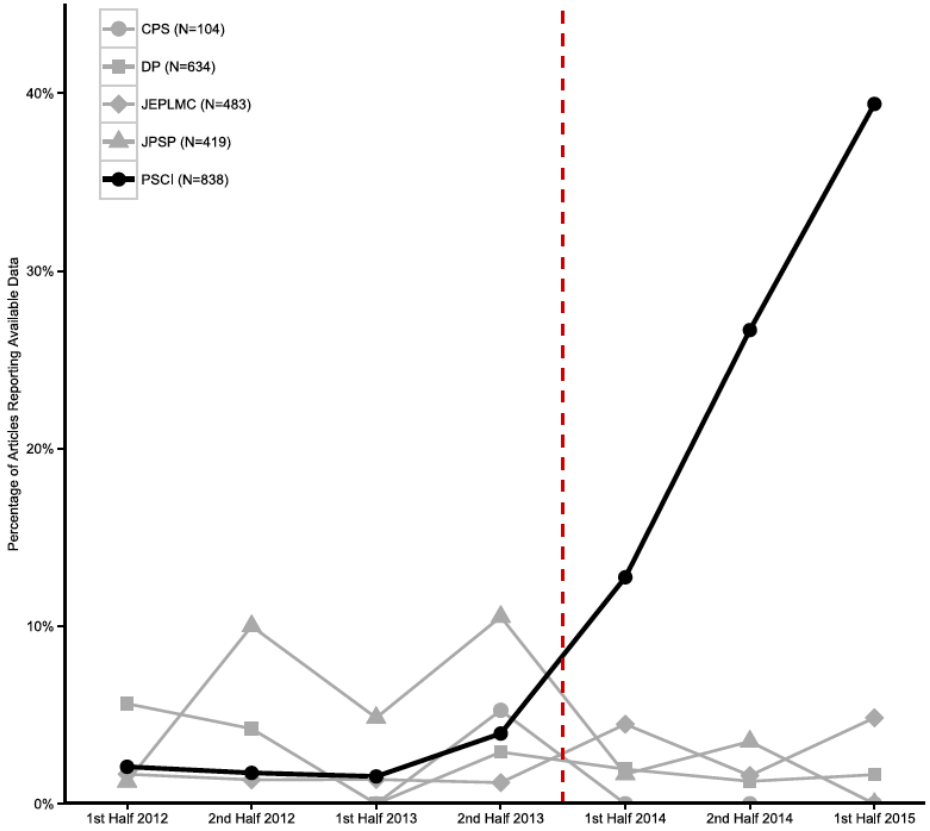
In 2000 the National Heart Lung, and Blood Institute required the registration of primary outcome on ClinicalTrials.gov for all their grant-funded activity



Kaplan & Irvin (2015). PLoS One, 10, e0132382.

Crisis or Opportunity?

Introduction of badges for open practices at *Psychological Science* followed by a steep increased in data sharing.



Kidwell et al. (2016). PLoS Biology, 14, e1002456.

Have We Been Here Before?

"This is an art of various forms, the object of which is to give ordinary observations the appearance and character of those of the highest degree of accuracy. One of its numerous processes is to make multitudes of observations, and out of these to select only those which agree, or very nearly agree. If a hundred observations are made, the cook must be very unhappy if he cannot pick out fifteen or twenty which will do for serving up."

Babbage (1830). Reflections on the Decline of Science in England.

Acknowledgements

marcus.munafo@bristol.ac.uk

@MarcusMunafo

@BristolTARG

http://www.bristol.ac.uk/expsych/research/brain/targ/

UKCTAS

UK Centre for Tobacco & Alcohol Studies

Tobacco and Alcohol Research Group:

Olivia Abrams	Research Assistant
Angela Attwood	Postdoc
Anna Blackwell	Research Assistant
Alex Board	Research Assistant
Emily Crowe	PhD Student
Katie Drax	Research Assistant
Maddy Dyer	PhD Student
Kayleigh Easey	PhD Student
Andy Eastwood	PhD Student
Meg Fluharty	PhD Student
Suzi Gage	Postdoc
Meryem Grabski	PhD Student
Eleanor Kennedy	PhD Student
Jasmine Khouja	PhD Student
Glenda Lassi	Postdoc
Rebecca Lawn	PhD Student
Jim Lumsden	PhD Student
Olivia Maynard	Postdoc
Hannah Sallis	Postdoc
Carlos Sillero	Research Assistant
Andy Skinner	Postdoc
Chris Stone	Research Assistant
Amy Taylor	Postdoc
Gemma Taylor	Postdoc
Daniel Toze	Research Assistant
David Troy	PhD Student

University of BRISTOL

MRC Integrative Epidemiology Unit