Implications of Soundness-Dependent Effects for Interventions in the Replication Crisis

Aydin Mohseni

University of California, Irvine

Department of Logic and Philosophy of Science

The replication crisis in the social & life sciences

RESEARCH

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estima
psycho

Open Science C

INTRODUCTIO ing feature of sci it characterizes c Scientific claims because of the originator but supporting evide plary quality ma ical findings beca error.

RATIONALE: Th and predictors o evidence. Potenti clude selective re insufficient speci essay or sufficie replication is the ditions believed

substantial decline. Ninety-seven percent of original studies had significant results ($P < .05$). Thirty-six percent of replications had signifi-

RESEARCH | REPORTS

nonradiatively [calculation in the supplementary materials (21)] in a perovskite film in air on glass. This value relates to a photon recycling–assisted average e (fig. S18). The ave hanced at larger d high fluences) and

In terms of e gest that the ave can travel in a p charge-carrier d recombination is in the film, the e can propagate ov creates a distincti diffusion lengths contradiction of r and long diffusio

What are the i presented here f solar cells (3, 6)? work provide va structures. Using developed above, circuit conditions 350 nm and non produces a doub sity under 1-sun be enhanced furt decay channels fluences, such as high bimolecular In the ideal case back mirror, phot nal photon densi 2.5) (31) in pero circuit conditions the use of highly imize photonic l surface, offers p photon recyclin efficiencies of pe Shockley-Queisse lead to higher int up of excited cha quasi-Fermi leve open-circuit volta

REFERENCES AN

1. A. Kojima, K. Teshi Soc. 131, 6050–60
2. J.-H. Im, C.-R. Lee, 3, 4088–4093 (20
3. M. M. Lee, J. Teus

16. L. S. Mattos et al., in Conference Record of the 38th IEEE Photovoltaic Specialists Conference, Austin, TX, 3 to 8 June 2012 (IEEE, 2012), pp. 3187–3190.
17. O. D. Miller, E. Yablonovitch, S. R. Kurtz, IEEE J. Photovoltaics

ACKNOWLEDGMENTS

We acknowledge financial support from the Engineering and Physical Sciences Research Council of the UK (EPSRC) and King Abdulaziz City for Science and Technology. L.M.P.-Q. thanks the

nature human behaviour

LETTERS

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

Eval
exp
201

Colin F.
Magnus
Thomas
Anup G
Julia Ro

Being abl entific pr experime and Scien follow an pre-regist high powe higher tha in the sam ies, and th 50% of th (57%) an indicators positive ra size of tru both false contribute that peer bility, sug

Biological Psychiatry

Review

Imaging
A Critica

Ryan Bogdan, E
Hugh Garavan,
Ned H. Kalin, a

ABSTRACT

Imaging genetics neural phenotype approaches its t studies using sm phenotypes for originally posited many new directi novel methods (but also introdu highlight potentia translational and genetics, pharma

CORRESPONDENCE

LINK TO ORIGINAL ARTICLE

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

Florian Prinz, Thomas Schlange and Khusru Asadullah

A recent report by Arrowsmith noted that the success rates for new development projects in Phase II trials have fallen from 28% to 18% in recent years, with insufficient efficacy being the most frequent reason for failure (Phase II failures: 2008–2010. *Nature Rev. Drug Discov.* **10**, 328–329 (2011))[1]. This indicates the limitations of the predictivity of disease models and also that the validity of the targets being investigated is frequently questionable, which to 'feasible/marketable', and the financial costs of pursuing a full-blown drug discovery and development programme for a particular target could ultimately be hundreds of millions of Euros. Even in the earlier stages, investments in activities such as high-throughput screening programmes are substantial, and thus the validity of published data on potential targets is crucial for companies when deciding to start novel projects.

results that are published are hard to reproduce. However, there is an imbalance between this apparently widespread impression and its public recognition (for example, see REFS 2,3), and the surprisingly few scientific publications dealing with this topic. Indeed, to our knowledge, so far there has been no published in-depth, systematic analysis that compares reproduced results with published results for wet-lab experiments related to target identification and validation.

Early research in the pharmaceutical industry, with a dedicated budget and scientists who mainly work on target validation to increase the confidence in a project, provides a unique opportunity to generate a broad data set on the reproducibility of published data. To substantiate our incidental observations that published reports are frequently not reproducible with quantitative data, we performed an analysis of our early (target identification and valida-

Problems

- Publication bias/file drawer problem
- P-hacking/multiple comparisons/data dredging/data fishing
- The garden of forking paths
- Optional stopping (of data collection)
- Post hoc storytelling (framing exploratory analysis as confirmatory analysis)
- Perverse incentives and “publish or perish”
- Fraud

Proposed solutions

- Lower significance threshold
- Require/encourage higher power studies
- Require/encourage preregistration
- Require/encourage sample-splitting
- Publication of null results & data-sharing
- Meta-analysis
- Replication studies
- Adversarial collaboration
- Replace/augment statistical significance

Proposed solutions

- Lower significance threshold
- Require/encourage higher power studies
- Require/encourage preregistration
- Require/encourage sample-splitting
- Publication of null results & data-sharing
- Meta-analysis
- Replication studies
- Adversarial collaboration
- Replace/augment statistical significance

Soundness-dependent effects

Aim of this talk

- To share a dynamic—*soundness-dependence*—with implications for some of the proposed interventions: i.a., voluntary *preregistration*, *sample splitting*, increasing *study power*, and *redefining statistical significance*.

- *Emphatically not:* to argue against any of these interventions.

Fact 1. Scientific studies vary in their methodological soundness.

Fact 1. Scientific studies vary in their methodological soundness.

Fact 2. Interventions in evidentiary standards and research practices can differentially affect studies as a function of their soundness.

Claim. The conjunction of these facts has unrecognized implications for proposed interventions in the replication crisis.

EXAMPLE A

Intervention. Voluntary preregistration

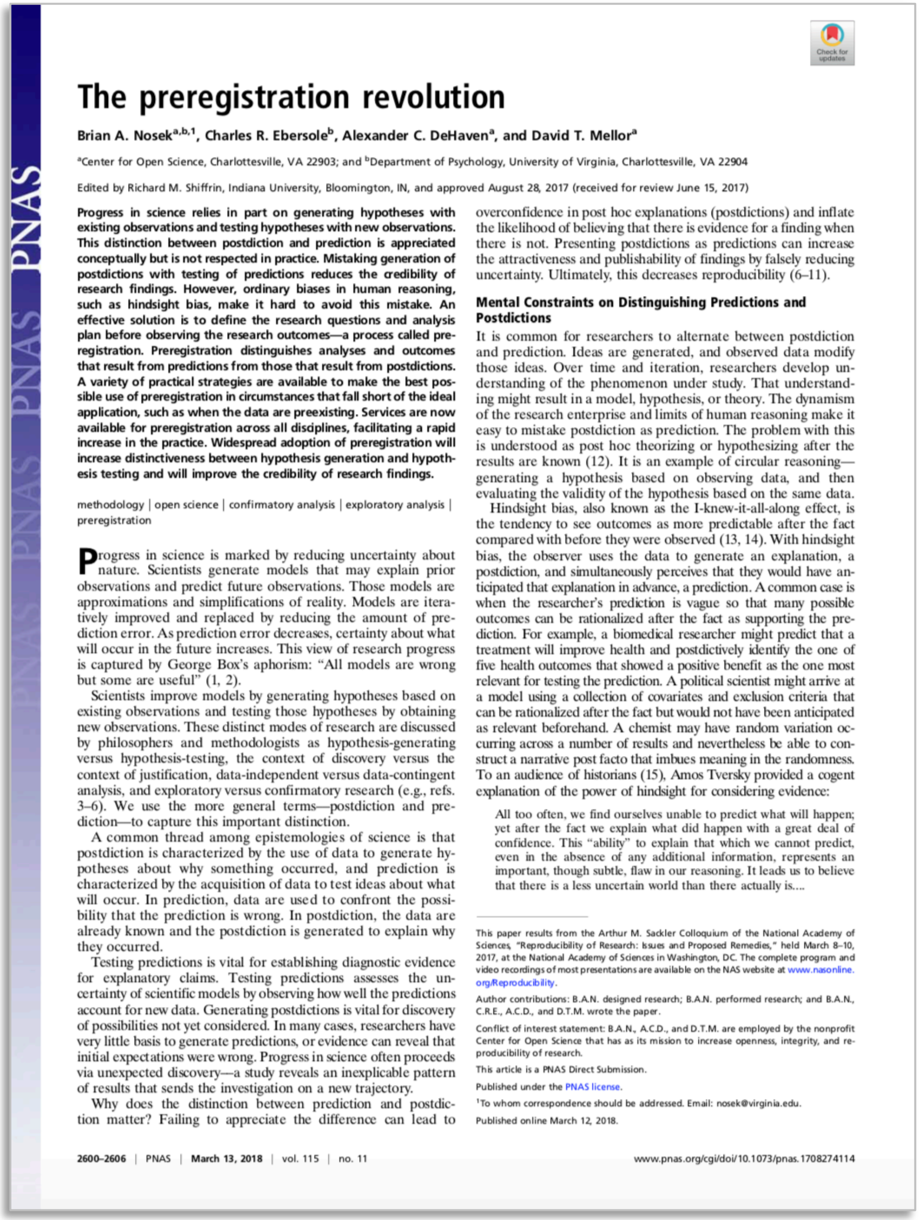
Background problem 1. Publication bias

Background problem 2. Questionable research practices producing differentials in study soundness

Observation. Voluntary preregistration and study soundness can be positively correlated

Preregistration

Specification of study design and analysis plan prior to conducting a study.



The preregistration revolution

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

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

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

Progress in science relies in part on generating hypotheses with existing observations and testing hypotheses with new observations. This distinction between postdiction and prediction is appreciated conceptually but is not respected in practice. Mistaking generation of postdictions with testing of predictions reduces the credibility of research findings. However, ordinary biases in human reasoning, such as hindsight bias, make it hard to avoid this mistake. An effective solution is to define the research questions and analysis plan before observing the research outcomes—a process called preregistration. Preregistration distinguishes analyses and outcomes that result from predictions from those that result from postdictions. A variety of practical strategies are available to make the best possible use of preregistration in circumstances that fall short of the ideal application, such as when the data are preexisting. Services are now available for preregistration across all disciplines, facilitating a rapid increase in the practice. Widespread adoption of preregistration will increase distinctiveness between hypothesis generation and hypothesis testing and will improve the credibility of research findings.

methodology | open science | confirmatory analysis | exploratory analysis | preregistration

Progress in science is marked by reducing uncertainty about nature. Scientists generate models that may explain prior observations and predict future observations. Those models are approximations and simplifications of reality. Models are iteratively improved and replaced by reducing the amount of prediction error. As prediction error decreases, certainty about what will occur in the future increases. This view of research progress is captured by George Box's aphorism: "All models are wrong but some are useful" (1, 2).

Scientists improve models by generating hypotheses based on existing observations and testing those hypotheses by obtaining new observations. These distinct modes of research are discussed by philosophers and methodologists as hypothesis-generating versus hypothesis-testing, the context of discovery versus the context of justification, data-independent versus data-contingent analysis, and exploratory versus confirmatory research (e.g., refs. 3–6). We use the more general terms—postdiction and prediction—to capture this important distinction.

A common thread among epistemologies of science is that postdiction is characterized by the use of data to generate hypotheses about why something occurred, and prediction is characterized by the acquisition of data to test ideas about what will occur. In prediction, data are used to confront the possibility that the prediction is wrong. In postdiction, the data are already known and the postdiction is generated to explain why they occurred.

Testing predictions is vital for establishing diagnostic evidence for explanatory claims. Testing predictions assesses the uncertainty of scientific models by observing how well the predictions account for new data. Generating postdictions is vital for discovery of possibilities not yet considered. In many cases, researchers have very little basis to generate predictions, or evidence can reveal that initial expectations were wrong. Progress in science often proceeds via unexpected discovery—a study reveals an inexplicable pattern of results that sends the investigation on a new trajectory.

Why does the distinction between prediction and postdiction matter? Failing to appreciate the difference can lead to overconfidence in post hoc explanations (postdictions) and inflate the likelihood of believing that there is evidence for a finding when there is not. Presenting postdictions as predictions can increase the attractiveness and publishability of findings by falsely reducing uncertainty. Ultimately, this decreases reproducibility (6–11).

Mental Constraints on Distinguishing Predictions and Postdictions

It is common for researchers to alternate between postdiction and prediction. Ideas are generated, and observed data modify those ideas. Over time and iteration, researchers develop understanding of the phenomenon under study. That understanding might result in a model, hypothesis, or theory. The dynamism of the research enterprise and limits of human reasoning make it easy to mistake postdiction as prediction. The problem with this is understood as post hoc theorizing or hypothesizing after the results are known (12). It is an example of circular reasoning—generating a hypothesis based on observing data, and then evaluating the validity of the hypothesis based on the same data.

Hindsight bias, also known as the I-knew-it-all-along effect, is the tendency to see outcomes as more predictable after the fact compared with before they were observed (13, 14). With hindsight bias, the observer uses the data to generate an explanation, a postdiction, and simultaneously perceives that they would have anticipated that explanation in advance, a prediction. A common case is when the researcher's prediction is vague so that many possible outcomes can be rationalized after the fact as supporting the prediction. For example, a biomedical researcher might predict that a treatment will improve health and postdictively identify the one of five health outcomes that showed a positive benefit as the one most relevant for testing the prediction. A political scientist might arrive at a model using a collection of covariates and exclusion criteria that can be rationalized after the fact but would not have been anticipated as relevant beforehand. A chemist may have random variation occurring across a number of results and nevertheless be able to construct a narrative post facto that imbues meaning in the randomness. To an audience of historians (15), Amos Tversky provided a cogent explanation of the power of hindsight for considering evidence:

> All too often, we find ourselves unable to predict what will happen; yet after the fact we explain what did happen with a great deal of confidence. This "ability" to explain that which we cannot predict, even in the absence of any additional information, represents an important, though subtle, flaw in our reasoning. It leads us to believe that there is a less uncertain world than there actually is....

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Reproducibility of Research: Issues and Proposed Remedies," held March 8–10, 2017, at the National Academy of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Reproducibility.

Author contributions: B.A.N. designed research; B.A.N. performed research; and B.A.N., C.R.E., A.C.D., and D.T.M. wrote the paper.

Conflict of interest statement: B.A.N., A.C.D., and D.T.M. are employed by the nonprofit Center for Open Science that has as its mission to increase openness, integrity, and reproducibility of research.

This article is a PNAS Direct Submission.

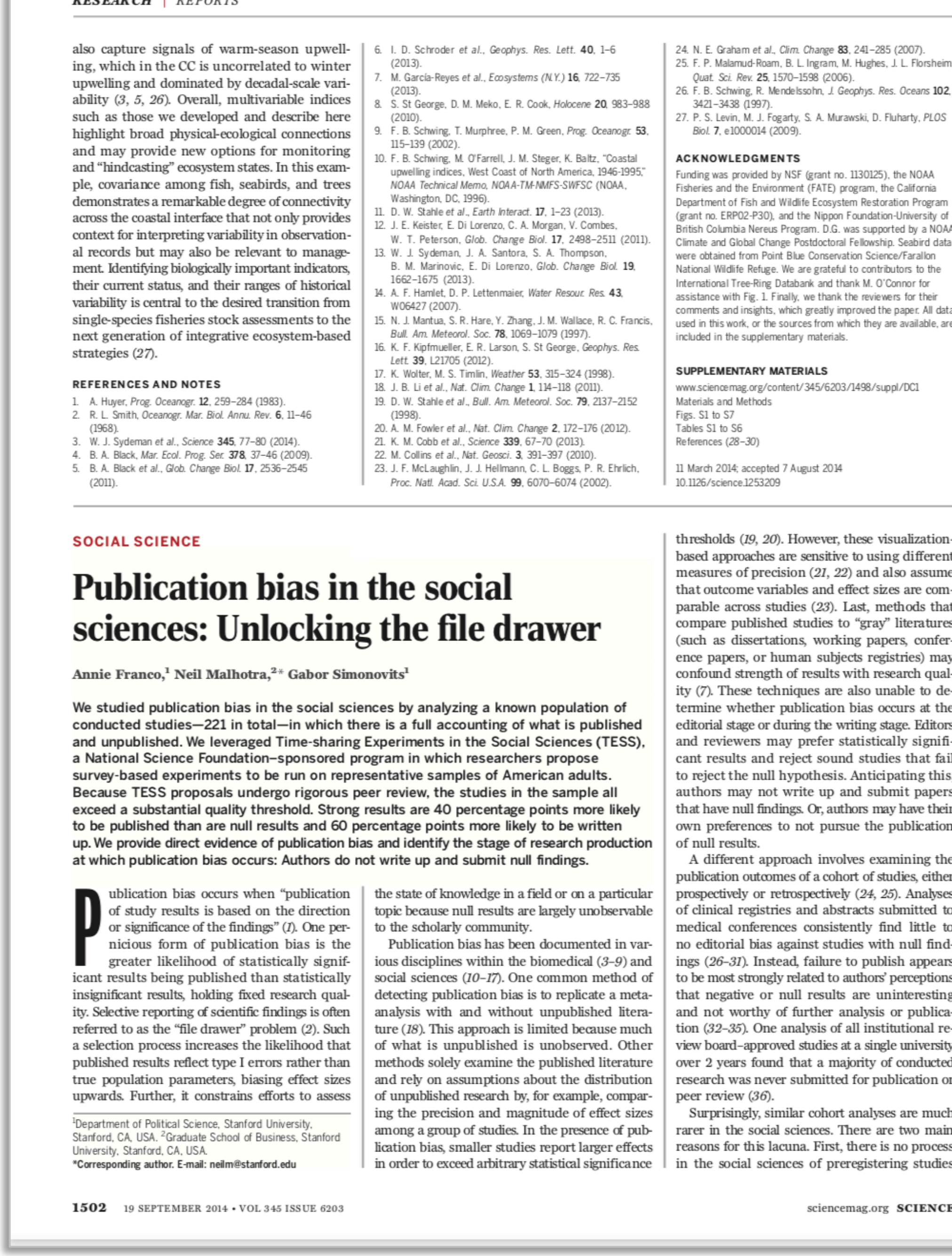
[1] To whom correspondence should be addressed. Email: nosek@virginia.edu.

Published online March 12, 2018.

www.pnas.org/cgi/doi/10.1073/pnas.1708274114

Publication bias

Results in a literature consisting largely in statistically significant (rather than null) results.



also capture signals of warm-season upwelling, which in the CC is uncorrelated to winter upwelling and dominated by decadal-scale variability (*3*, *5*, *26*). Overall, multivariable indices such as those we developed and describe here highlight broad physical-ecological connections and may provide new options for monitoring and "hindcasting" ecosystem states. In this example, covariance among fish, seabirds, and trees demonstrates a remarkable degree of connectivity across the coastal interface that not only provides context for interpreting variability in observational records but may also be relevant to management. Identifying biologically important indicators, their current status, and their ranges of historical variability is central to the desired transition from single-species fisheries stock assessments to the next generation of integrative ecosystem-based strategies (*27*).

REFERENCES AND NOTES

1. A. Huyer, *Prog. Oceanogr.* **12**, 259–284 (1983).
2. R. L. Smith, *Oceanogr. Mar. Biol. Annu. Rev.* **6**, 11–46 (1968).
3. W. J. Sydeman *et al.*, *Science* **345**, 77–80 (2014).
4. B. A. Black, *Mar. Ecol. Prog. Ser.* **378**, 37–46 (2009).
5. B. A. Black *et al.*, *Glob. Change Biol.* **17**, 2536–2545 (2011).
6. I. D. Schroder *et al.*, *Geophys. Res. Lett.* **40**, 1–6 (2013).
7. M. García-Reyes *et al.*, *Ecosystems (N.Y.)* **16**, 722–735 (2013).
8. S. St George, D. M. Meko, E. R. Cook, *Holocene* **20**, 983–988 (2010).
9. F. B. Schwing, T. Murphree, P. M. Green, *Prog. Oceanogr.* **53**, 115–139 (2002).
10. F. B. Schwing, M. O'Farrell, J. M. Steger, K. Baltz, "Coastal upwelling indices, West Coast of North America, 1946-1995," *NOAA Technical Memo, NOAA-TM-NMFS-SWFSC* (NOAA, Washington, DC, 1996).
11. D. W. Stahle *et al.*, *Earth Interact.* **17**, 1–23 (2013).
12. J. E. Keister, E. Di Lorenzo, C. A. Morgan, V. Combes, W. T. Peterson, *Glob. Change Biol.* **17**, 2498–2511 (2011).
13. W. J. Sydeman, J. A. Santora, S. A. Thompson, B. M. Marinovic, E. Di Lorenzo, *Glob. Change Biol.* **19**, 1662–1675 (2013).
14. A. F. Hamlet, D. P. Lettenmaier, *Water Resour. Res.* **43**, W06427 (2007).
15. N. J. Mantua, S. R. Hare, Y. Zhang, J. M. Wallace, R. C. Francis, *Bull. Am. Meteorol. Soc.* **78**, 1069–1079 (1997).
16. K. F. Kipfmueller, E. R. Larson, S. St George, *Geophys. Res. Lett.* **39**, L21705 (2012).
17. K. Wolter, M. S. Timlin, *Weather* **53**, 315–324 (1998).
18. J. B. Li *et al.*, *Nat. Clim. Change* **1**, 114–118 (2011).
19. D. W. Stahle *et al.*, *Bull. Am. Meteorol. Soc.* **79**, 2137–2152 (1998).
20. A. M. Fowler *et al.*, *Nat. Clim. Change* **2**, 172–176 (2012).
21. K. M. Cobb *et al.*, *Science* **339**, 67–70 (2013).
22. M. Collins *et al.*, *Nat. Geosci.* **3**, 391–397 (2010).
23. J. F. McLaughlin, J. J. Hellmann, C. L. Boggs, P. R. Ehrlich, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 6070–6074 (2002).
24. N. E. Graham *et al.*, *Clim. Change* **83**, 241–285 (2007).
25. F. P. Malamud-Roam, B. L. Ingram, M. Hughes, J. L. Florsheim, *Quat. Sci. Rev.* **25**, 1570–1598 (2006).
26. F. B. Schwing, R. Mendelssohn, *J. Geophys. Res. Oceans* **102**, 3421–3438 (1997).
27. P. S. Levin, M. J. Fogarty, S. A. Murawski, D. Fluharty, *PLOS Biol.* **7**, e1000014 (2009).

ACKNOWLEDGMENTS

Funding was provided by NSF (grant no. 1130125), the NOAA Fisheries and the Environment (FATE) program, the California Department of Fish and Wildlife Ecosystem Restoration Program (grant no. ERP02-P30), and the Nippon Foundation-University of British Columbia Nereus Program. D.G. was supported by a NOAA Climate and Global Change Postdoctoral Fellowship. Seabird data were obtained from Point Blue Conservation Science/Farallon National Wildlife Refuge. We are grateful to contributors to the International Tree-Ring Databank and thank M. O'Connor for assistance with Fig. 1. Finally, we thank the reviewers for their comments and insights, which greatly improved the paper. All data used in this work, or the sources from which they are available, are included in the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/345/6203/1498/suppl/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S6
References (*28–30*)

11 March 2014; accepted 7 August 2014
10.1126/science.1253209

SOCIAL SCIENCE

Publication bias in the social sciences: Unlocking the file drawer

Annie Franco,[1] Neil Malhotra,[2]* Gabor Simonovits[1]

We studied publication bias in the social sciences by analyzing a known population of conducted studies—221 in total—in which there is a full accounting of what is published and unpublished. We leveraged Time-sharing Experiments in the Social Sciences (TESS), a National Science Foundation–sponsored program in which researchers propose survey-based experiments to be run on representative samples of American adults. Because TESS proposals undergo rigorous peer review, the studies in the sample all exceed a substantial quality threshold. Strong results are 40 percentage points more likely to be published than are null results and 60 percentage points more likely to be written up. We provide direct evidence of publication bias and identify the stage of research production at which publication bias occurs: Authors do not write up and submit null findings.

Publication bias occurs when "publication of study results is based on the direction or significance of the findings" (*1*). One pernicious form of publication bias is the greater likelihood of statistically significant results being published than statistically insignificant results, holding fixed research quality. Selective reporting of scientific findings is often referred to as the "file drawer" problem (*2*). Such a selection process increases the likelihood that published results reflect type I errors rather than true population parameters, biasing effect sizes upwards. Further, it constrains efforts to assess the state of knowledge in a field or on a particular topic because null results are largely unobservable to the scholarly community.

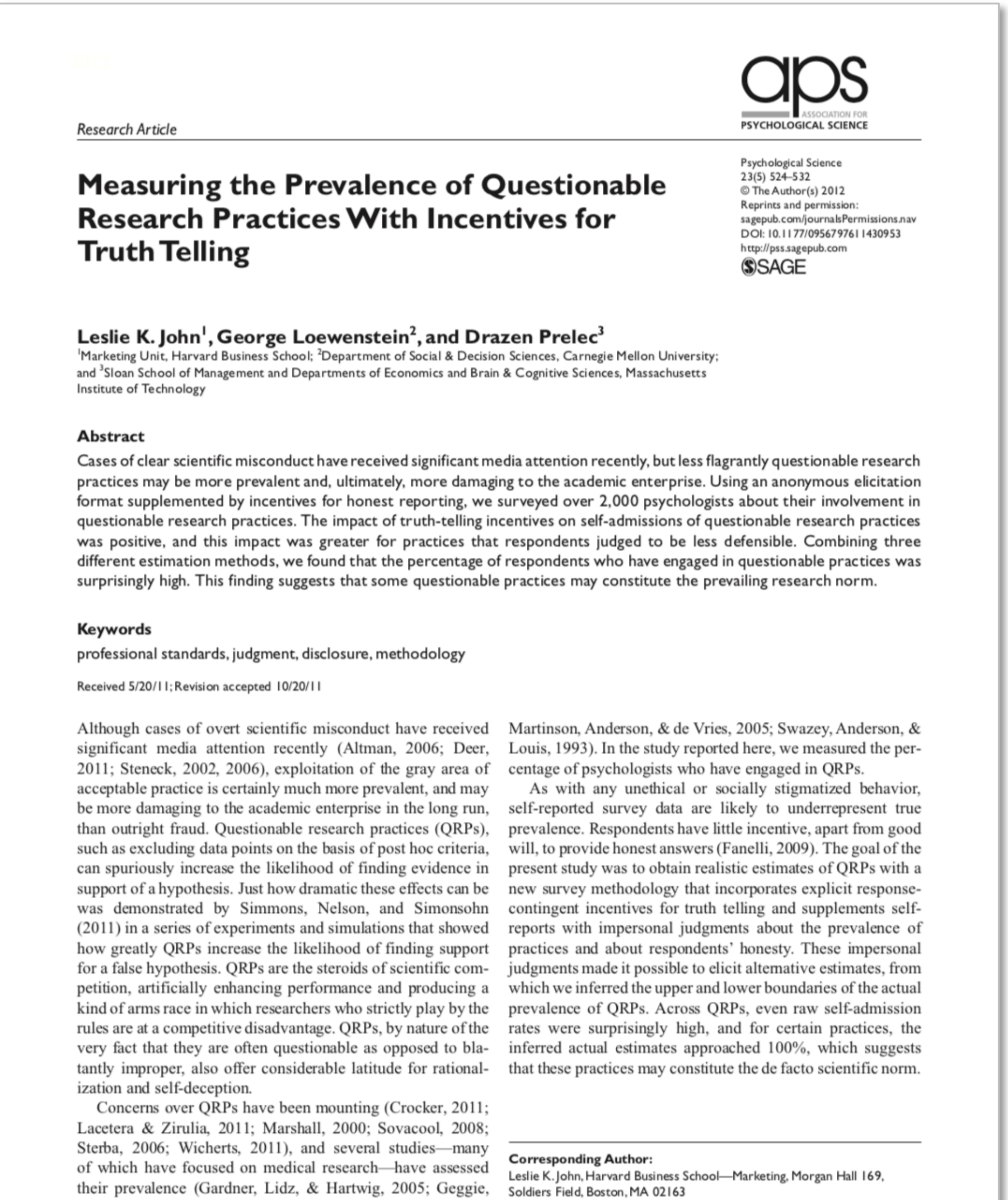
Publication bias has been documented in various disciplines within the biomedical (*3–9*) and social sciences (*10–17*). One common method of detecting publication bias is to replicate a meta-analysis with and without unpublished literature (*18*). This approach is limited because much of what is unpublished is unobserved. Other methods solely examine the published literature and rely on assumptions about the distribution of unpublished research by, for example, comparing the precision and magnitude of effect sizes among a group of studies. In the presence of publication bias, smaller studies report larger effects in order to exceed arbitrary statistical significance thresholds (*19*, *20*). However, these visualization-based approaches are sensitive to using different measures of precision (*21*, *22*) and also assume that outcome variables and effect sizes are comparable across studies (*23*). Last, methods that compare published studies to "gray" literatures (such as dissertations, working papers, conference papers, or human subjects registries) may confound strength of results with research quality (*7*). These techniques are also unable to determine whether publication bias occurs at the editorial stage or during the writing stage. Editors and reviewers may prefer statistically significant results and reject sound studies that fail to reject the null hypothesis. Anticipating this, authors may not write up and submit papers that have null findings. Or, authors may have their own preferences to not pursue the publication of null results.

A different approach involves examining the publication outcomes of a cohort of studies, either prospectively or retrospectively (*24*, *25*). Analyses of clinical registries and abstracts submitted to medical conferences consistently find little to no editorial bias against studies with null findings (*26–31*). Instead, failure to publish appears to be most strongly related to authors' perceptions that negative or null results are uninteresting and not worthy of further analysis or publication (*32–35*). One analysis of all institutional review board–approved studies at a single university over 2 years found that a majority of conducted research was never submitted for publication or peer review (*36*).

Surprisingly, similar cohort analyses are much rarer in the social sciences. There are two main reasons for this lacuna. First, there is no process in the social sciences of preregistering studies

[1] Department of Political Science, Stanford University, Stanford, CA, USA. [2] Graduate School of Business, Stanford University, Stanford, CA, USA.
*Corresponding author. E-mail: neilm@stanford.edu

QRPs



Research Article

aps
ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
23(5) 524–532
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797611430953
http://pss.sagepub.com
SAGE

Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling

Leslie K. John[1], George Loewenstein[2], and Drazen Prelec[3]

[1]Marketing Unit, Harvard Business School; [2]Department of Social & Decision Sciences, Carnegie Mellon University; and [3]Sloan School of Management and Departments of Economics and Brain & Cognitive Sciences, Massachusetts Institute of Technology

Abstract

Cases of clear scientific misconduct have received significant media attention recently, but less flagrantly questionable research practices may be more prevalent and, ultimately, more damaging to the academic enterprise. Using an anonymous elicitation format supplemented by incentives for honest reporting, we surveyed over 2,000 psychologists about their involvement in questionable research practices. The impact of truth-telling incentives on self-admissions of questionable research practices was positive, and this impact was greater for practices that respondents judged to be less defensible. Combining three different estimation methods, we found that the percentage of respondents who have engaged in questionable practices was surprisingly high. This finding suggests that some questionable practices may constitute the prevailing research norm.

Keywords

professional standards, judgment, disclosure, methodology

Received 5/20/11; Revision accepted 10/20/11

Although cases of overt scientific misconduct have received significant media attention recently (Altman, 2006; Deer, 2011; Steneck, 2002, 2006), exploitation of the gray area of acceptable practice is certainly much more prevalent, and may be more damaging to the academic enterprise in the long run, than outright fraud. Questionable research practices (QRPs), such as excluding data points on the basis of post hoc criteria, can spuriously increase the likelihood of finding evidence in support of a hypothesis. Just how dramatic these effects can be was demonstrated by Simmons, Nelson, and Simonsohn (2011) in a series of experiments and simulations that showed how greatly QRPs increase the likelihood of finding support for a false hypothesis. QRPs are the steroids of scientific competition, artificially enhancing performance and producing a kind of arms race in which researchers who strictly play by the rules are at a competitive disadvantage. QRPs, by nature of the very fact that they are often questionable as opposed to blatantly improper, also offer considerable latitude for rationalization and self-deception.

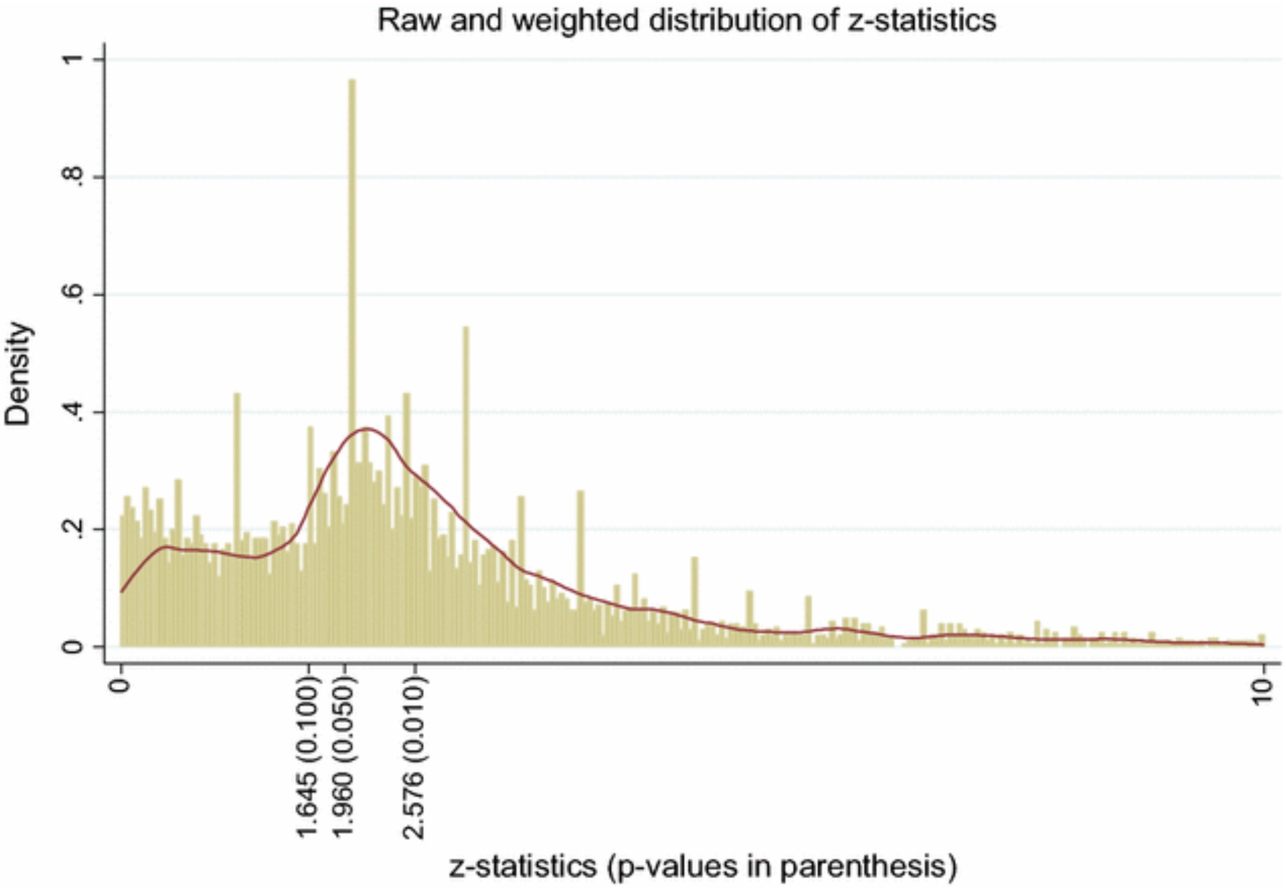
Concerns over QRPs have been mounting (Crocker, 2011; Lacetera & Zirulia, 2011; Marshall, 2000; Sovacool, 2008; Sterba, 2006; Wicherts, 2011), and several studies—many of which have focused on medical research—have assessed their prevalence (Gardner, Lidz, & Hartwig, 2005; Geggie, Martinson, Anderson, & de Vries, 2005; Swazey, Anderson, & Louis, 1993). In the study reported here, we measured the percentage of psychologists who have engaged in QRPs.

As with any unethical or socially stigmatized behavior, self-reported survey data are likely to underrepresent true prevalence. Respondents have little incentive, apart from good will, to provide honest answers (Fanelli, 2009). The goal of the present study was to obtain realistic estimates of QRPs with a new survey methodology that incorporates explicit response-contingent incentives for truth telling and supplements self-reports with impersonal judgments about the prevalence of practices and about respondents' honesty. These impersonal judgments made it possible to elicit alternative estimates, from which we inferred the upper and lower boundaries of the actual prevalence of QRPs. Across QRPs, even raw self-admission rates were surprisingly high, and for certain practices, the inferred actual estimates approached 100%, which suggests that these practices may constitute the de facto scientific norm.

Corresponding Author:
Leslie K. John, Harvard Business School—Marketing, Morgan Hall 169, Soldiers Field, Boston, MA 02163

QRPs produce *methodological bias* such that more studies obtain positive results when the null hypothesis is true.

QRPs


Raw and weighted distribution of z-statistics
Density
0
.2
.4
.6
.8
1
0
1.645 (0.100)
1.960 (0.050)
2.576 (0.010)
10
z-statistics (p-values in parenthesis)

The model

Consider a unit mass of independent studies.

A *study* consists of a hypothesis tests between a pair of null (H_0) and alternative (H_A) hypotheses.

Let $\phi \in (0, 1)$ denote the *fraction* of null hypotheses that are true.

Adaptation of a classic model of bias in science (Ioannidis, 2005; Maniadis, Tufano and List, 2014)

False discovery

A *false positive* occurs when a study yields a significant outcome when the null hypothesis is true.

The *false positive ratio* is equal to the number of false positive outcomes over the number of all significant outcomes.

$$Pr(H_0 \mid \text{significant}) = \frac{Pr(H_0, \text{significant})}{Pr(\text{significant})}$$

Methodological soundness

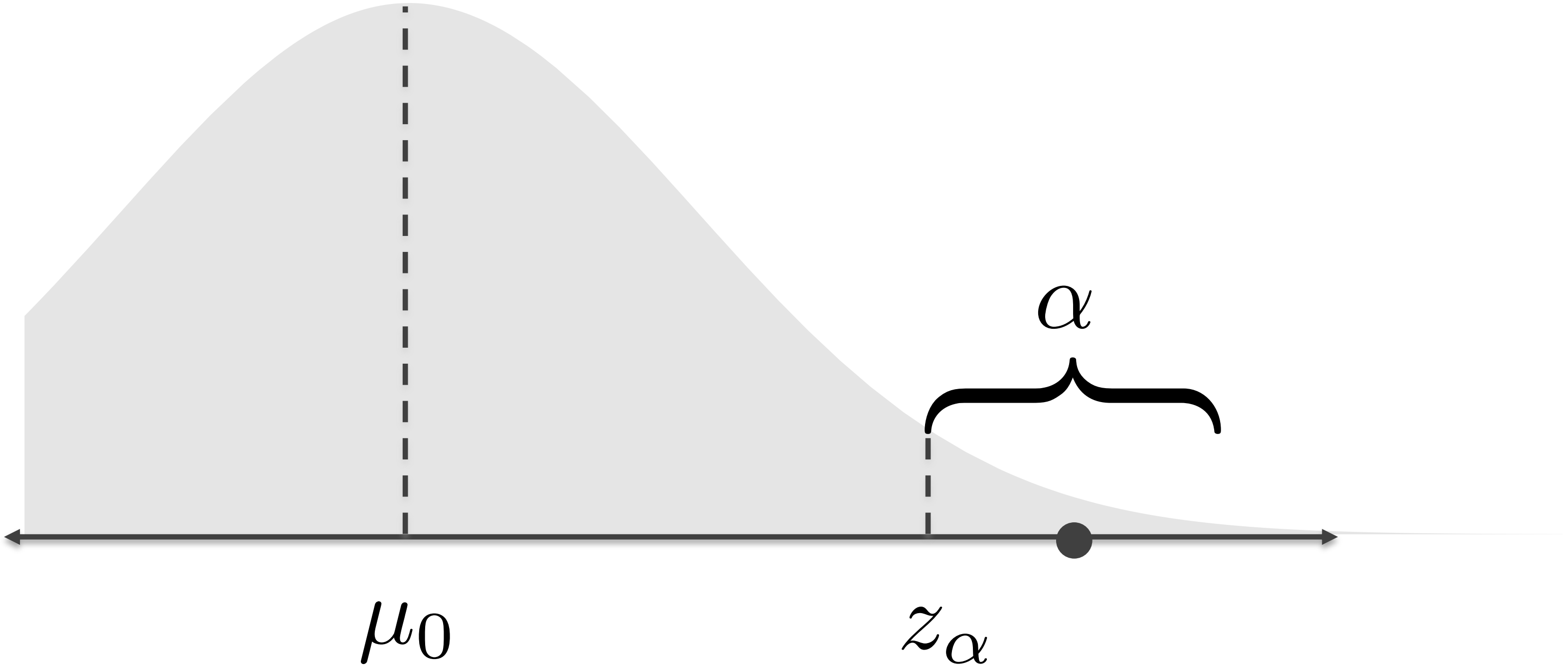
The population of studies exhibits some distribution of methodological soundness.

$$f_S$$

Assumption. Full support.

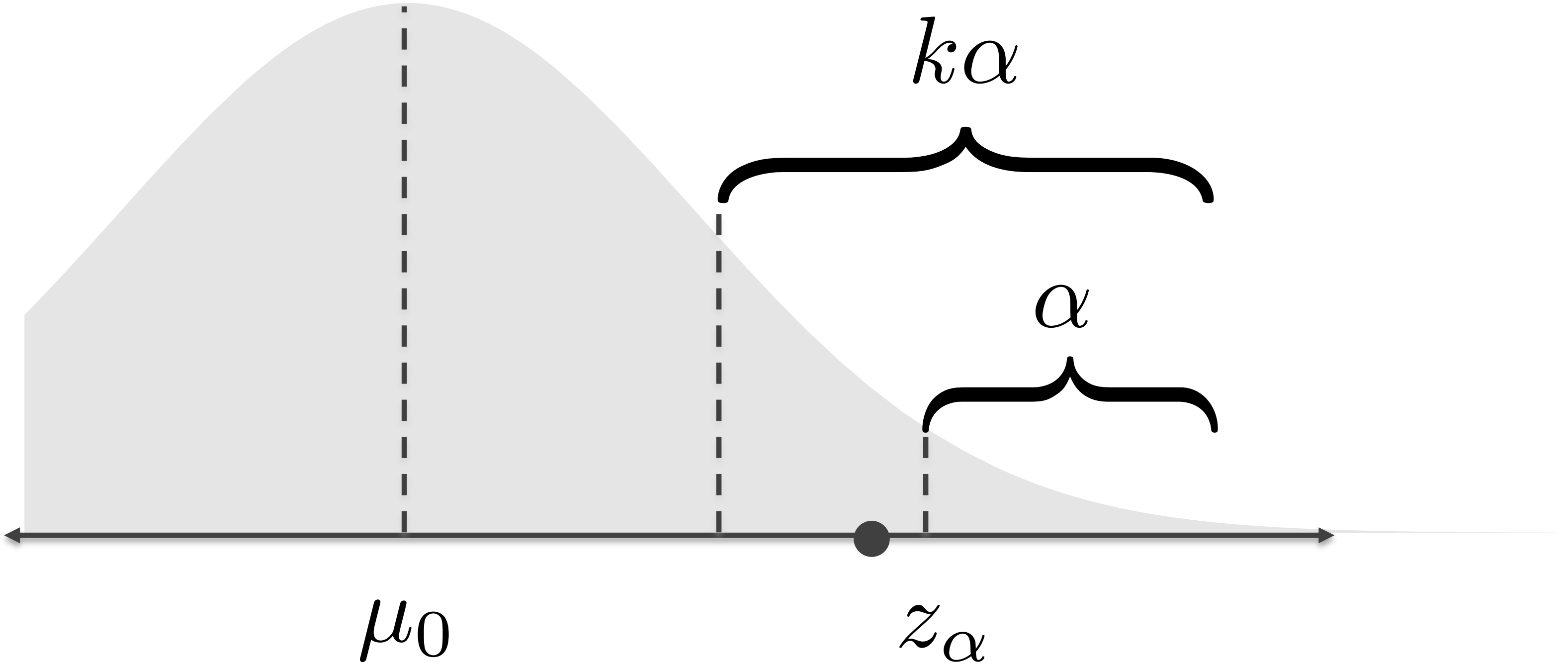
Methodological soundness

A *perfectly-sound* study attains significance when the null hypothesis is true with a frequency corresponding to the significance threshold.



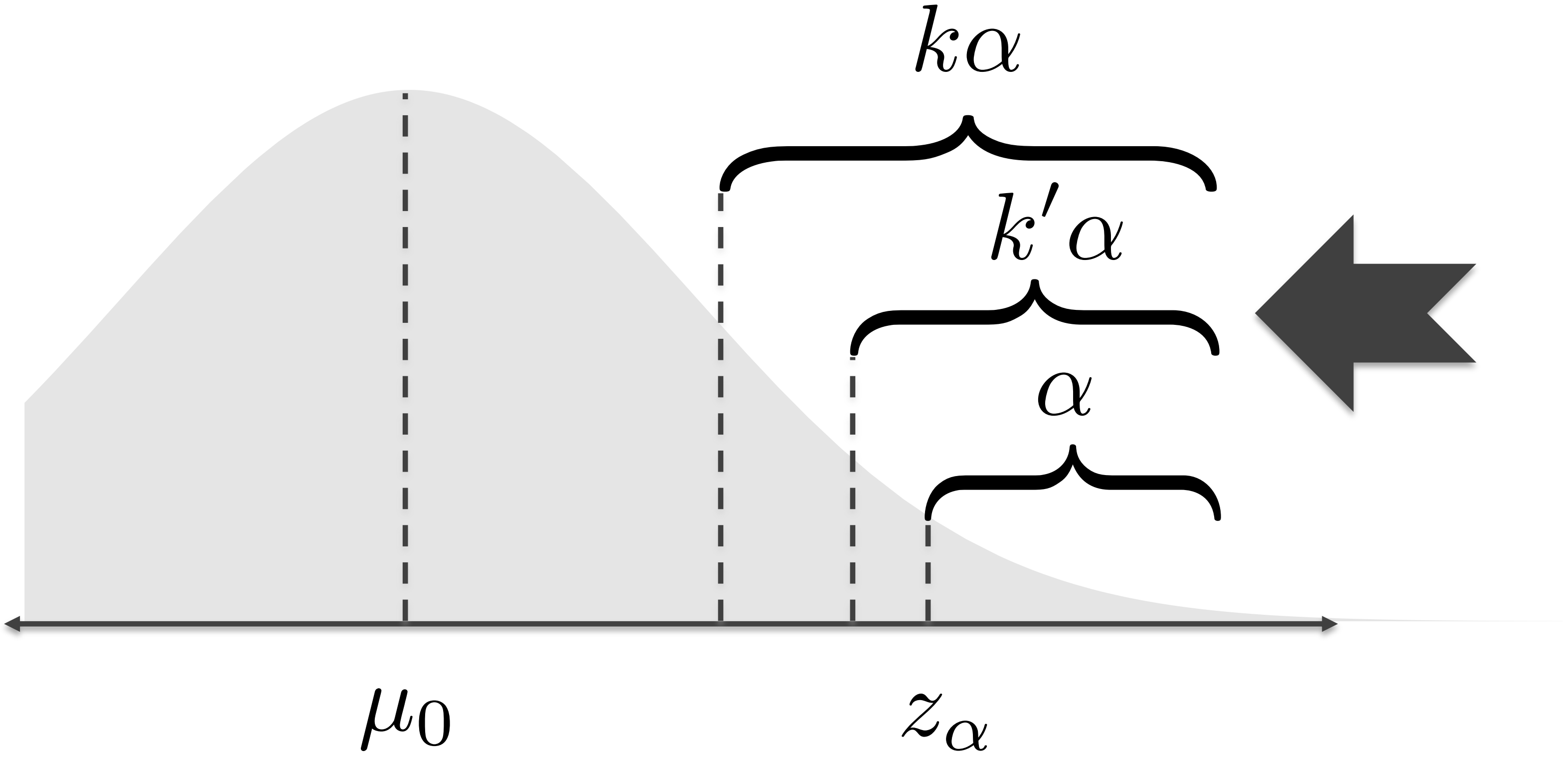
Methodological soundness

A *less-than-perfectly-sound* study will attain significance when the null hypothesis is true with a frequency *greater* than the significance threshold by some factor $k > 1$.



Modeling preregistration

When a study is *preregistered* we model this as having its unsoundness k decreased to k' nearer the optimal value such that $k > k' \geq 1$.



Study-level effects of preregistration

1. Preregistered studies produce *fewer false positive outcomes*.

2. Preregistered studies produce *fewer positive outcomes* as well.

A population-level problem

Voluntary preregistration can make it so:

A population-level problem

Voluntary preregistration can make it so:

Assumption 1. Researchers who are, on average, more likely to preregister their studies are also more likely to already employ more-sound methods.

A population-level problem

Voluntary preregistration can make it so:

Assumption 1. Researchers who are, on average, more likely to preregister their studies are also more likely to already employ more-sound methods.

This will produce *soundness-dependent filtration* of studies that can diminish (or reverse) desired improvement.

Proposition 1. If preregistration and study soundness are sufficiently strongly correlated, then *voluntary preregistration* will increase the false discovery rate of a literature.

How this works

For simplicity, assume there are two types of studies:

Relatively *sound studies* with $k = 2$.

(So, at $a = 0.05$, $ka = 0.1$)

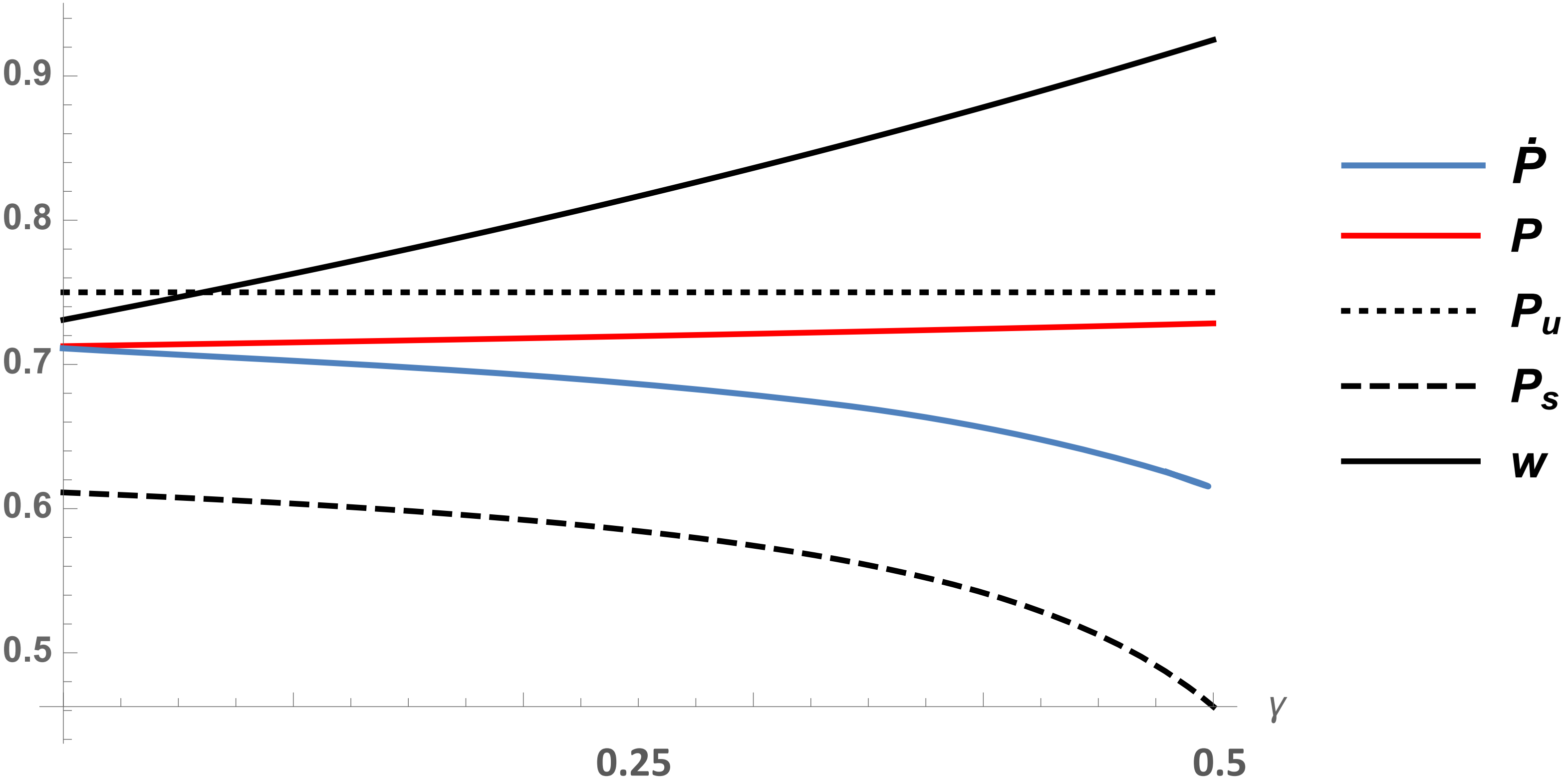
Unsound studies that always attain significance.

Stipulate that only sound studies are likely to be preregistered.

Preregistration of a study improves its soundness to the maximal value $k' = 1$.

Some reasonable further assumptions: studies are moderately powered with medium average effect sizes $d = 0.5$; and three-quarters of null hypotheses are true.

Effect of *soundness–preregistration* correlation on false discovery



Assumption 2. Researchers who are more confident in the truth of their studies are, on average, more likely to preregister their studies.

Assumption 3. Researchers' confidence in the truth of their study hypotheses is, on average, positively correlated with the truth of their hypotheses.

Proposition 2. If preregistration and prior odds of study hypotheses are sufficiently strongly related, then *voluntary preregistration* will increase the false discovery rate of a literature.

How this works

For simplicity, assume there are two types of studies:

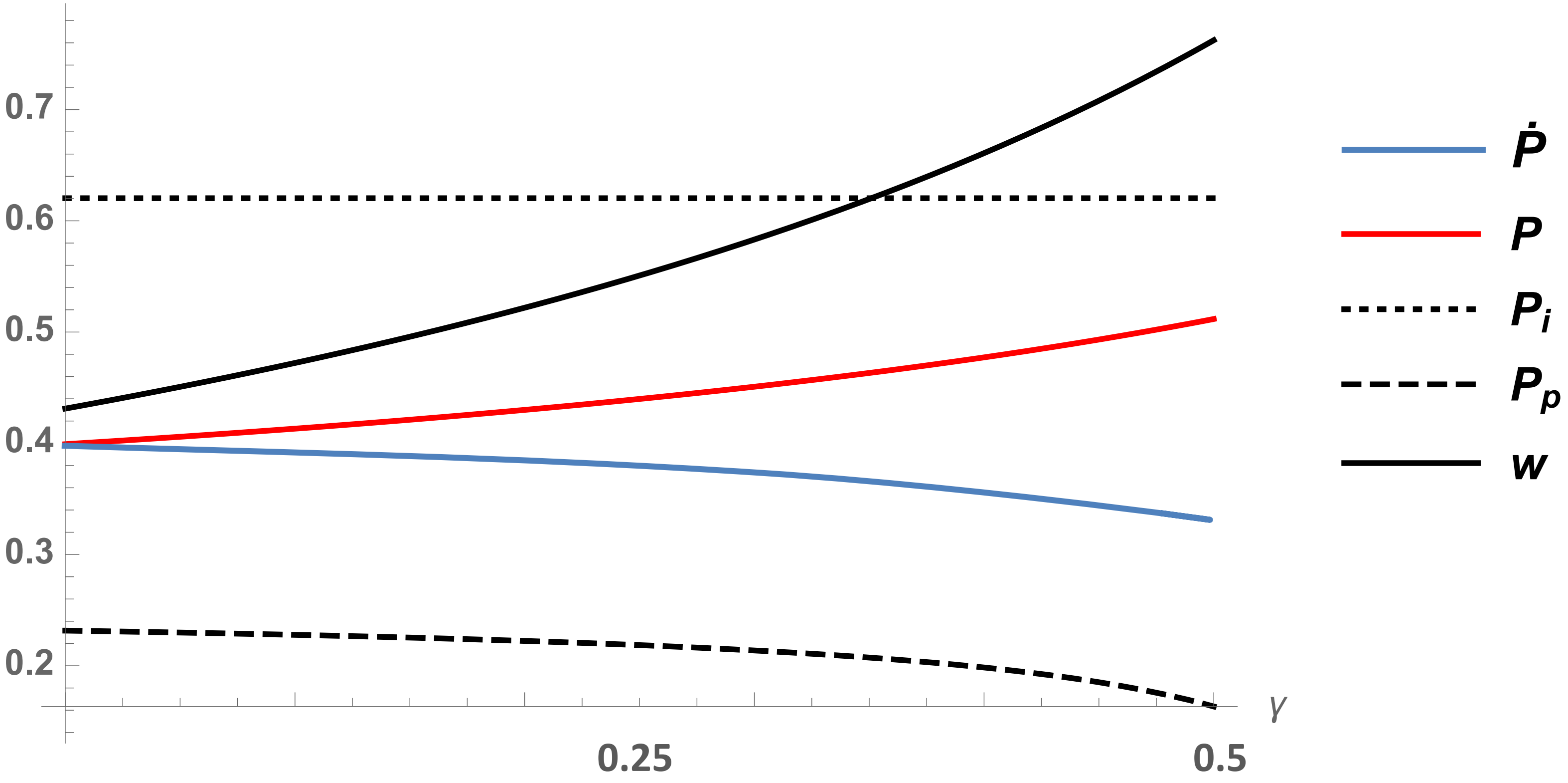
Plausible studies where, on average, 1/3 of the null hypotheses are true, and the studies are well-powered with medium average effect sizes $d = 0.5$.

Implausible studies where, on average, 2/3 of the null hypotheses are true, and the studies are modestly-powered with small average effect sizes $d = 0.2$.

Stipulate that only plausible studies are likely to be preregistered.

Both types of studies are equally sound with $k = 4$, and preregistration of a study improves its soundness to the maximal value $k' = 1$.

Effect of *confidence–preregistration* correlation on false discovery

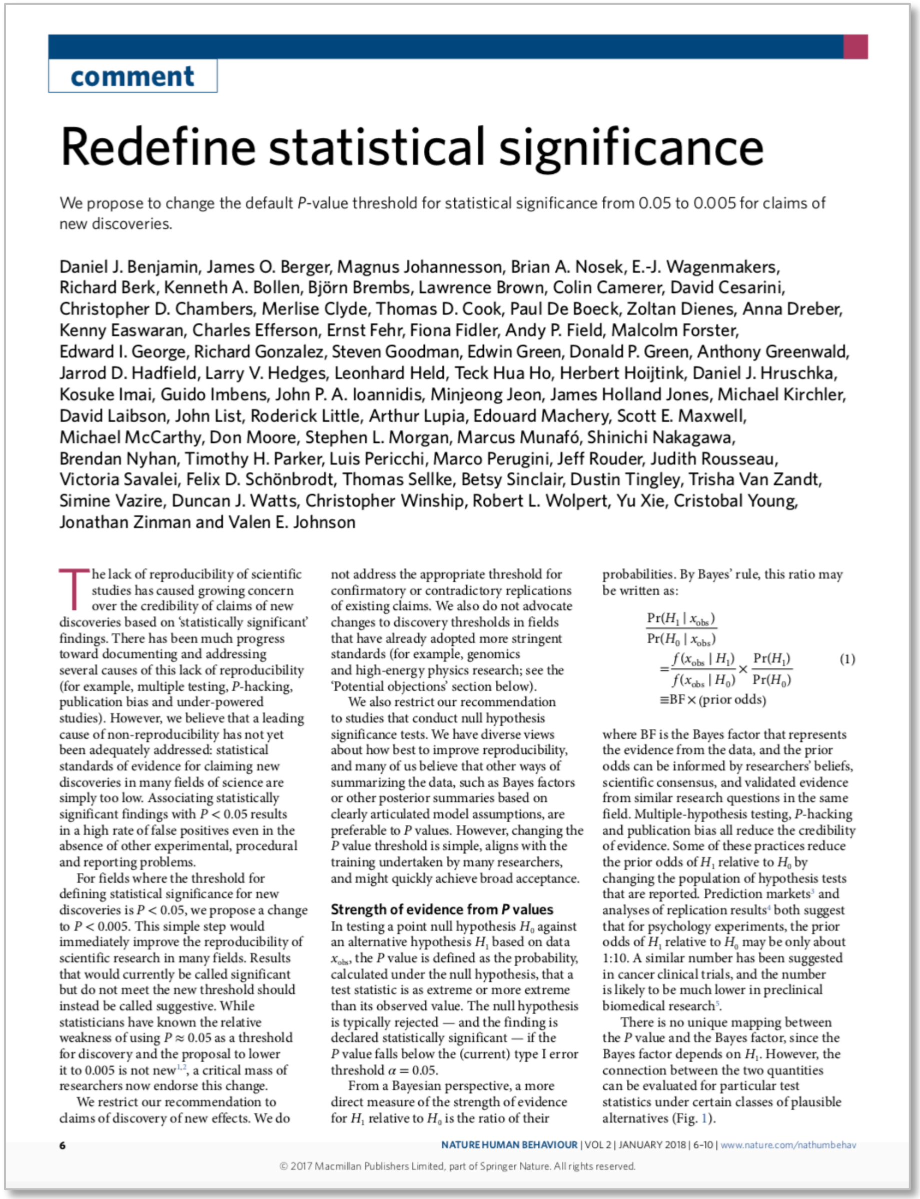


Analogous propositions on the overestimation of improvement and the possibility of backfire effects obtain for each:

- voluntary *sample-splitting*, and
- voluntary *increase in study power*.

Redefining statistical significance

Proposal to change the default p-value threshold for statistical significance from 0.05 to 0.005.



comment

Redefine statistical significance

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

The lack of reproducibility of scientific studies has caused growing concern over the credibility of claims of new discoveries based on 'statistically significant' findings. There has been much progress toward documenting and addressing several causes of this lack of reproducibility (for example, multiple testing, *P*-hacking, publication bias and under-powered studies). However, we believe that a leading cause of non-reproducibility has not yet been adequately addressed: statistical standards of evidence for claiming new discoveries in many fields of science are simply too low. Associating statistically significant findings with $P < 0.05$ results in a high rate of false positives even in the absence of other experimental, procedural and reporting problems.

For fields where the threshold for defining statistical significance for new discoveries is $P < 0.05$, we propose a change to $P < 0.005$. This simple step would immediately improve the reproducibility of scientific research in many fields. Results that would currently be called significant but do not meet the new threshold should instead be called suggestive. While statisticians have known the relative weakness of using $P \approx 0.05$ as a threshold for discovery and the proposal to lower it to 0.005 is not new[1,2], a critical mass of researchers now endorse this change.

We restrict our recommendation to claims of discovery of new effects. We do not address the appropriate threshold for confirmatory or contradictory replications of existing claims. We also do not advocate changes to discovery thresholds in fields that have already adopted more stringent standards (for example, genomics and high-energy physics research; see the 'Potential objections' section below).

We also restrict our recommendation to studies that conduct null hypothesis significance tests. We have diverse views about how best to improve reproducibility, and many of us believe that other ways of summarizing the data, such as Bayes factors or other posterior summaries based on clearly articulated model assumptions, are preferable to *P* values. However, changing the *P* value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance.

Strength of evidence from *P* values

In testing a point null hypothesis H_0 against an alternative hypothesis H_1 based on data x_{obs}, the *P* value is defined as the probability, calculated under the null hypothesis, that a test statistic is as extreme or more extreme than its observed value. The null hypothesis is typically rejected — and the finding is declared statistically significant — if the *P* value falls below the (current) type I error threshold $\alpha = 0.05$.

From a Bayesian perspective, a more direct measure of the strength of evidence for H_1 relative to H_0 is the ratio of their probabilities. By Bayes' rule, this ratio may be written as:

$$\begin{aligned}&\frac{\Pr(H_1 \mid x_{\mathrm{obs}})}{\Pr(H_0 \mid x_{\mathrm{obs}})}\\ &\quad = \frac{f(x_{\mathrm{obs}} \mid H_1)}{f(x_{\mathrm{obs}} \mid H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)}\\ &\quad \equiv \mathrm{BF} \times (\text{prior odds})\end{aligned} \tag{1}$$

where BF is the Bayes factor that represents the evidence from the data, and the prior odds can be informed by researchers' beliefs, scientific consensus, and validated evidence from similar research questions in the same field. Multiple-hypothesis testing, *P*-hacking and publication bias all reduce the credibility of evidence. Some of these practices reduce the prior odds of H_1 relative to H_0 by changing the population of hypothesis tests that are reported. Prediction markets[3] and analyses of replication results[4] both suggest that for psychology experiments, the prior odds of H_1 relative to H_0 may be only about 1:10. A similar number has been suggested in cancer clinical trials, and the number is likely to be much lower in preclinical biomedical research[5].

There is no unique mapping between the *P* value and the Bayes factor, since the Bayes factor depends on H_1. However, the connection between the two quantities can be evaluated for particular test statistics under certain classes of plausible alternatives (Fig. 1).

Proposition 3. If the distribution of methodological soundness of studies exhibits full support, then there will be a critical value beyond which lowering the statistical significance threshold will increase in the false discovery rate of a literature.

How this works

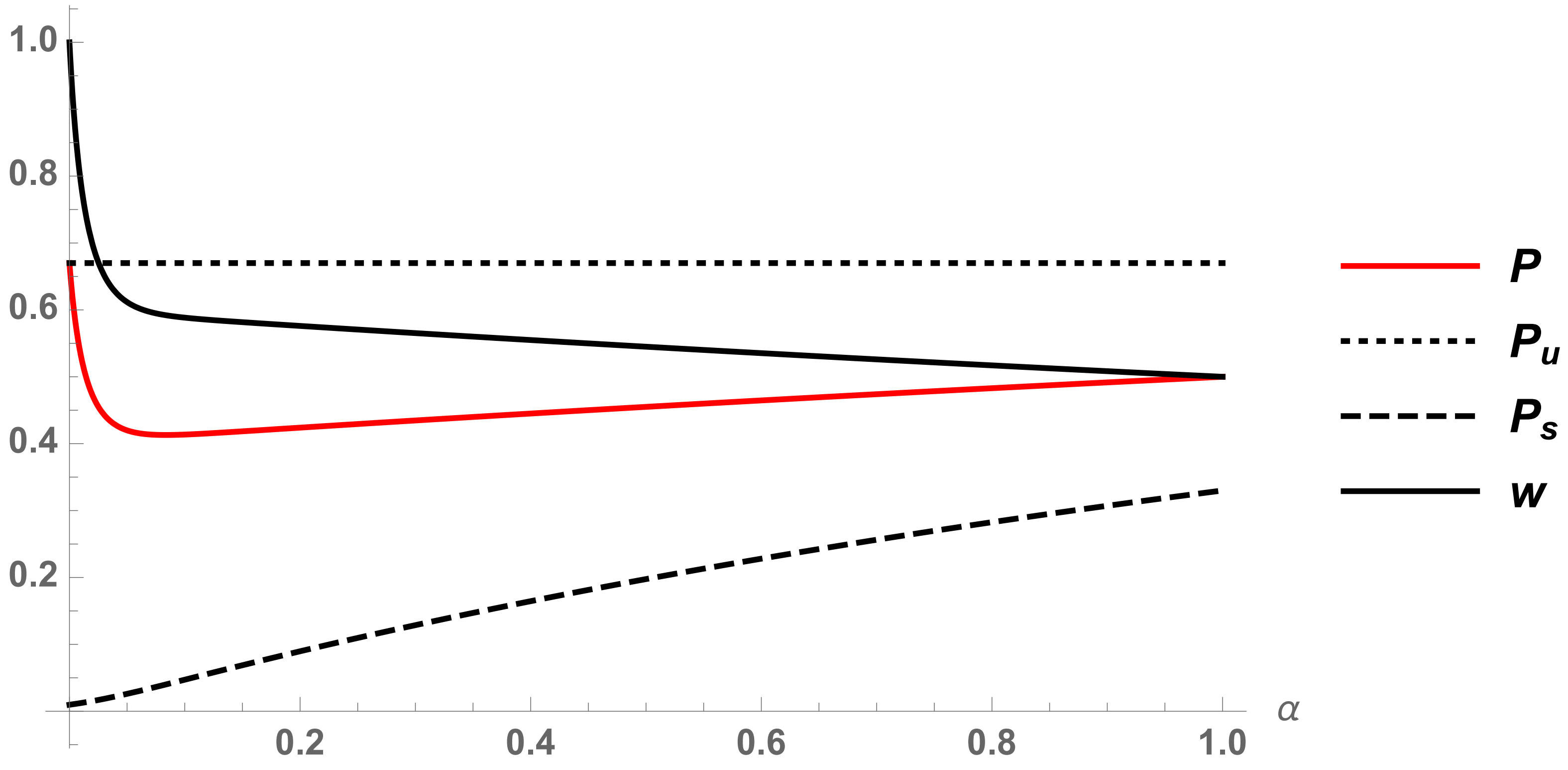
For simplicity, assume there are two types of studies:

Sound studies with $k = 1$.

Unsound studies that always attain significance.

Further assumptions: studies are moderately powered with medium average effect sizes $d = 0.5$; and two-thirds of null hypotheses are true.

Effect of *soundness-dependence and* on false discovery



Discussion

- Soundness-dependent filtration is *always at play*. E.g., when some researchers choose to employ higher power methods.

- Soundness-dependent filtration has implications for the natural selection of bad science (Cf. Smaldino 2016; Grimes et al 2017).

- A necessary condition for soundness-dependent filtration is *publication bias*.

- (Aside: subtle effects on *sign & magnitude errors*.)

Discussion

- Consider soundness-dependent effects in choosing between interventions (and in choosing b/w *voluntary & mandatory* versions thereof).

- Assess and *report replication rates* of studies in a literature *by method* (rather than in aggregate).

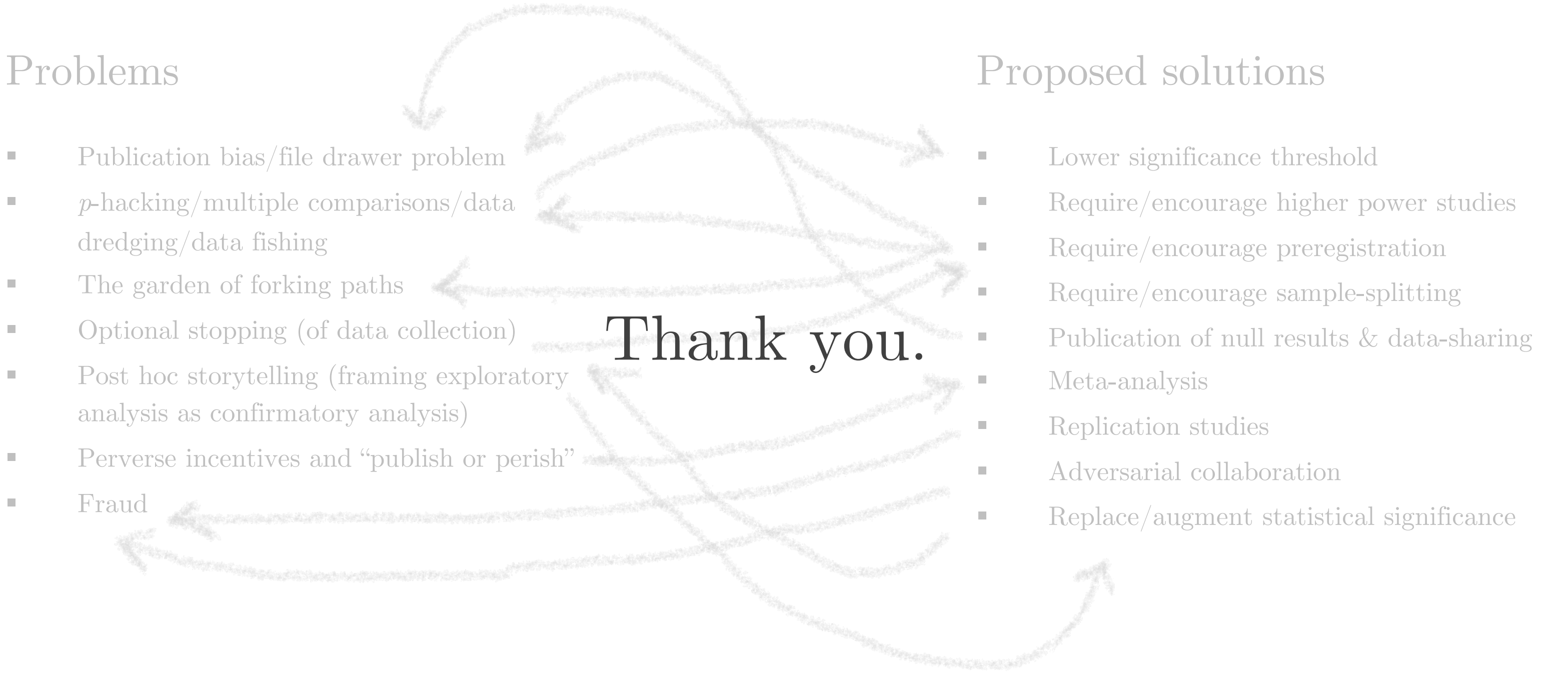
- *Adjust expectations* for improvement from interventions that will likely produce soundness-dependent filtration.

Moral

Interventions in evidentiary standards and research practices *need not affect all studies equally*.

Indeed, whenever the effects of interventions are *not entirely uncorrelated* with study soundness, we should be alert to the interaction between the two.

Thank you.



Problems

- Publication bias/file drawer problem
- p-hacking/multiple comparisons/data dredging/data fishing
- The garden of forking paths
- Optional stopping (of data collection)
- Post hoc storytelling (framing exploratory analysis as confirmatory analysis)
- Perverse incentives and “publish or perish”
- Fraud

Proposed solutions

- Lower significance threshold
- Require/encourage higher power studies
- Require/encourage preregistration
- Require/encourage sample-splitting
- Publication of null results & data-sharing
- Meta-analysis
- Replication studies
- Adversarial collaboration
- Replace/augment statistical significance