Interventions & Soundness-Dependence

Aydin Mohseni

Carnegie Mellon University
Department of Philosophy





A puzzle

A puzzle



Top scientist: We will find the 'God particle'

The "father" of an elusive subatomic particle said Monday he is almost sure it will be discovered in the next year in a race between powerful research equipment in the United States and Europe.

British physicist Perer Higgs maintains that particle known as "Higgs bosoen" will be found when the Large Hadron Collider, where he is shown here, starts fully operating early next year. Alan Wal / University of Edinburgh

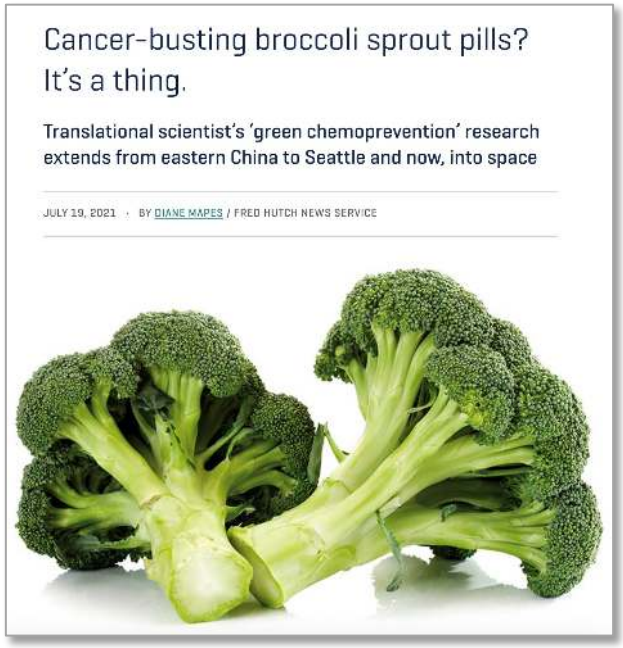
A puzzle



Top scientist: We will find the 'God particle'

The "father" of an elusive subatomic particle said Monday he is almost sure it will be discovered in the next year in a race between powerful research equipment in the United States and Europe.

British physicist Peter Higgs maintains that particle known as "Higgs boson" will be found when the Large Hadron Collider, where he is shown here, starts fully operating early next year. Alan Wal / University of Edinburgh



Cancer-busting broccoli sprout pills? It's a thing.

Translational scientist's 'green chemoprevention' research extends from eastern China to Seattle and now, into space

JULY 19, 2021 · BY DIANE MAPES / FRED HUTCH NEWS SERVICE

A puzzle





Why was it better to bet on the Higgs boson (pre test) than on broccoli (post test)?

You may have the answer. Let's see.

Aims of this talk

Meta level. Present framework for modeling interventions with research questions.

Aims of this talk

Meta level. Present framework for modeling interventions with research questions.

Object level. Demonstrate the value of the framework by using it to reveal insights for proposed interventions.

Brief history of the replication crisis



Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

© 2011 American Psychological Association
0022-3514/11/$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (*d*) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 4 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are discussed.

Keywords: psi, parapsychology, ESP, precognition, retrocausation

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. The term is purely descriptive; it neither implies that such phenomena are paranormal nor connotes anything about their underlying mechanisms. Alleged psi phenomena include *telepathy*, the apparent transfer of information from one person to another without the mediation of any known channel of sensory communication; *clairvoyance* (sometimes called *remote viewing*), the apparent perception of objects or events that do not provide a stimulus to the known senses; *psychokinesis*, the apparent influence of thoughts or intentions on physical or biological processes; and *precognition* (conscious cognitive awareness) or *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process.

This article was published Online First January 31, 2011.

I am grateful to the students who served as lead research assistants and laboratory coordinators for their enthusiasm and dedication to this controversial enterprise: Ben Edelman, Rebecca Epskin, Dan Feldman, Jamison Fisher, Eric Hartman, Kelly Lin, Universe Mowers, Brittany Nemer, and Jude Wu. I am also indebted to the 30 other students who served as friendly and reliable experimenters over the course of this research program. Dean Radin, senior scientist at the Institute of Noetic Sciences (IONS), and David Sherman, professor of psychology at the University of California, Santa Barbara, provided valuable guidance in the preparation of this article.

Correspondence concerning this article should be addressed to Daryl J. Bem, Department of Psychology, Uris Hall, Cornell University, Ithaca, NY 14853. E-mail: djb3@cornell.edu

Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports nine experiments designed to test for such retroactive influence by "time-reversing" several well-established psychological effects, so that the individual's responses are obtained before the putatively causal stimulus events occur.

Psi is a controversial subject, and most academic psychologists do not believe that psi phenomena are likely to exist. A survey of 1,100 college professors in the United States found that psychologists were much more skeptical about the existence of psi than were their colleagues in the natural sciences, the other social sciences, or the humanities (Wagner & Monnet, 1979). In fact, 34% of the psychologists in the sample declared psi to be impossible, a view expressed by only 2% of all other respondents. Although our colleagues in other disciplines would probably agree with the oft-quoted dictum that "extraordinary claims require extraordinary evidence," we psychologists are more likely to be familiar with the methodological and statistical requirements for sustaining such claims and aware of previous claims that failed either to meet those requirements or to survive the test of successful replication. Several other reasons for our greater skepticism are discussed by Bem and Honorton (1994, pp. 4–5).

There are two major challenges for psi researchers, one empirical and one theoretical. The major empirical challenge, of course, is to provide well-controlled demonstrations of psi that can be replicated by independent investigators. That is the major goal in the research program reported in this article. Accordingly, the

Deryl Bem's 2011 paper, 'Feeling the Future' purporting show compelling evidence for extra-sensory perception.

Brief history of the replication crisis



Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,100 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (*d*) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are discussed.

Keywords: psi, parapsychology, ESP, precognition, nonconscious

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. The term is purely descriptive; it neither implies that such phenomena are paranormal nor connotes anything about their underlying mechanisms. Alleged psi phenomena include *telepathy*, the apparent transfer of information from one person to another without the mediation of any known channel of sensory communication; *clairvoyance* (sometimes called *remote viewing*), the apparent perception of objects or events that do not provide a stimulus to the known senses; *psychokinesis*, the apparent influence of thoughts or intentions on physical or biological processes; and *precognition* (conscious cognitive awareness) or *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process.

This article was published Online First January 31, 2011.

I am grateful to the students who served as lead research assistants and laboratory coordinators for their enthusiasm and dedication to this controversial enterprise: Ben Feldman, Dan Feldman, Jeanine Hohr, Eric Hoffman, Kelly Lin, Brianne Meeuwis, Brittney Turner, and Jade Wu. I am also indebted to the 50 other students who served so literally and reliably *experimenters* over the course of this research program. Dean Radin, senior scientist at the Institute of Noetic Sciences (IONS), and David Sharman, professor of psychology at the University of California, Santa Barbara, provided valuable guidance in the preparation of this article.

Correspondence concerning this article should be addressed to Daryl J. Bem, Department of Psychology, Uris Hall, Cornell University, Ithaca, NY 14853. E-mail: djb3@cornell.edu

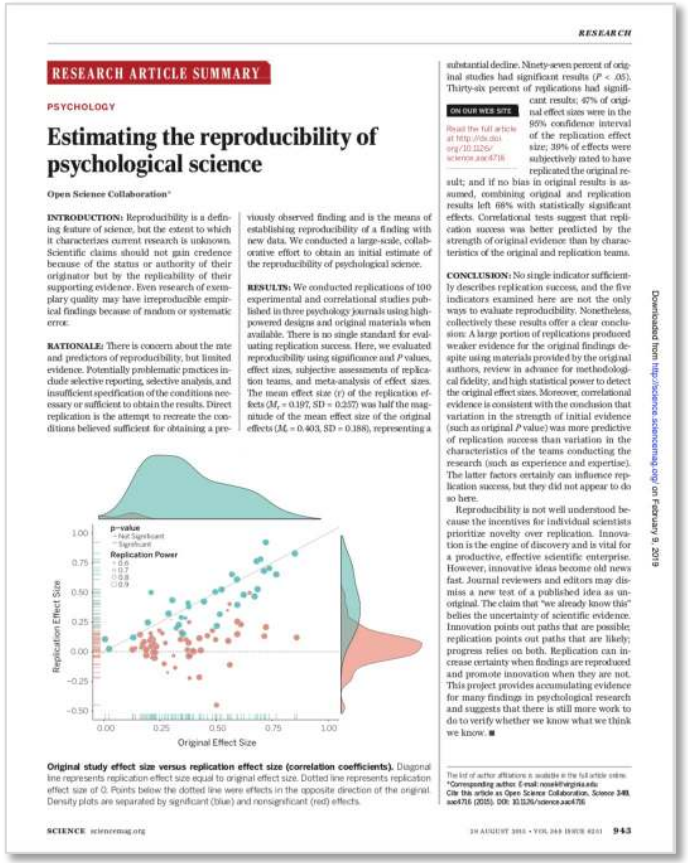
Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports nine experiments designed to test for such retroactive influence by "time-reversing" several well-established psychological effects, so that the individual's responses are obtained before the putatively causal stimulus events occur.

Psi is a controversial subject, and most academic psychologists do not believe that psi phenomena are likely to exist. A survey of 1,100 college professors in the United States found that psychologists were much more skeptical about the existence of psi than were their colleagues in the natural sciences, the other social sciences, or the humanities (Wagner & Monnet, 1979). In fact, 34% of the psychologists in the sample declared psi to be impossible, a view expressed by only 2% of all other respondents. Although our colleagues in other disciplines would probably agree with the oft-quoted dictum that "extraordinary claims require extraordinary evidence," we psychologists are more likely to be familiar with the methodological and statistical requirements for assessing such claims and aware of previous claims that failed either to meet those requirements or to survive the test of successful replication. Several other reasons for our greater skepticism are discussed by Bem and Honorton (1994, pp. 4–5).

There are two major challenges for psi researchers, one empirical and one theoretical. The major empirical challenge, of course, is to provide well-controlled demonstrations of psi that can be replicated by independent investigators. That is the major goal in the research program reported in this article. Accordingly, the

There was something wrong with existing practices.

Brief history of the replication crisis



RESEARCH

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

INTRODUCTION: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

RATIONALE: There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (*r*) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a substantial decline. Ninety-seven percent of original studies had significant results ($P < .05$). Thirty-six percent of replications had 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.

ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.aac4716

CONCLUSION: No single indicator sufficiently describes replication success, and the five indicators examined here are not the only ways to evaluate reproducibility. Nonetheless, collectively these results offer a clear conclusion: A large portion of replications produced weaker evidence for the original findings despite using materials provided by the original authors, review in advance for methodological fidelity, and high statistical power to detect the original effect sizes. Moreover, correlational evidence is consistent with the conclusion that variation in the strength of initial evidence (such as original *P* value) was more predictive of replication success than variation in the characteristics of the teams conducting the research (such as experience and expertise). The latter factors certainly can influence replication success, but they did not appear to do so here.

Reproducibility is not well understood because the incentives for individual scientists prioritize novelty over replication. Innovation is the engine of discovery and is vital for a productive, effective scientific enterprise. However, innovative ideas become old news fast. Journal reviewers and editors may dismiss a new test of a published idea as unoriginal. The claim that "we already know this" belies the uncertainty of scientific evidence. Innovation points out paths that are possible; replication points out paths that are likely; progress relies on both. Replication can increase certainty when findings are reproduced and promote innovation when they are not. This project provides accumulating evidence for many findings in psychological research and suggests that there is still more work to do to verify whether we know what we think we know. ■

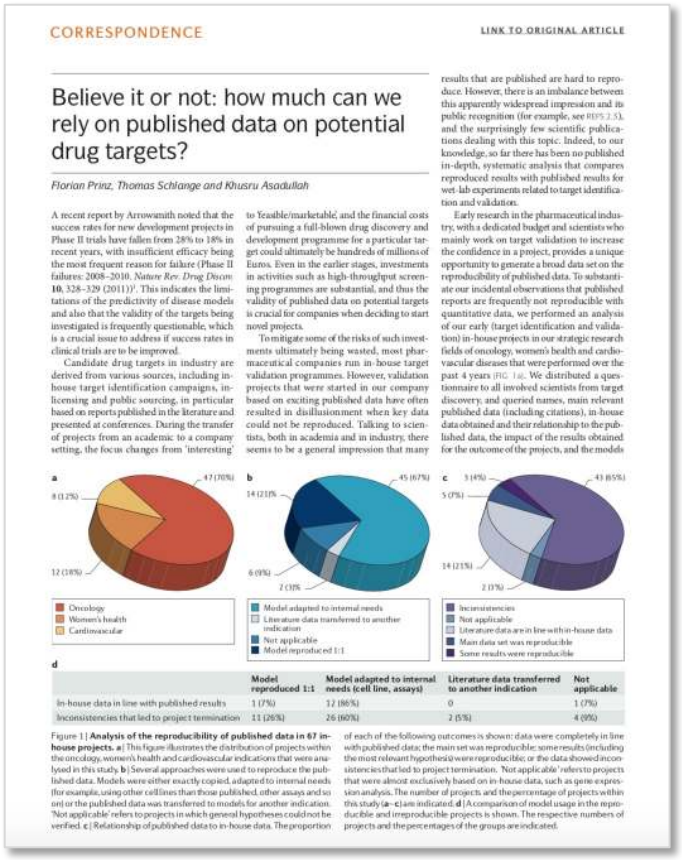
The list of author affiliations is available in the full article online.
*Corresponding author. E-mail: nosek@virginia.edu
Cite this article as Open Science Collaboration, *Science* 349, aac4716 (2015). DOI: 10.1126/science.aac4716

Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Downloaded from http://science.sciencemag.org/ on February 9, 2019

A large-scale effort to reproduce 100 findings in psychology found that only 39 held up.

Brief history of the replication crisis



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 is a crucial issue to address if success rates in clinical trials are to be improved.

Candidate drug targets in industry are derived from various sources, including in-house target identification campaigns, in-licensing and public sourcing, in particular based on reports published in the literature and presented at conferences. During the transfer of projects from an academic to a company setting, the focus changes from 'interesting' 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.

To mitigate some of the risks of such investments ultimately being wasted, most pharmaceutical companies run in-house target validation programmes. However, validation projects that were started in our company based on exciting published data have often resulted in disillusionment when key data could not be reproduced. Talking to scientists, both in academia and in industry, there seems to be a general impression that many 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.

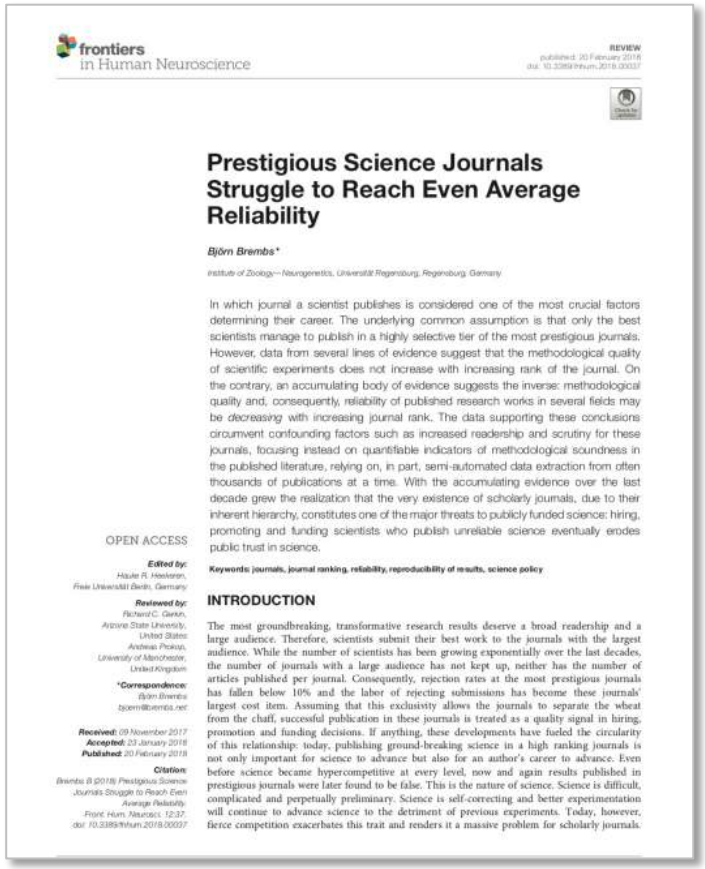
Early research in the pharmaceutical industry, with a dedicated budget and scientists who mainly work on target validation to increase the confidence in a project, provides a unique opportunity to generate a broad data set on the reproducibility of published data. To substantiate our incidental observations that published reports are frequently not reproducible with quantitative data, we performed an analysis of our early (target identification and validation) in-house projects in our strategic research fields of oncology, women's health and cardiovascular diseases that were performed over the past 4 years (FIG. 1a). We distributed a questionnaire to all involved scientists from target discovery, and queried names, main relevant published data (including citations), in-house data obtained and their relationship to the published data, the impact of the results obtained for the outcome of the projects, and the models

d

	Model reproduced 1:1	Model adapted to internal needs (cell line, assays)	Literature data transferred to another indication	Not applicable
In-house data in line with published results	1 (7%)	12 (86%)	0	1 (7%)
Inconsistencies that led to project termination	11 (26%)	26 (60%)	2 (5%)	4 (9%)

Figure 1 | **Analysis of the reproducibility of published data in 67 in-house projects.** **a** | This figure illustrates the distribution of projects within the oncology, women's health and cardiovascular indications that were analysed in this study. **b** | Several approaches were used to reproduce the published data. Models were either exactly copied, adapted to internal needs (for example, using other cell lines than those published, other assays and so on) or the published data was transferred to models for another indication. 'Not applicable' refers to projects in which general hypotheses could not be verified. **c** | Relationship of published data to in-house data. The proportion of each of the following outcomes is shown: data were completely in line with published data; the main set was reproducible; some results (including the most relevant hypothesis) were reproducible; or the data showed inconsistencies that led to project termination. 'Not applicable' refers to projects that were almost exclusively based on in-house data, such as gene expression analysis. The number of projects and the percentage of projects within this study (**a**–**c**) are indicated. **d** | A comparison of model usage in the reproducible and irreproducible projects is shown. The respective numbers of projects and the percentages of the groups are indicated.

An effort to reproduce 100 findings in biomedicine found that only 20 held up.



frontiers in Human Neuroscience

REVIEW

published: 20 February 2018

doi: 10.3389/fnhum.2018.00037

Prestigious Science Journals Struggle to Reach Even Average Reliability

*Björn Brembs**

Institute of Zoology — Neurogenetics, Universität Regensburg, Regensburg, Germany

In which journal a scientist publishes is considered one of the most crucial factors determining their career. The underlying common assumption is that only the best scientists manage to publish in a highly selective tier of the most prestigious journals. However, data from several lines of evidence suggest that the methodological quality of scientific experiments does not increase with increasing rank of the journal. On the contrary, an accumulating body of evidence suggests the inverse: methodological quality and, consequently, reliability of published research works in several fields may be *decreasing* with increasing journal rank. The data supporting these conclusions circumvent confounding factors such as increased readership and scrutiny for these journals, focusing instead on quantifiable indicators of methodological soundness in the published literature, relying on, in part, semi-automated data extraction from often thousands of publications at a time. With the accumulating evidence over the last decade grew the realization that the very existence of scholarly journals, due to their inherent hierarchy, constitutes one of the major threats to publicly funded science: hiring, promoting and funding scientists who publish unreliable science eventually erodes public trust in science.

Keywords: journals, journal ranking, reliability, reproducibility of results, science policy

OPEN ACCESS

Edited by:
Hauke R. Heekeren,
Freie Universität Berlin, Germany

Reviewed by:
Richard C. Gerkin,
Arizona State University, United States
Andreas Prokop,
University of Manchester, United Kingdom

****Correspondence:***
Björn Brembs
bjoern@brembs.net

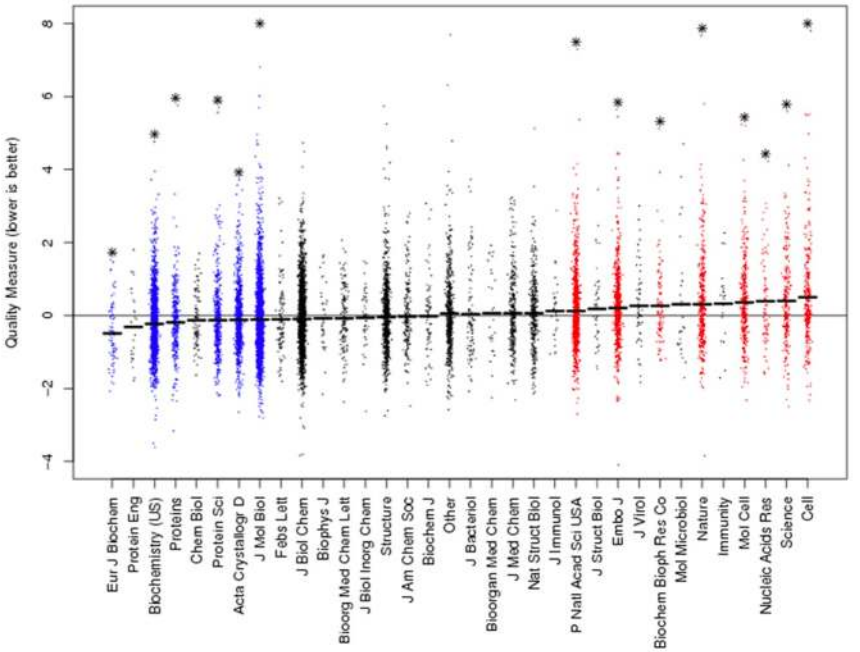
Received: *09 November 2017*
Accepted: *23 January 2018*
Published: *20 February 2018*

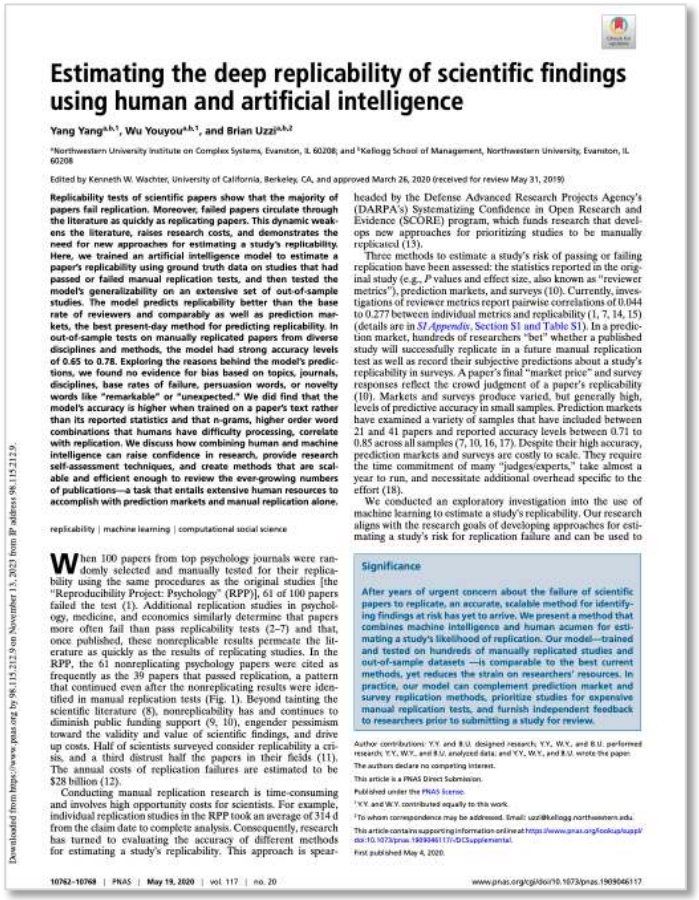
Citation:
Brembs B (2018) Prestigious Science Journals Struggle to Reach Even Average Reliability. Front. Hum. Neurosci. 12:37. doi: 10.3389/fnhum.2018.00037

INTRODUCTION

The most groundbreaking, transformative research results deserve a broad readership and a large audience. Therefore, scientists submit their best work to the journals with the largest audience. While the number of scientists has been growing exponentially over the last decades, the number of journals with a large audience has not kept up, neither has the number of articles published per journal. Consequently, rejection rates at the most prestigious journals has fallen below 10% and the labor of rejecting submissions has become these journals' largest cost item. Assuming that this exclusivity allows the journals to separate the wheat from the chaff, successful publication in these journals is treated as a quality signal in hiring, promotion and funding decisions. If anything, these developments have fueled the circularity of this relationship: today, publishing ground-breaking science in a high ranking journals is not only important for science to advance but also for an author's career to advance. Even before science became hypercompetitive at every level, now and again results published in prestigious journals were later found to be false. This is the nature of science. Science is difficult, complicated and perpetually preliminary. Science is self-correcting and better experimentation will continue to advance science to the detriment of previous experiments. Today, however, fierce competition exacerbates this trait and renders it a massive problem for scholarly journals.

Journal prestige appears to anti-correlate with replicability of findings.





Estimating the deep replicability of scientific findings using human and artificial intelligence

Yang Yang[a,b,1], Wu Youyou[a,b,1], and Brian Uzzi[a,b,2]

[a]Northwestern University Institute on Complex Systems, Evanston, IL 60208; and [b]Kellogg School of Management, Northwestern University, Evanston, IL 60208

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved March 26, 2020 (received for review May 31, 2019)

Replicability tests of scientific papers show that the majority of papers fail replication. Moreover, failed papers circulate through the literature as quickly as replicating papers. This dynamic weakens the literature, raises research costs, and demonstrates the need for new approaches for estimating a study's replicability. Here, we trained an artificial intelligence model to estimate a paper's replicability using ground truth data on studies that had passed or failed manual replication tests, and then tested the model's generalizability on an extensive set of out-of-sample studies. The model predicts replicability better than the base rate of reviewers and comparably as well as prediction markets, the best present-day method for predicting replicability. In out-of-sample tests on manually replicated papers from diverse disciplines and methods, the model had strong accuracy levels of 0.65 to 0.78. Exploring the reasons behind the model's predictions, we found no evidence for bias based on topics, journals, disciplines, base rates of failure, persuasion words, or novelty words like "remarkable" or "unexpected." We did find that the model's accuracy is higher when trained on a paper's text rather than its reported statistics and that n-grams, higher order word combinations that humans have difficulty processing, correlate with replication. We discuss how combining human and machine intelligence can raise confidence in research, provide research self-assessment techniques, and create methods that are scalable and efficient enough to review the ever-growing numbers of publications—a task that entails extensive human resources to accomplish with prediction markets and manual replication alone.

replicability | machine learning | computational social science

When 100 papers from top psychology journals were randomly selected and manually tested for their replicability using the same procedures as the original studies [the "Reproducibility Project: Psychology" (RPP)], 61 of 100 papers failed the test (1). Additional replication studies in psychology, medicine, and economics similarly determine that papers more often fail than pass replicability tests (2–7) and that, once published, these nonreplicable results permeate the literature as quickly as the results of replicating studies. In the RPP, the 61 nonreplicating psychology papers were cited as frequently as the 39 papers that passed replication, a pattern that continued even after the nonreplicating results were identified in manual replication tests (Fig. 1). Beyond tainting the scientific literature (8), nonreplicability has and continues to diminish public funding support (9, 10), engender pessimism toward the validity and value of scientific findings, and drive up costs. Half of scientists surveyed consider replicability a crisis, and a third distrust half the papers in their fields (11). The annual costs of replication failures are estimated to be $28 billion (12).

Conducting manual replication research is time-consuming and involves high opportunity costs for scientists. For example, individual replication studies in the RPP took an average of 314 d from the claim date to complete analysis. Consequently, research has turned to evaluating the accuracy of different methods for estimating a study's replicability. This approach is spearheaded by the Defense Advanced Research Projects Agency's (DARPA's) Systematizing Confidence in Open Research and Evidence (SCORE) program, which funds research that develops new approaches for prioritizing studies to be manually replicated (13).

Three methods to estimate a study's risk of passing or failing replication have been assessed: the statistics reported in the original study (e.g., P values and effect size, also known as "reviewer metrics"), prediction markets, and surveys (10). Currently, investigations of reviewer metrics report pairwise correlations of 0.044 to 0.277 between individual metrics and replicability (1, 7, 14, 15) (details are in *SI Appendix*, Section S1 and Table S1). In a prediction market, hundreds of researchers "bet" whether a published study will successfully replicate in a future manual replication test as well as record their subjective predictions about a study's replicability in surveys. A paper's final "market price" and survey responses reflect the crowd judgment of a paper's replicability (10). Markets and surveys produce varied, but generally high, levels of predictive accuracy in small samples. Prediction markets have examined a variety of samples that have included between 21 and 41 papers and reported accuracy levels between 0.71 to 0.85 across all samples (7, 10, 16, 17). Despite their high accuracy, prediction markets and surveys are costly to scale. They require the time commitment of many "judges/experts," take almost a year to run, and necessitate additional overhead specific to the effort (18).

We conducted an exploratory investigation into the use of machine learning to estimate a study's replicability. Our research aligns with the research goals of developing approaches for estimating a study's risk for replication failure and can be used to

Significance

After years of urgent concern about the failure of scientific papers to replicate, an accurate, scalable method for identifying findings at risk has yet to arrive. We present a method that combines machine intelligence and human acumen for estimating a study's likelihood of replication. Our model—trained and tested on hundreds of manually replicated studies and out-of-sample datasets —is comparable to the best current methods, yet reduces the strain on researchers' resources. In practice, our model can complement prediction market and survey replication methods, prioritize studies for expensive manual replication tests, and furnish independent feedback to researchers prior to submitting a study for review.

Author contributions: Y.Y. and B.U. designed research; Y.Y., W.Y., and B.U. performed research; Y.Y., W.Y., and B.U. analyzed data; and Y.Y., W.Y., and B.U. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

[1]Y.Y. and W.Y. contributed equally to this work.

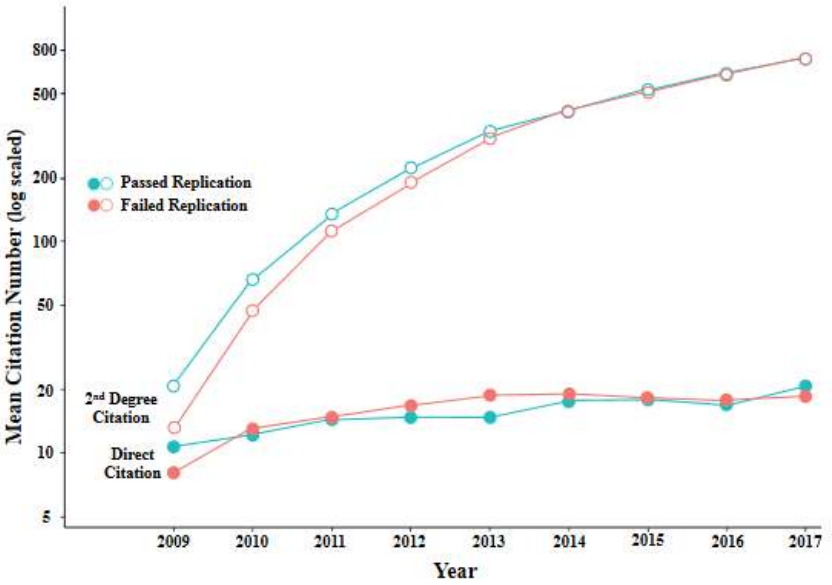
[2]To whom correspondence may be addressed. Email: uzzi@kellogg.northwestern.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1909046117/-/DCSupplemental.

First published May 4, 2020.

Downloaded from https://www.pnas.org by 98.115.212.9 on November 13, 2023 from IP address 98.115.212.9.

Replicability of studies is uncorrelated with their rate of citation.



RESEARCH ARTICLE

Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine

Tzu-Kun Hsiao and Jodi Schneider

School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign IL, USA

Keywords: citation analysis, citation context analysis, intentional postretraction citation, postretraction citation PubMed Central Open Access Subset, retraction

ABSTRACT

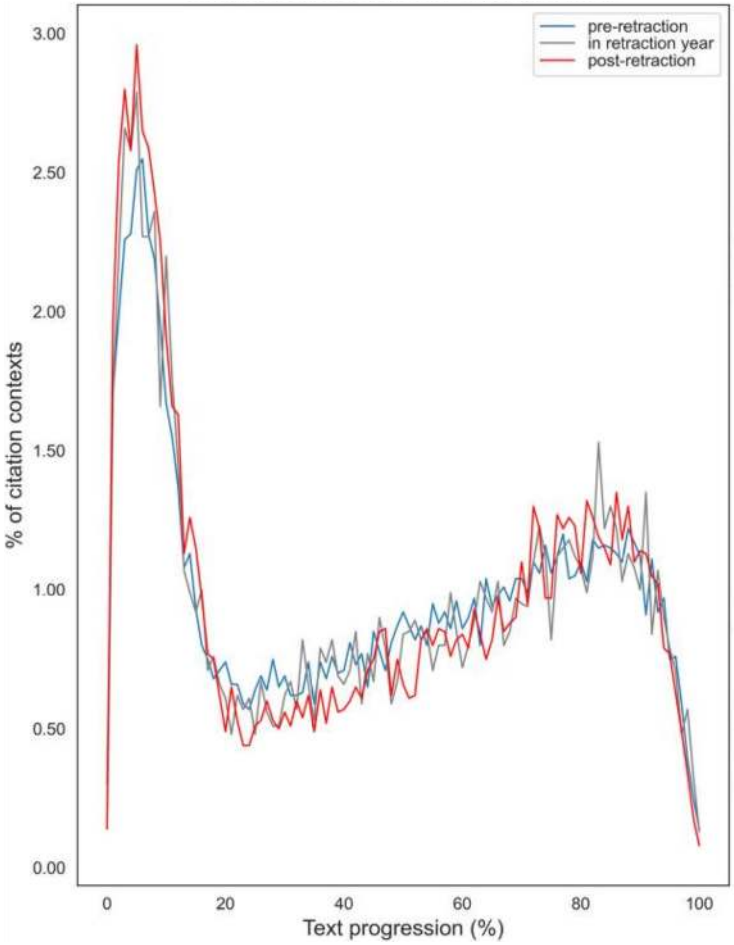
We present the first database-wide study on the citation contexts of retracted papers, which covers 7,813 retracted papers indexed in PubMed, 169,434 citations collected from iCite, and 48,134 citation contexts identified from the XML version of the PubMed Central Open Access Subset. Compared with previous citation studies that focused on comparing citation counts using two time frames (i.e., preretraction and postretraction), our analyses show the longitudinal trends of citations to retracted papers in the past 60 years (1960–2020). Our temporal analyses show that retracted papers continued to be cited, but that old retracted papers stopped being cited as time progressed. Analysis of the text progression of pre- and postretraction citation contexts shows that retraction did not change the way the retracted papers were cited. Furthermore, among the 13,252 postretraction citation contexts, only 722 (5.4%) citation contexts acknowledged the retraction. In these 722 citation contexts, the retracted papers were most commonly cited as related work or as an example of problematic science. Our findings deepen the understanding of why retraction does not stop citation and demonstrate that the vast majority of postretraction citations in biomedicine do not document the retraction.

1. INTRODUCTION

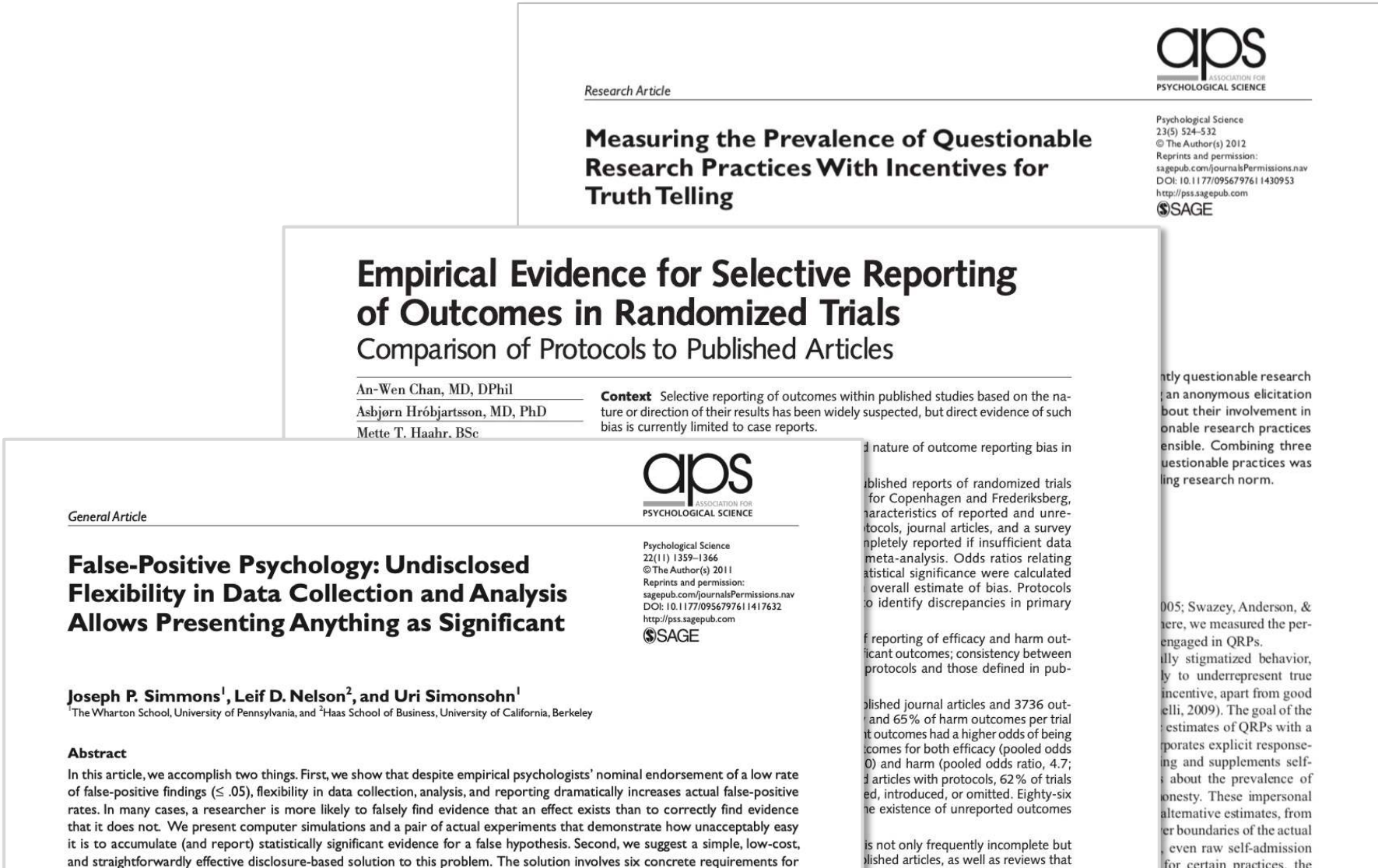
Although retraction is intended to remove published research from the citable literature, retraction does not stop the diffusion of the retracted paper. Empirical studies have shown that papers continue to be cited after being retracted (Bar-Ilan & Halevi, 2017; Bolland, Grey, & Avenell, 2021; Candal-Pedreira, Ruano-Ravina et al., 2020; Dal-Ré & Ayuso, 2020; Mott, Fairhurst, & Torgerson, 2019; Pfeifer & Snodgrass, 1990; Theis-Mahon & Bakker, 2020; van der Vet & Nijveen, 2016). For example, Dal-Ré and Ayuso (2020) studied citations to 460 retracted genetics articles and found that 23% of the citations were postretraction citations.

Although citations to retracted papers have been widely discussed in previous studies, the bulk of studies focused on a small number of retracted papers (e.g., 15 retracted papers in Bar-Ilan and Halevi [2017]) or a single retracted paper (Schneider, Ye et al., 2020; Suelzer, Deal et al., 2019; van der Vet & Nijveen, 2016). Only citation counts and citation networks have

Citation rates appear uncorrelated with study retraction.



Imputed causes of the crisis



Research Article

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

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

Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials
Comparison of Protocols to Published Articles

An-Wen Chan, MD, DPhil
Asbjørn Hróbjartsson, MD, PhD
Mette T. Haahr, BSc

Context Selective reporting of outcomes within published studies based on the nature or direction of their results has been widely suspected, but direct evidence of such bias is currently limited to case reports.

General Article

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

Psychological Science
22(11) 1359–1366
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797611417632
http://pss.sagepub.com

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

Abstract

In this article, we accomplish two things. First, we show that despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤ .05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. We present computer simulations and a pair of actual experiments that demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis. Second, we suggest a simple, low-cost, and straightforwardly effective disclosure-based solution to this problem. The solution involves six concrete requirements for

Publication bias

Produces a literature consisting predominantly in statistically significant (rather than null) results.



RESEARCH | REPORTS

also capture signals of warm-season upwelling, which in the CC is uncorrelated to winter upwelling and dominated by decadal-scale variability (*9*, *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. Hazen, *Prog. Oceanogr.* **12**, 265–284 (1983).
2. R. L. Smith, *Oceanogr. Mar. Biol. Annu. Rev.* **6**, 31–46 (1968).
3. W. J. Sydeman *et al.*, *Science* **345**, 77–80 (2014).
4. B. A. Black, *Mar. Ecol. Prog. Ser.* **398**, 37–49 (2009).
5. B. A. Black *et al.*, *Glob. Change Biol.* **17**, 2536–2545 (2011).
6. I. D. Schroeder *et al.*, *Geophys. Res. Lett.* **40**, 1–6 (2013).
7. M. García-Reyes *et al.*, *Ecosystems (N.Y.)* **18**, 722–735 (2015).
8. E. 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, 1946–2995," NOAA Technical Memo. NOAA-TM-NMFS-SWFSC (NOAA, Washington, DC, 1996).
11. D. W. Stahle *et al.*, *Earth Interact.* **17**, 1–23 (2013).
12. I. F. Koehler, F. Di Lorenzo, C. A. Morgan, V. Combes, W. T. Peterson, *Glob. Change Biol.* **17**, 2481–2501 (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. Kaplanis, E. R. Larson, S. St George, *Geophys. Res. Lett.* **39**, L17705 (2012).
17. K. Wolter, M. S. Timlin, *Weather* **53**, 315–324 (1998).
18. J. B. Li *et al.*, *Nat. Clim. Change* **1**, 134–138 (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. Mallard, Rolain, B. L. Ingram, M. Hughes, J. L. Fordham, *Quat. Sci. Rev.* **25**, 1370–1386 (2006).
26. F. B. Schwing, R. Mendelssohn, *J. Geophys. Res. Oceans* **102**, 3403–3418 (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. 1338225), 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. D'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 S5
References (*28*–*30*)

11 March 2014; accepted 7 August 2014
10.1126/science.1253999

Downloaded from http://science.sciencemag.org/ on May 17, 2019

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.

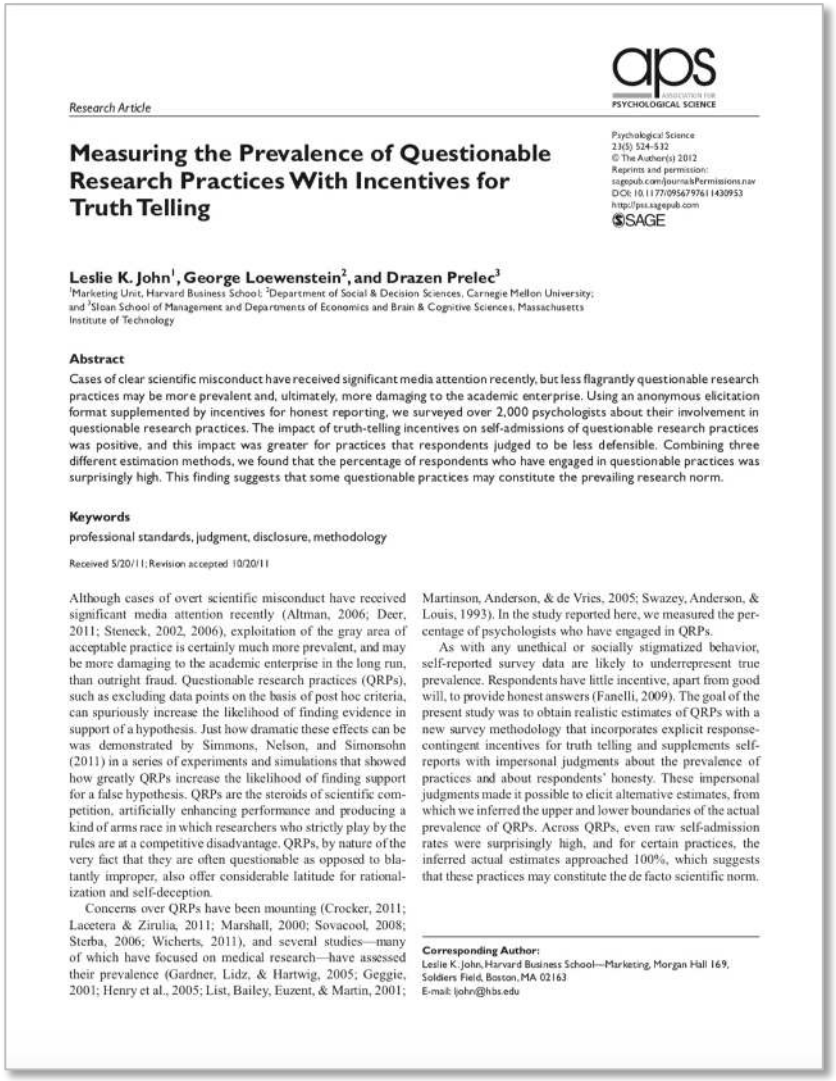
Publication bias occurs when "publication of study results is based on the direction or significance of the findings" (*1*). One pernicious form of publication bias is the greater likelihood of statistically significant results being published than statistically insignificant results, holding fixed research 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



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, 2001; Henry et al., 2005; List, Bailey, Euzent, & Martin, 2001; 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.

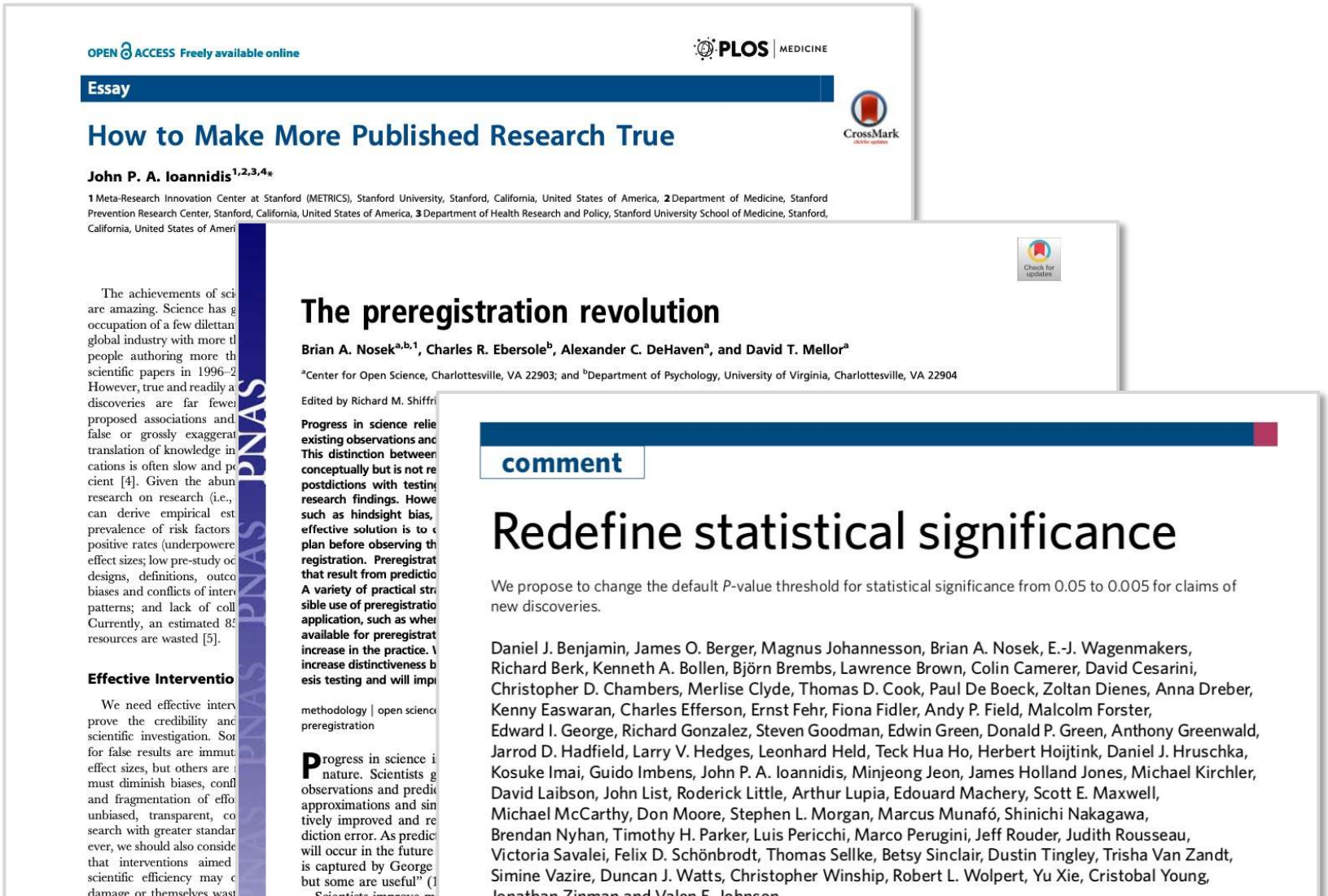
As with any unethical or socially stigmatized behavior, self-reported survey data are likely to underrepresent true prevalence. Respondents have little incentive, apart from good will, to provide honest answers (Fanelli, 2009). The goal of the present study was to obtain realistic estimates of QRPs with 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
E-mail: ljohn@hbs.edu

QRPs

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

Proposals



OPEN ACCESS Freely available online

PLOS | MEDICINE

Essay

How to Make More Published Research True

John P. A. Ioannidis[1,2,3,4]*

1 Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, United States of America, 2 Department of Medicine, Stanford Prevention Research Center, Stanford, California, United States of America, 3 Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California, United States of Ameri

The achievements of sci
are amazing. Science has g
occupation of a few dilettan
global industry with more th
people authoring more th
scientific papers in 1996–2
However, true and readily a
discoveries are far fewe
proposed associations and
false or grossly exaggerat
translation of knowledge in
cations is often slow and p
cient [4]. Given the abun
research on research (i.e.,
can derive empirical est
prevalence of risk factors
positive rates (underpowere
effect sizes; low pre-study od
designs, definitions, outco
biases and conflicts of inter
patterns; and lack of coll
Currently, an estimated 85
resources are wasted [5].

Effective Interventio

We need effective interv
prove the credibility and
scientific investigation. Sor
for false results are immut
effect sizes, but others are
must diminish biases, confl
and fragmentation of effo
unbiased, transparent, co
search with greater standar
ever, we should also conside
that interventions aimed
scientific efficiency may c
damage or themselves wast

PNAS PNAS PNAS

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. Shiffri

**Progress in science relie
existing observations and
This distinction between
conceptually but is not re
postdictions with testin
research findings. Howe
such as hindsight bias,
effective solution is to
plan before observing th
registration. Preregistrat
that result from predictio
A variety of practical str
sible use of preregistratio
application, such as wher
available for preregistrat
increase in the practice.
increase distinctiveness b
esis testing and will impr**

methodology | open scienc
preregistration

Progress in science i
nature. Scientists g
observations and predi
approximations and sin
tively improved and re
diction error. As predic
will occur in the future
is captured by George
but some are useful" (1

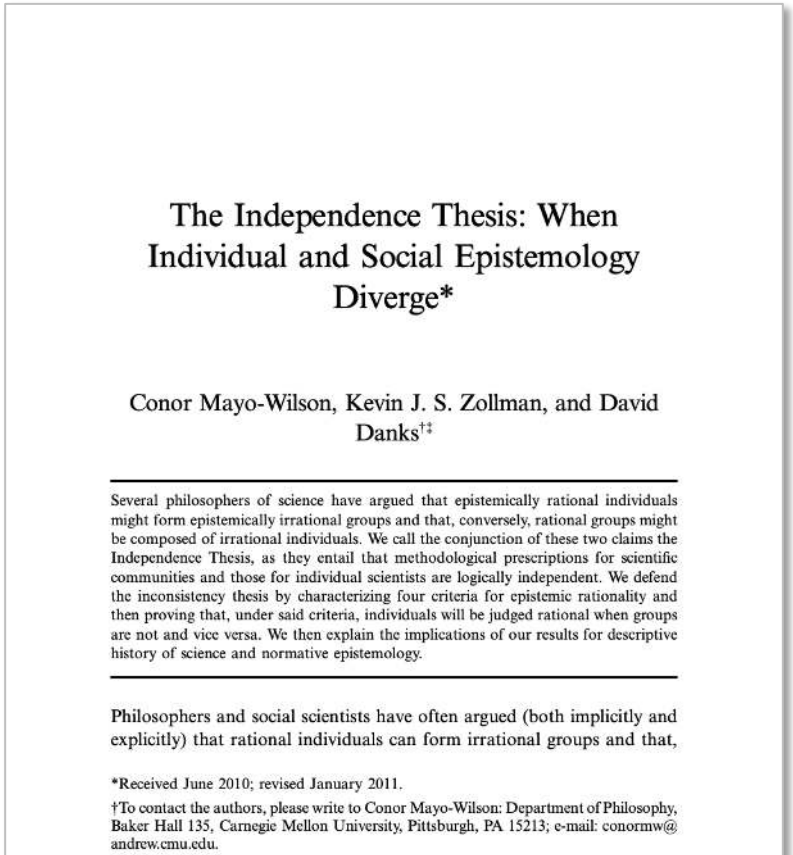
comment

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,

We will see how interventions that can look good at the level of individual studies can be bad when deployed at the level of a community of studies.



The Independence Thesis: When Individual and Social Epistemology Diverge*

Conor Mayo-Wilson, Kevin J. S. Zollman, and David Danks†‡

Several philosophers of science have argued that epistemically rational individuals might form epistemically irrational groups and that, conversely, rational groups might be composed of irrational individuals. We call the conjunction of these two claims the Independence Thesis, as they entail that methodological prescriptions for scientific communities and those for individual scientists are logically independent. We defend the inconsistency thesis by characterizing four criteria for epistemic rationality and then proving that, under said criteria, individuals will be judged rational when groups are not and vice versa. We then explain the implications of our results for descriptive history of science and normative epistemology.

Philosophers and social scientists have often argued (both implicitly and explicitly) that rational individuals can form irrational groups and that,

*Received June 2010; revised January 2011.

†To contact the authors, please write to Conor Mayo-Wilson: Department of Philosophy, Baker Hall 135, Carnegie Mellon University, Pittsburgh, PA 15213; e-mail: conormw@andrew.cmu.edu.

(Soundness-Dependent Selection) An intervention can potentially worsen replicability under the following conditions:

- Variation in the methodological soundness of studies;
- Variation in how studies are affected by intervention;
- Publication bias.

My aim is *emphatically not* to argue against the interventions considered here.

Judiciously deployed, I believe each intervention has its place in our repertoire of interventions.

Upshot 1. Under plausible conditions, we should expect soundness-dependent selection to obtain; this can effect our assessment and expectations of interventions.

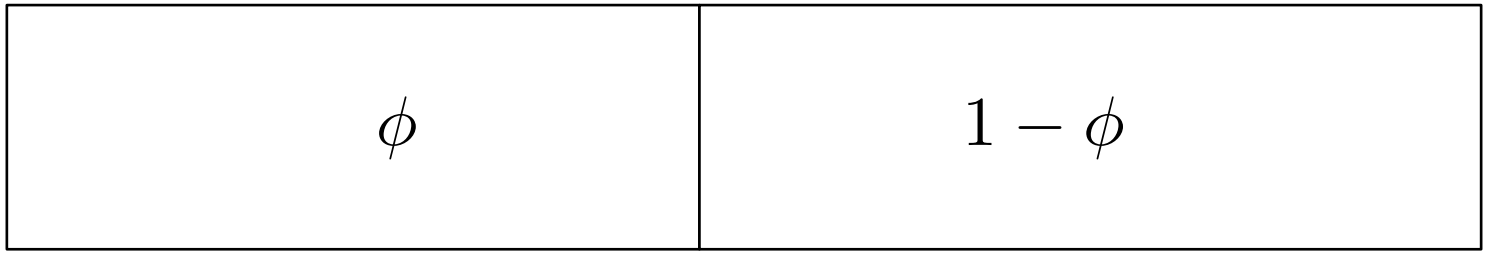
Upshot 2. Certain forms of *voluntary* improvement in evidentiary standards and practices—like voluntary preregistration—may be acutely vulnerable to adverse selection.

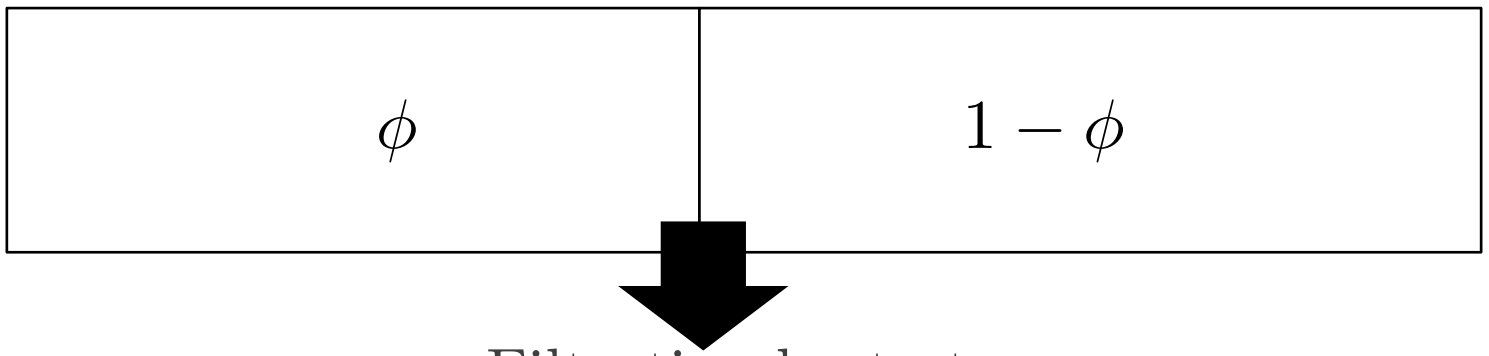
Upshot 3. Interventions which mitigate publication bias—like registered reports—should be favored in such cases.

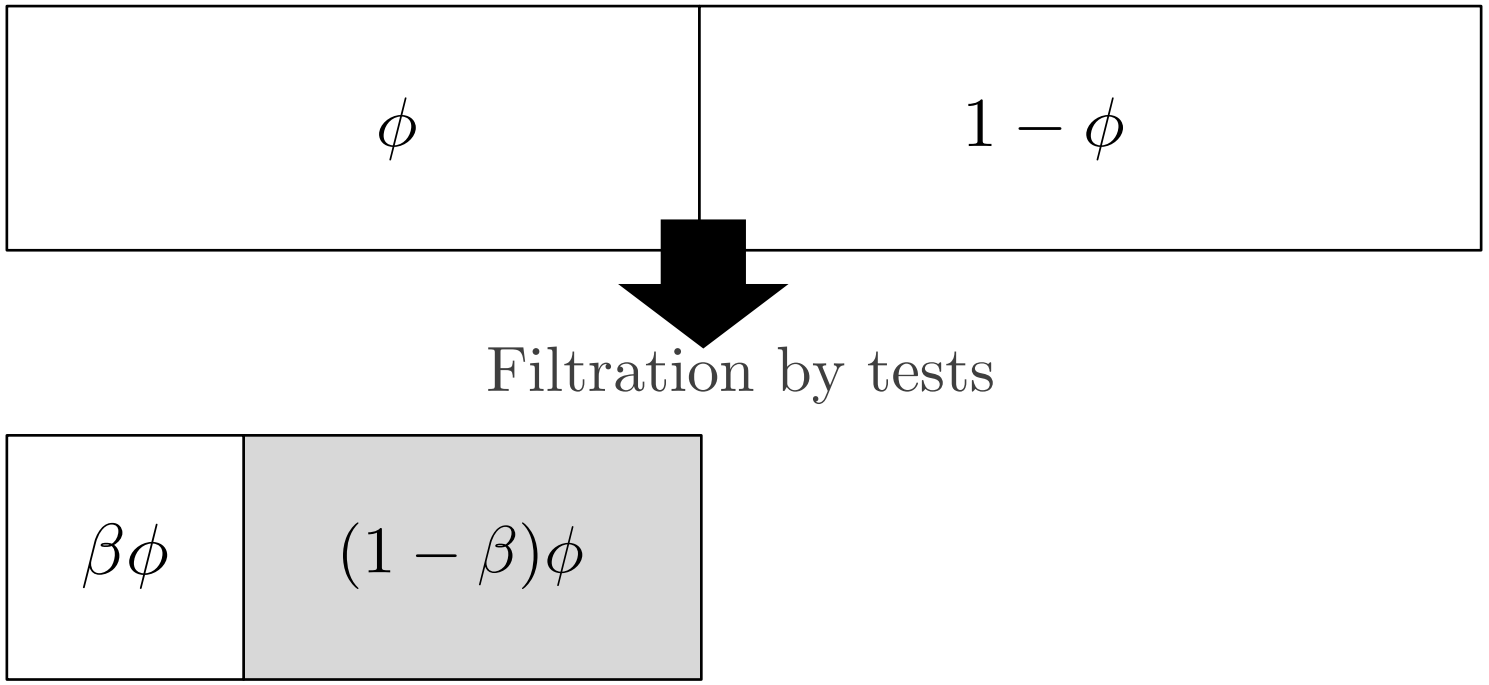
The filter model of science

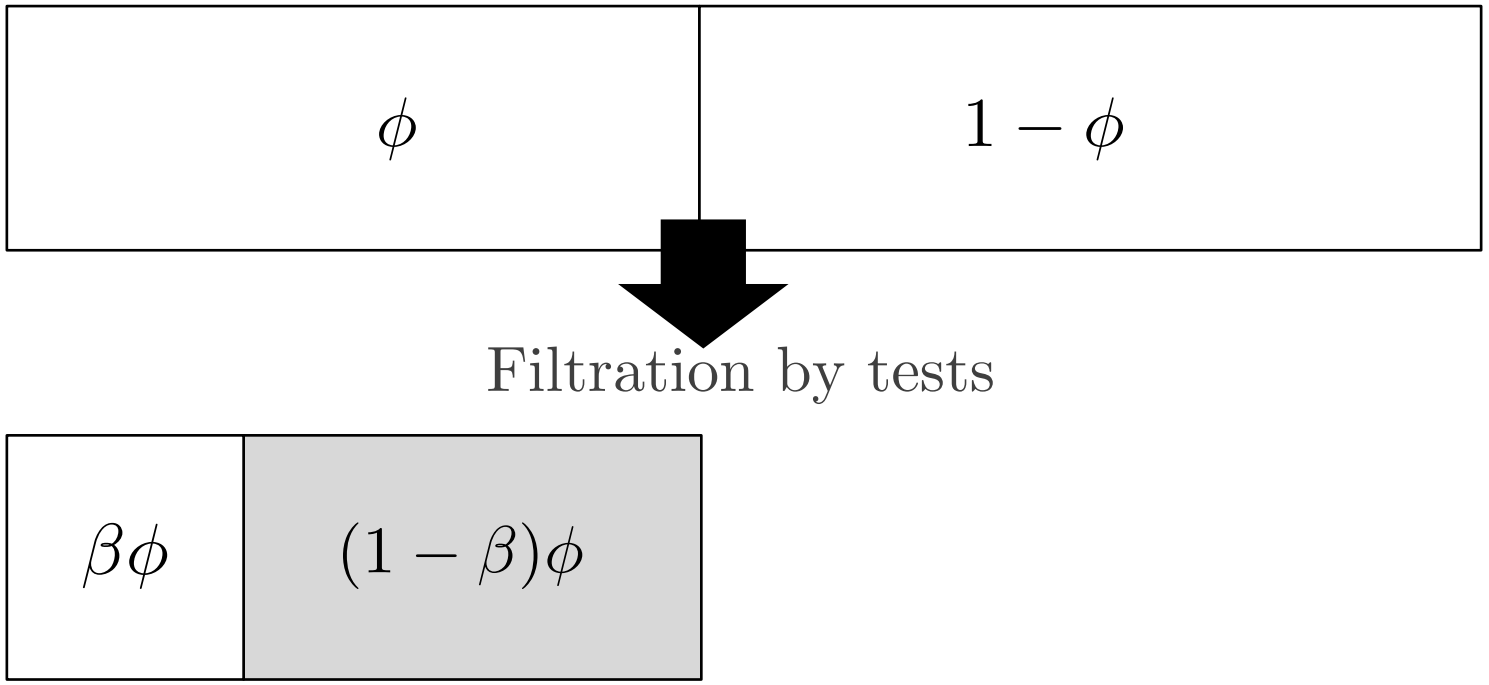
Adaptation of a classic model of bias in science (Ioannidis, 2005; Maniadis, Tufano, and List, 2014)

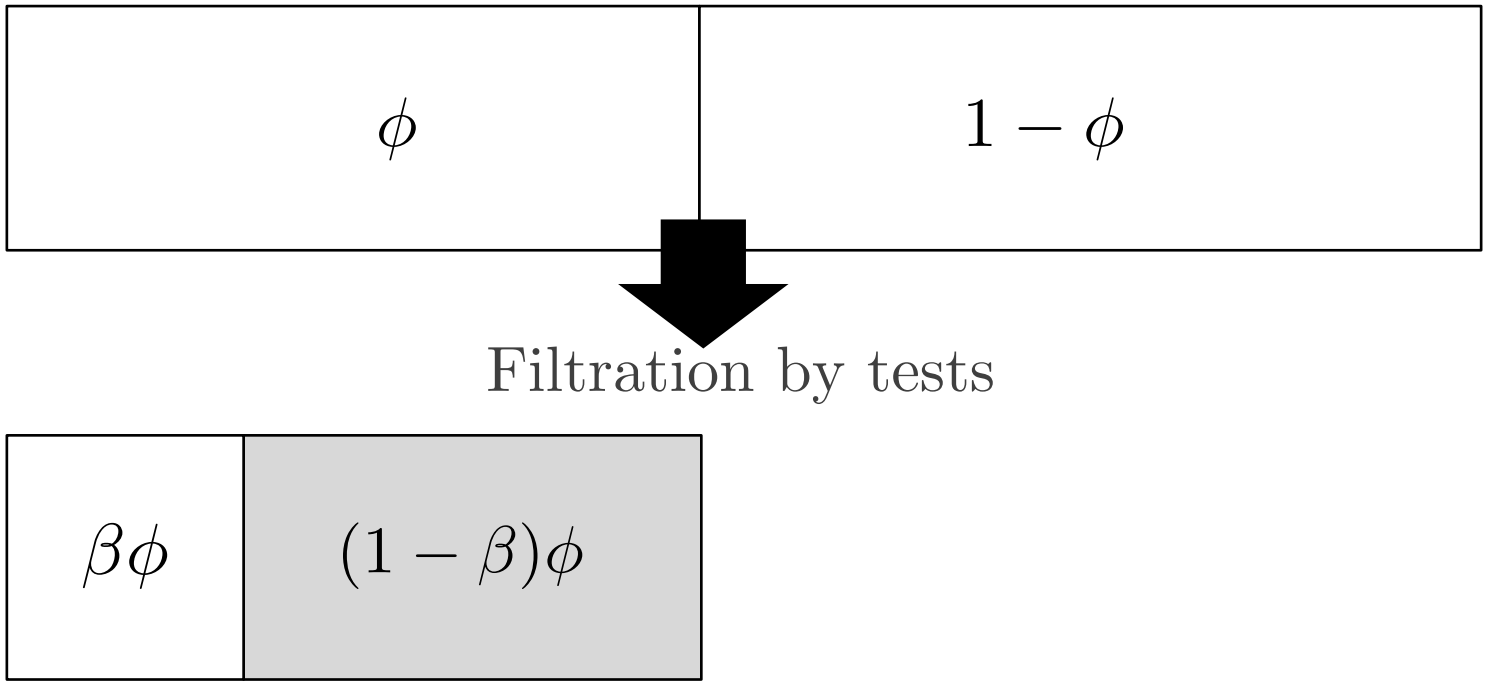
Population of studies

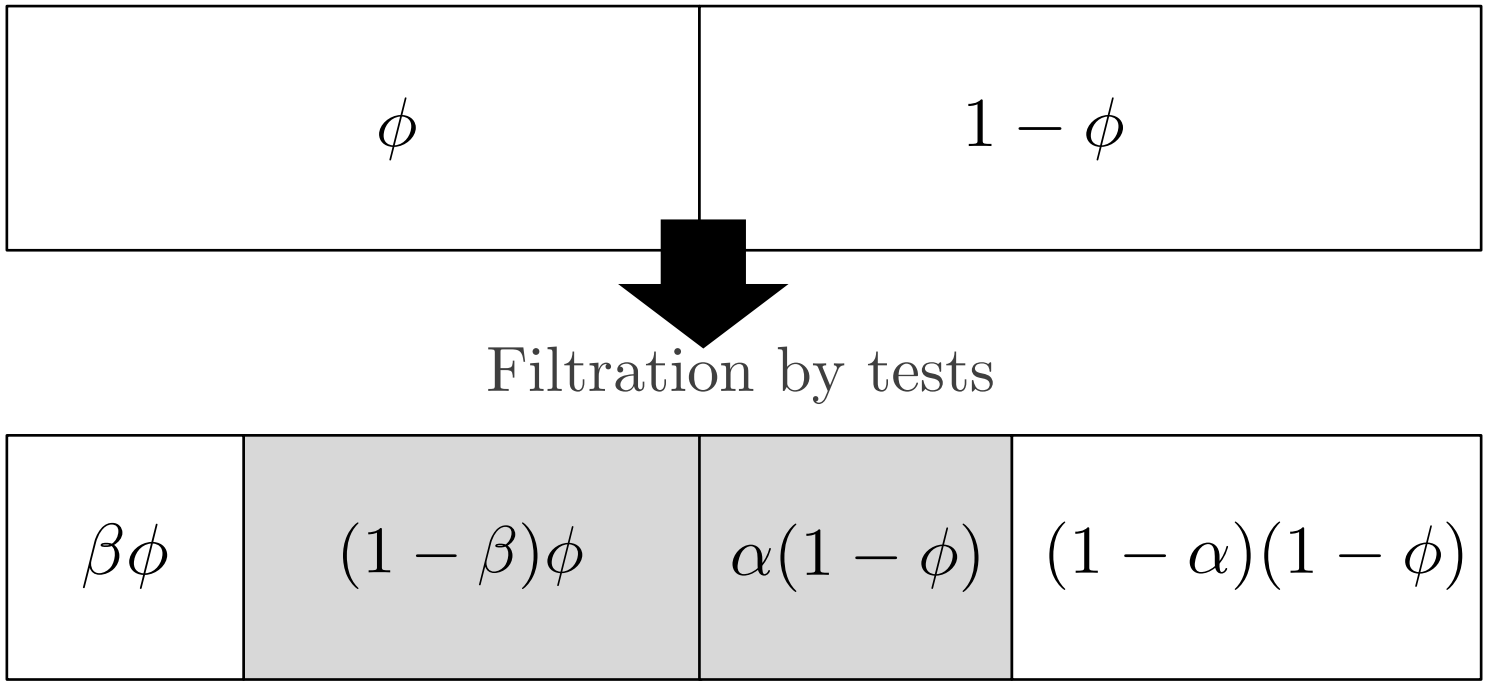


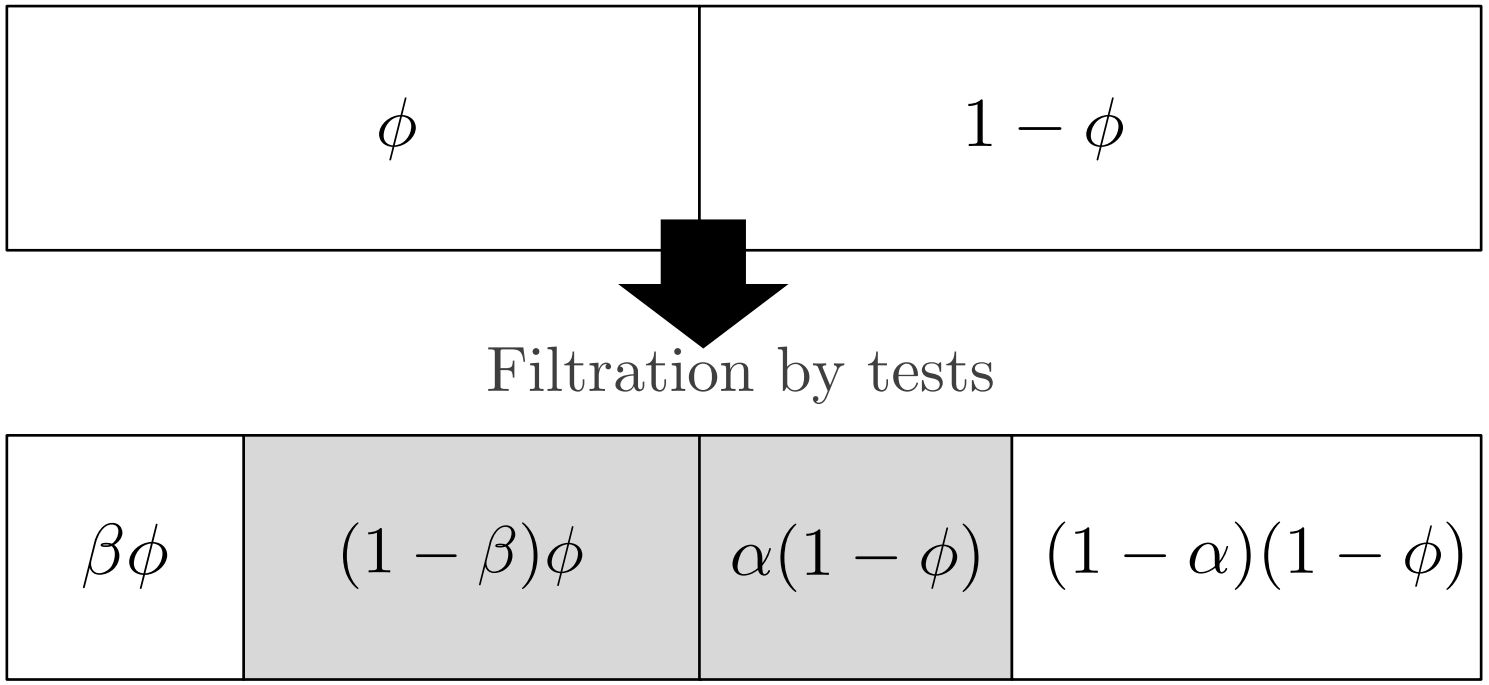

Population of studies
ϕ
$1-\phi$
Filtration by tests

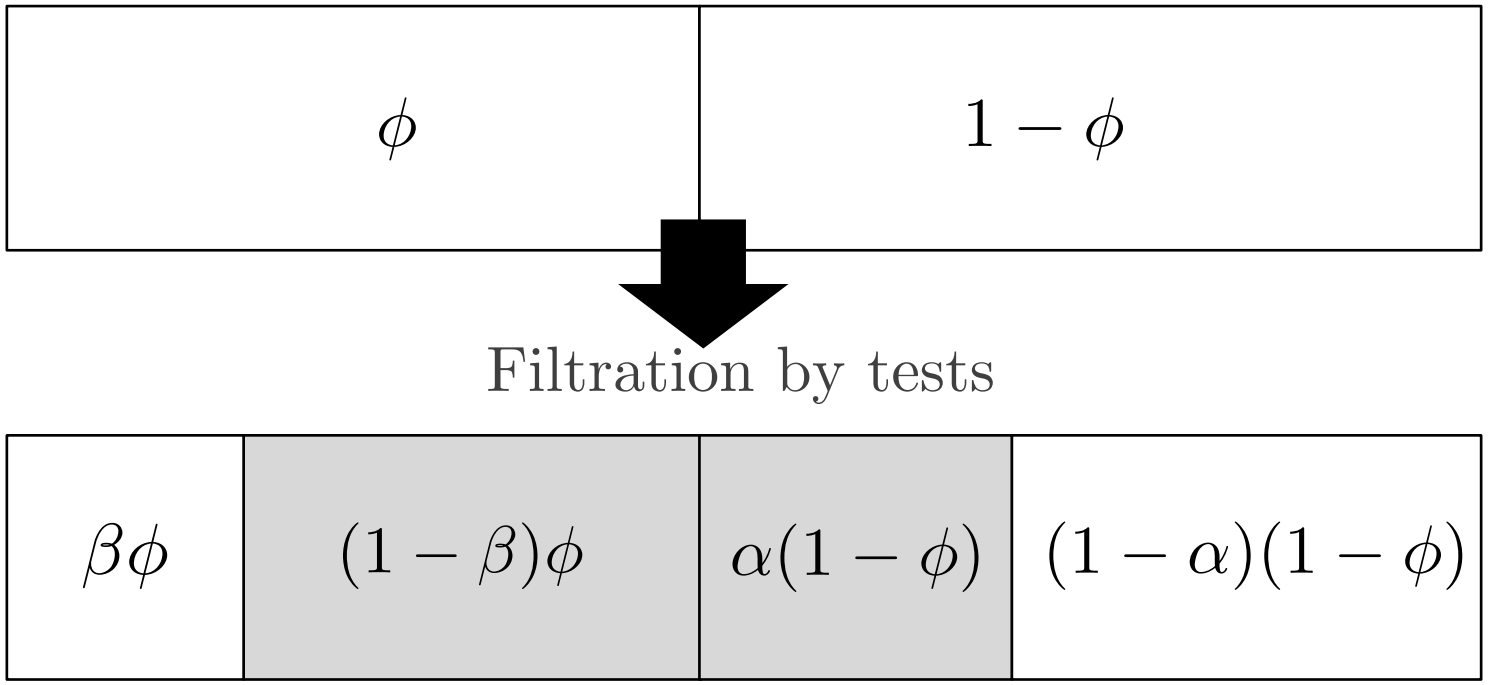

Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$

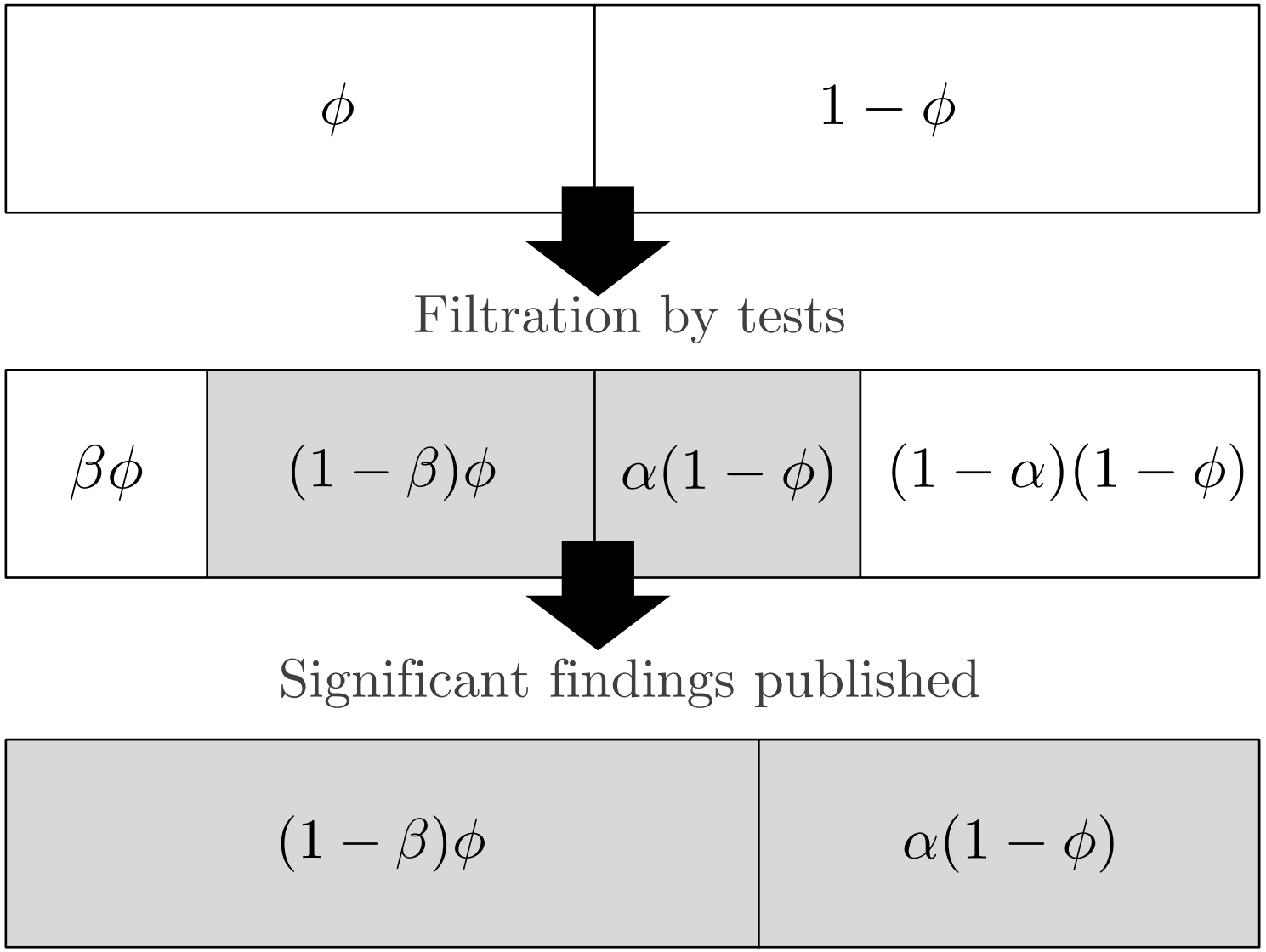

Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
False –


Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
False –
True +


Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
False –
True +


Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
False –
True +
False +


Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
False –
True +
False +
True –


Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
Significant findings published
$(1-\beta)\phi$
$\alpha(1-\phi)$

False discovery rate of published findings

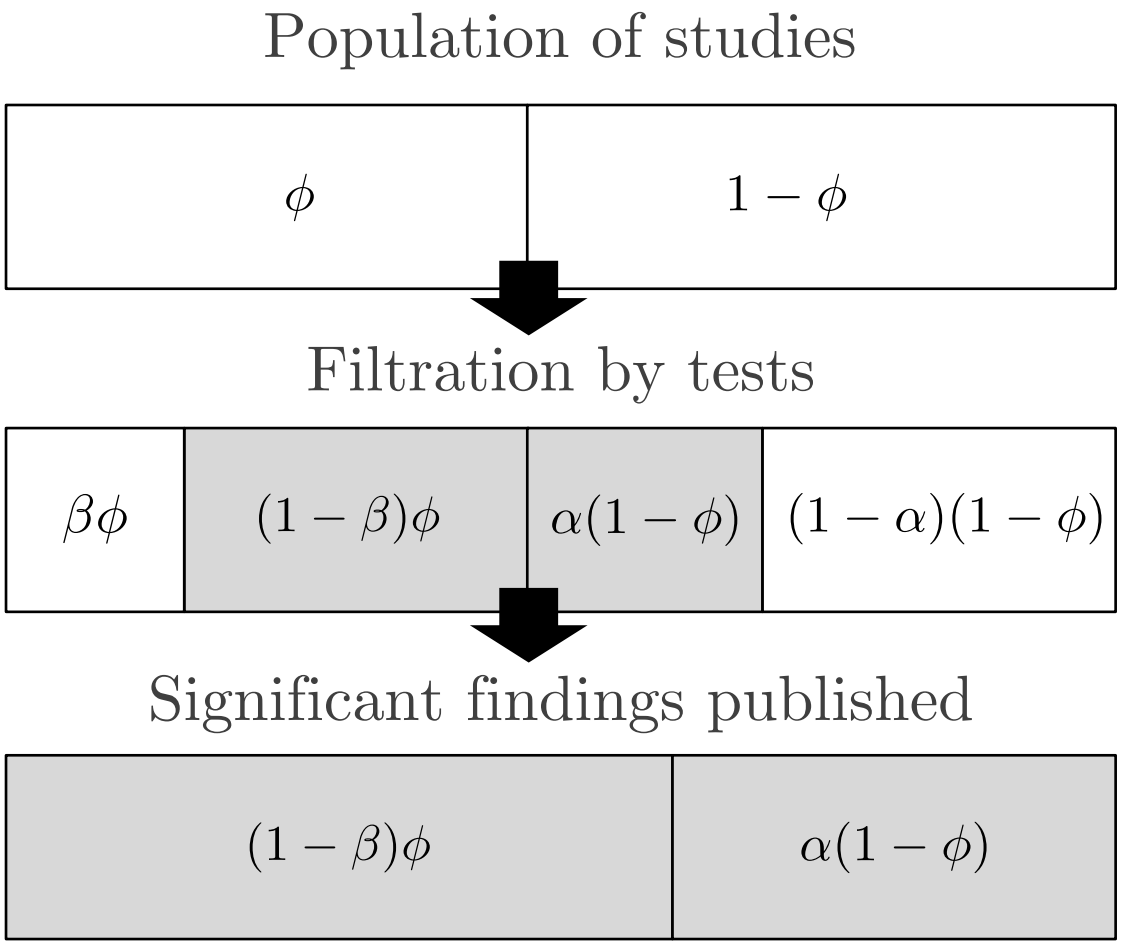
$$\frac{\alpha(1-\phi)}{(1-\beta)\phi+\alpha(1-\phi)}$$

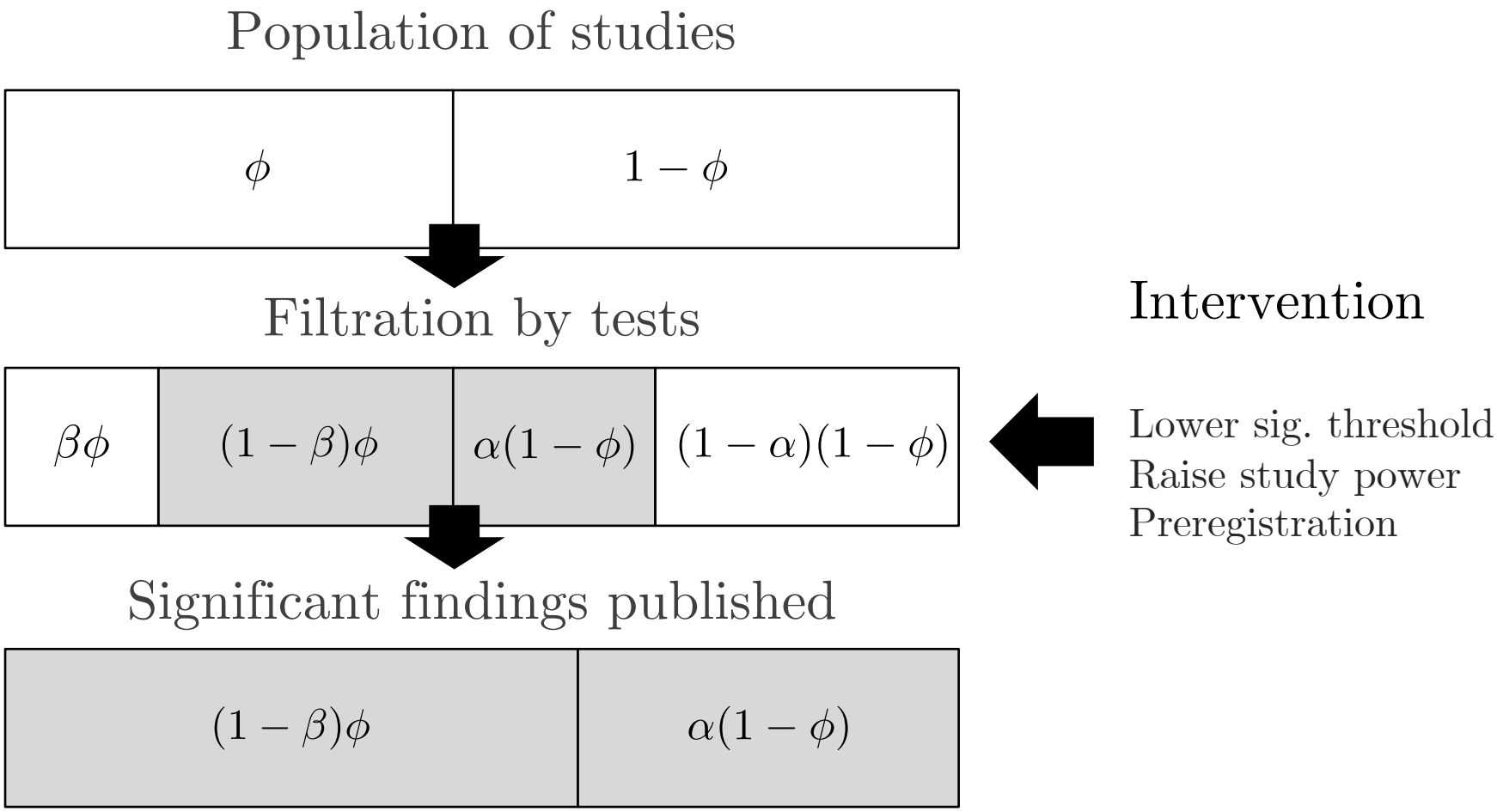
False discovery rate of published findings

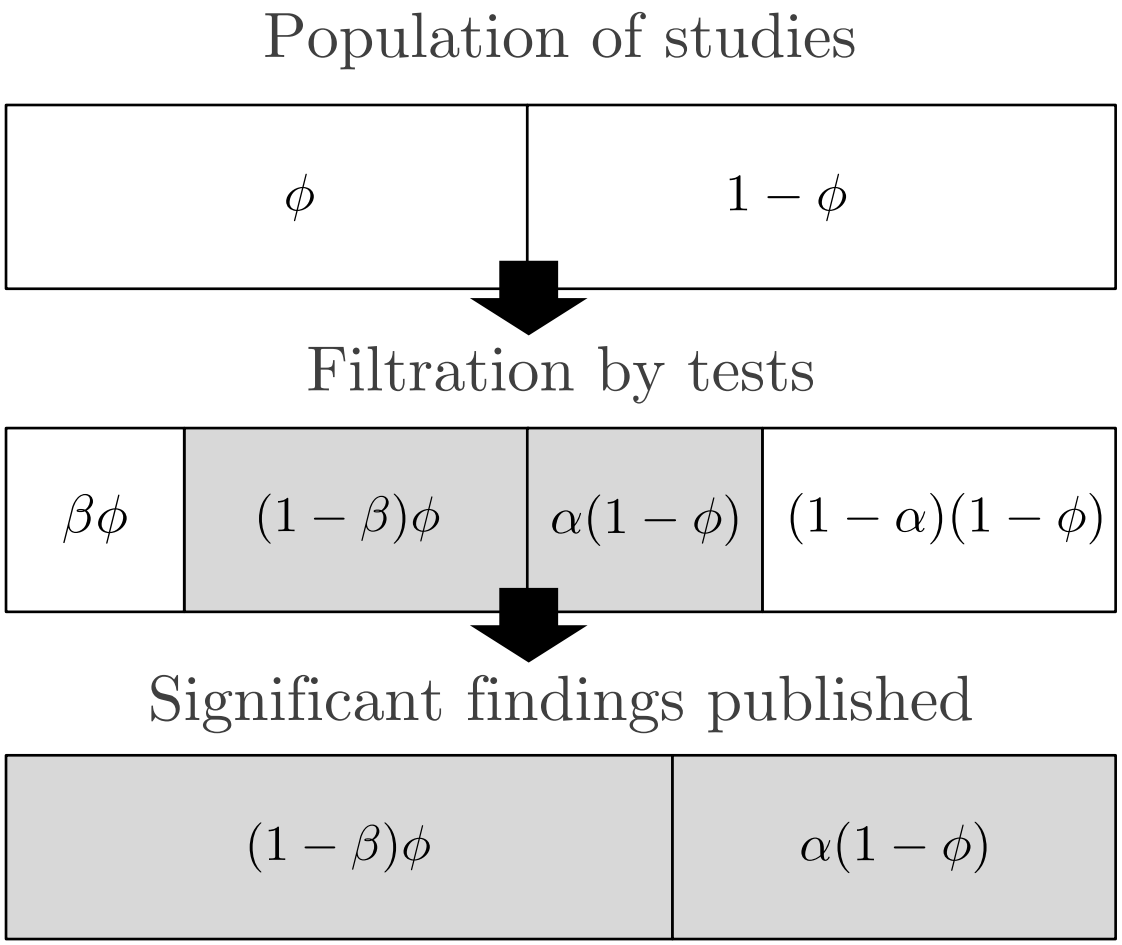
$$\frac{\alpha(1-\phi)}{(1-\beta)\phi+\alpha(1-\phi)}$$

signif. & false

signif.

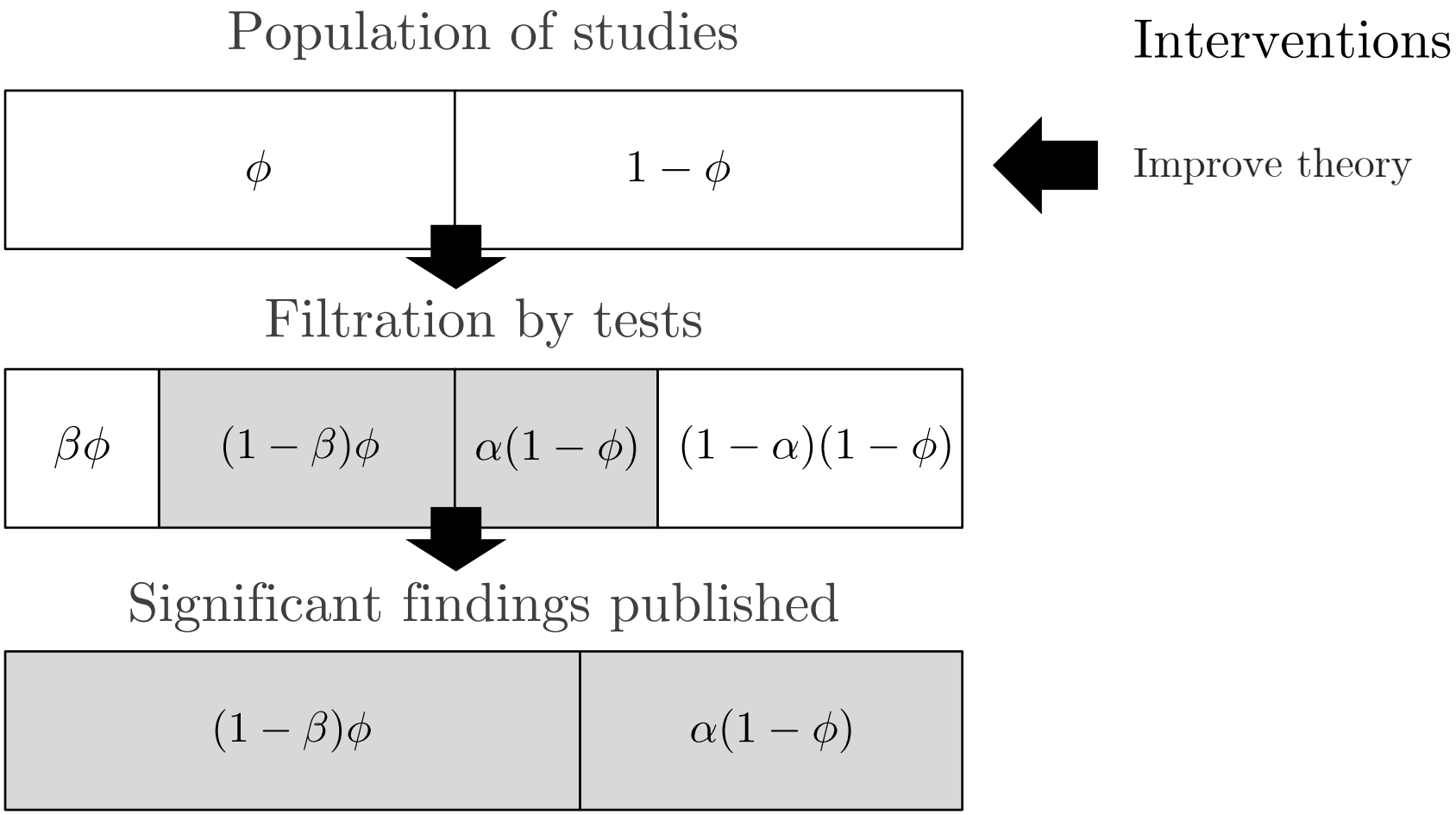

Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
Significant findings published
$(1-\beta)\phi$
$\alpha(1-\phi)$


Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
Significant findings published
$(1-\beta)\phi$
$\alpha(1-\phi)$
Intervention
Lower sig. threshold
Raise study power
Preregistration



Interventions

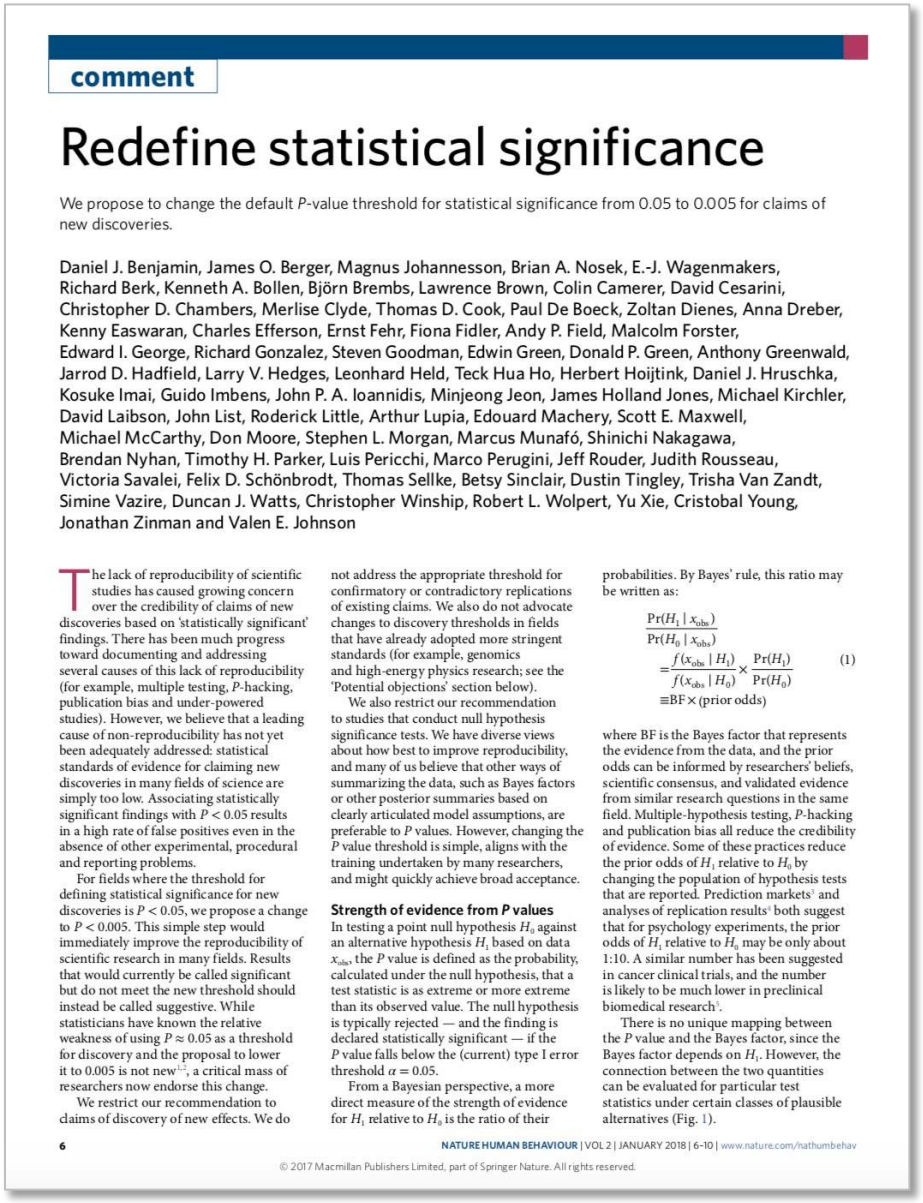
Journals of null results
Registered reports*


Population of studies
Interventions
ϕ
$1-\phi$
Improve theory
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
Significant findings published
$(1-\beta)\phi$
$\alpha(1-\phi)$

Soundness-Dependent Selection

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[1–3], 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:

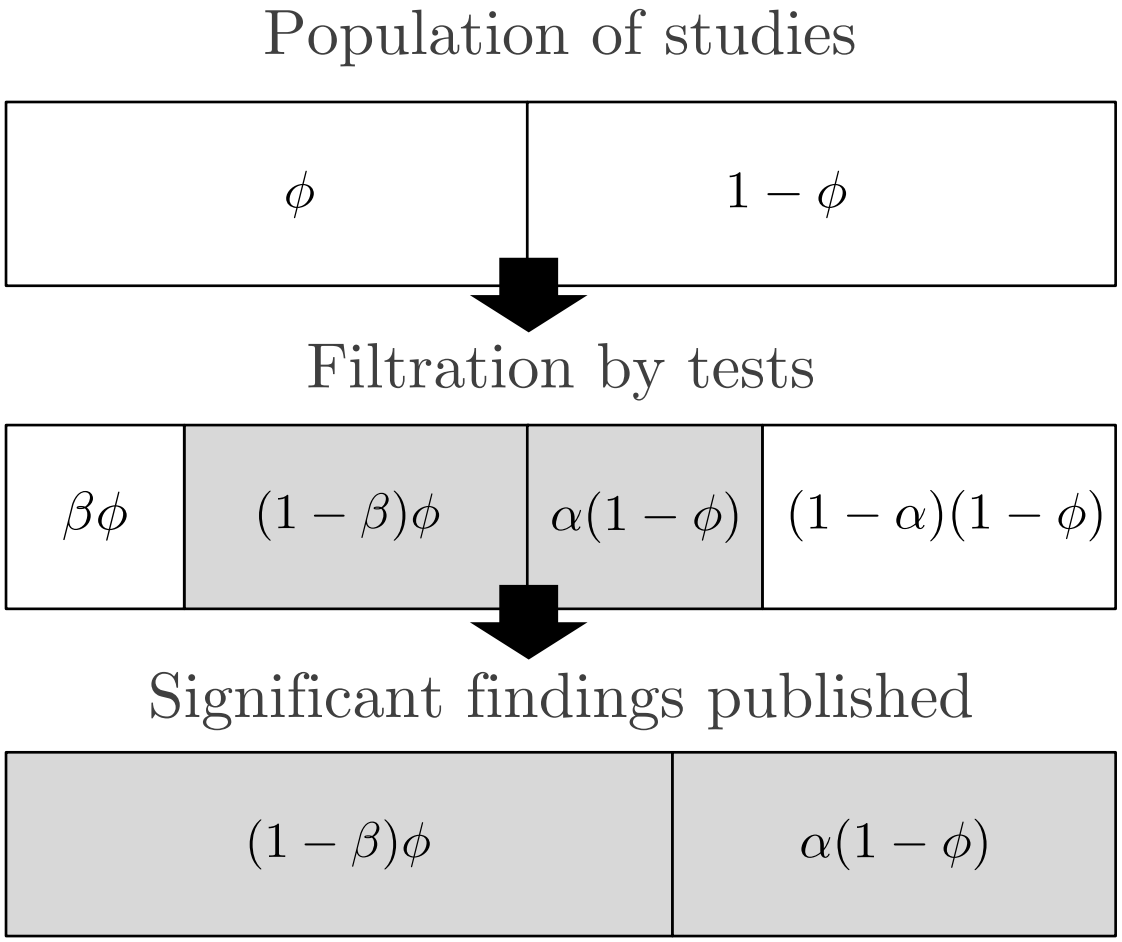
$$\frac{\Pr(H_1 \mid x_{obs})}{\Pr(H_0 \mid x_{obs})} = \frac{f(x_{obs} \mid H_1)}{f(x_{obs} \mid H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)} \equiv \mathrm{BF} \times (\text{prior odds}) \quad (1)$$

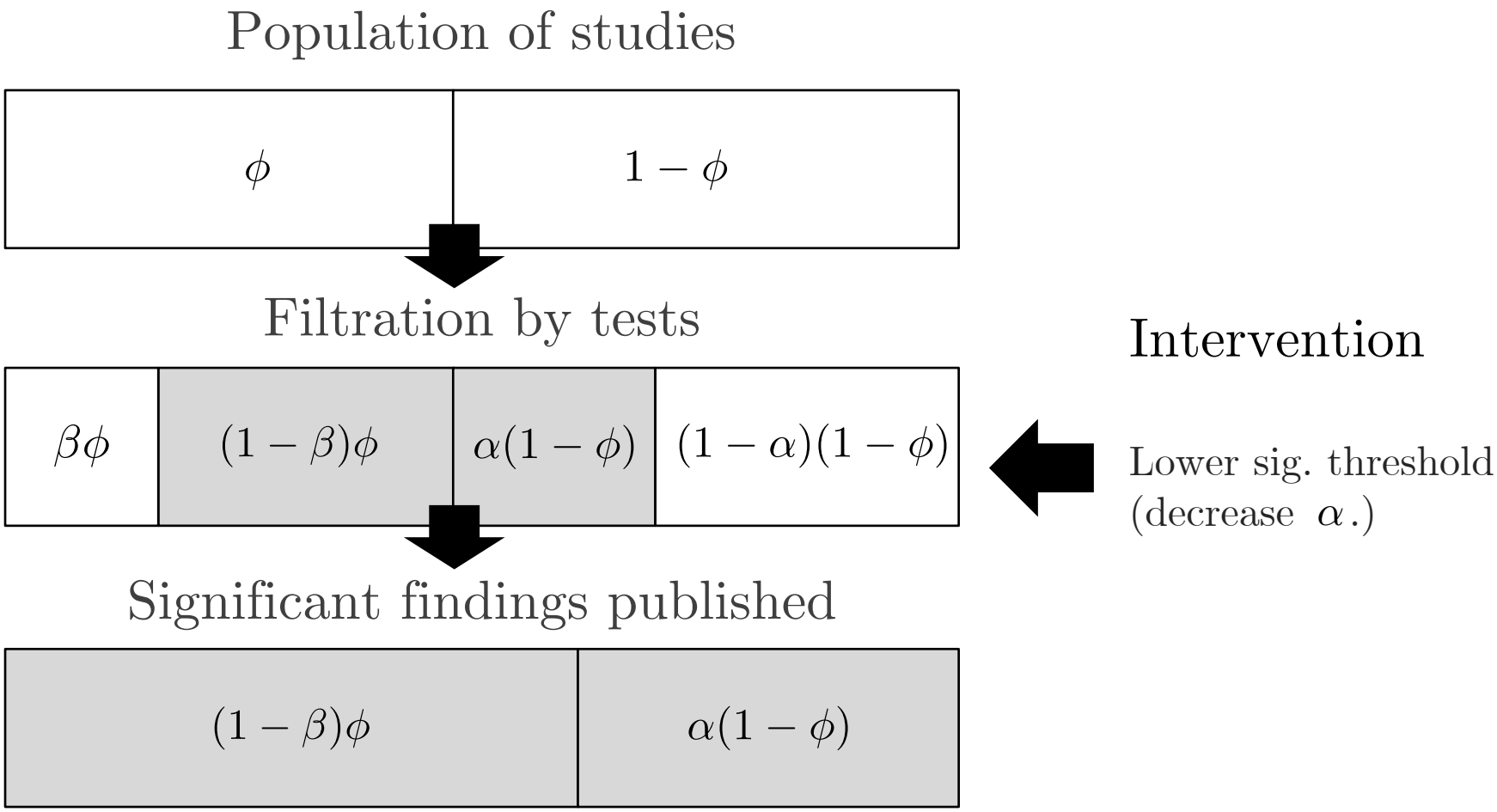
where BF is the Bayes factor that represents the evidence from the data, and the prior odds can be informed by researchers' beliefs, scientific consensus, and validated evidence from similar research questions in the same field. Multiple-hypothesis testing, *P*-hacking and publication bias all reduce 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].

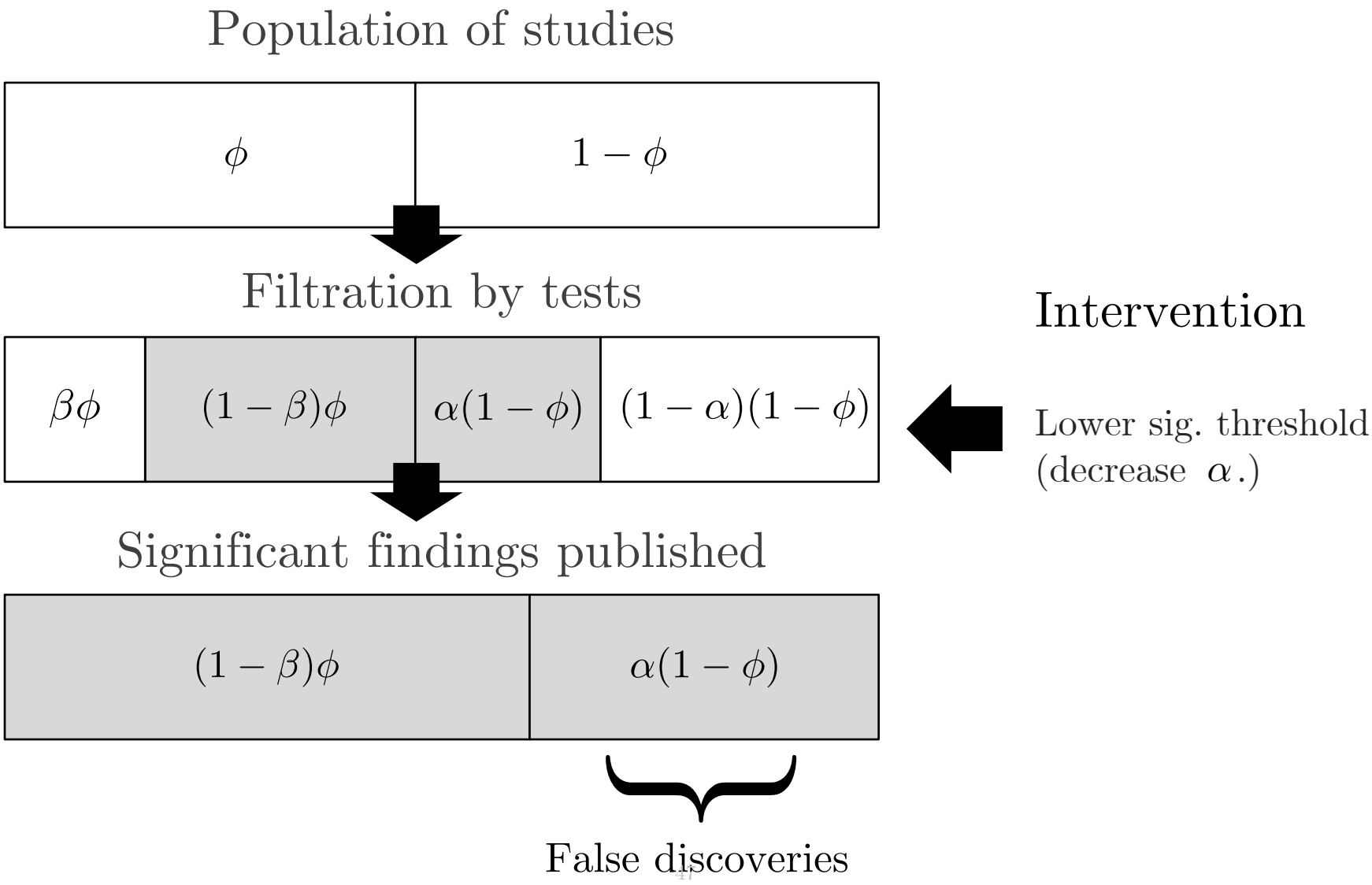
There is no unique mapping between the *P* value and the Bayes factor, since the Bayes factor depends on H_1. However, the connection between the two quantities can be evaluated for particular test statistics under certain classes of plausible alternatives (Fig. 1).

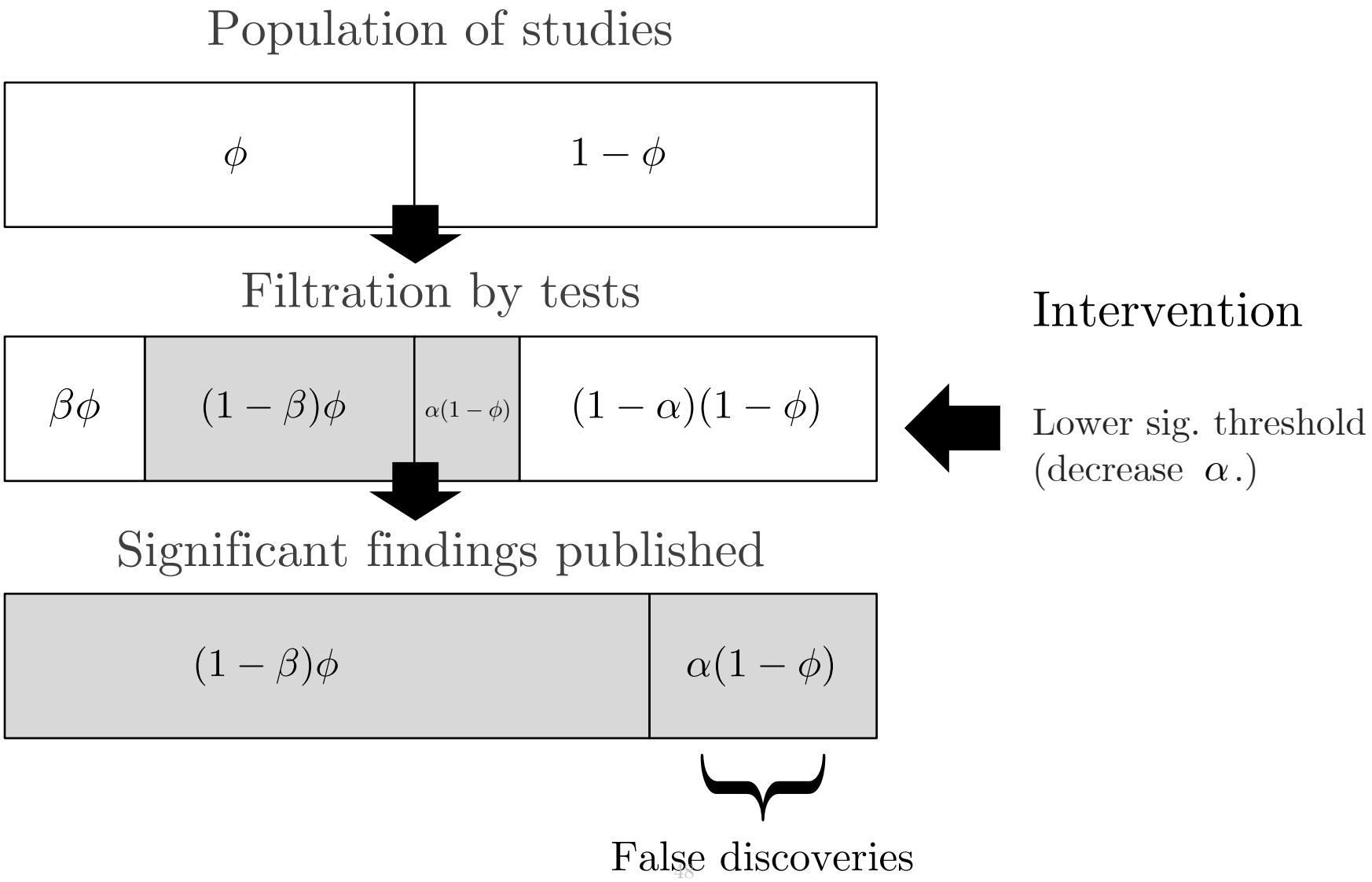
© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

This is not an unreasonable proposal.


Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
Significant findings published
$(1-\beta)\phi$
$\alpha(1-\phi)$

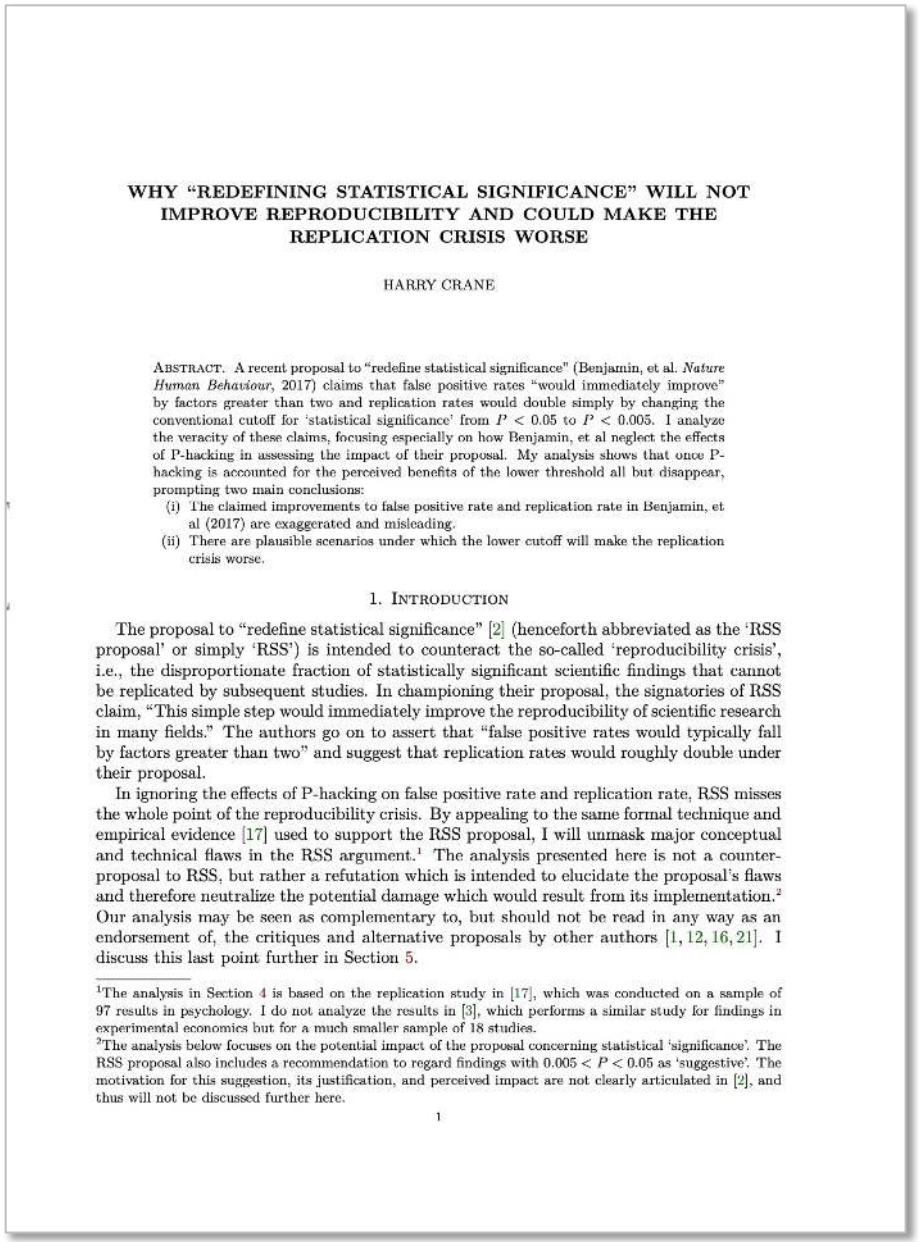

Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
Intervention
Lower sig. threshold
(decrease α.)
Significant findings published
$(1-\beta)\phi$
$\alpha(1-\phi)$


Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
Intervention
Lower sig. threshold
(decrease α.)
Significant findings published
$(1-\beta)\phi$
$\alpha(1-\phi)$
False discoveries


Population of studies
ϕ
$1-\phi$
Filtration by tests
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
Intervention
Lower sig. threshold
(decrease α.)
Significant findings published
$(1-\beta)\phi$
$\alpha(1-\phi)$
False discoveries

Why "redefining statistical significance will not improve reproducibility"

QRPs can cause the intervention to worsen false discovery rates.



WHY "REDEFINING STATISTICAL SIGNIFICANCE" WILL NOT IMPROVE REPRODUCIBILITY AND COULD MAKE THE REPLICATION CRISIS WORSE

HARRY CRANE

ABSTRACT. A recent proposal to "redefine statistical significance" (Benjamin, et al. *Nature Human Behaviour*, 2017) claims that false positive rates "would immediately improve" by factors greater than two and replication rates would double simply by changing the conventional cutoff for 'statistical significance' from $P < 0.05$ to $P < 0.005$. I analyze the veracity of these claims, focusing especially on how Benjamin, et al neglect the effects of P-hacking in assessing the impact of their proposal. My analysis shows that once P-hacking is accounted for the perceived benefits of the lower threshold all but disappear, prompting two main conclusions:

(i) The claimed improvements to false positive rate and replication rate in Benjamin, et al (2017) are exaggerated and misleading.

(ii) There are plausible scenarios under which the lower cutoff will make the replication crisis worse.

1. INTRODUCTION

The proposal to "redefine statistical significance" [2] (henceforth abbreviated as the 'RSS proposal' or simply 'RSS') is intended to counteract the so-called 'reproducibility crisis', i.e., the disproportionate fraction of statistically significant scientific findings that cannot be replicated by subsequent studies. In championing their proposal, the signatories of RSS claim, "This simple step would immediately improve the reproducibility of scientific research in many fields." The authors go on to assert that "false positive rates would typically fall by factors greater than two" and suggest that replication rates would roughly double under their proposal.

In ignoring the effects of P-hacking on false positive rate and replication rate, RSS misses the whole point of the reproducibility crisis. By appealing to the same formal technique and empirical evidence [17] used to support the RSS proposal, I will unmask major conceptual and technical flaws in the RSS argument.[1] The analysis presented here is not a counter-proposal to RSS, but rather a refutation which is intended to elucidate the proposal's flaws and therefore neutralize the potential damage which would result from its implementation.[2] Our analysis may be seen as complementary to, but should not be read in any way as an endorsement of, the critiques and alternative proposals by other authors [1, 12, 16, 21]. I discuss this last point further in Section 5.

[1]The analysis in Section 4 is based on the replication study in [17], which was conducted on a sample of 97 results in psychology. I do not analyze the results in [3], which performs a similar study for findings in experimental economics but for a much smaller sample of 18 studies.

[2]The analysis below focuses on the potential impact of the proposal concerning statistical 'significance'. The RSS proposal also includes a recommendation to regard findings with $0.005 < P < 0.05$ as 'suggestive'. The motivation for this suggestion, its justification, and perceived impact are not clearly articulated in [2], and thus will not be discussed further here.

1

"Adopting lower p-value thresholds may. . . [produce] collateral harms. . . bias may escalate rather than decrease if researchers. . . try to find ways to make the results have lower p-values."

—Ioannidis 2018, "The proposal to lower P value thresholds to .005," p. 1430

EXAMPLE A

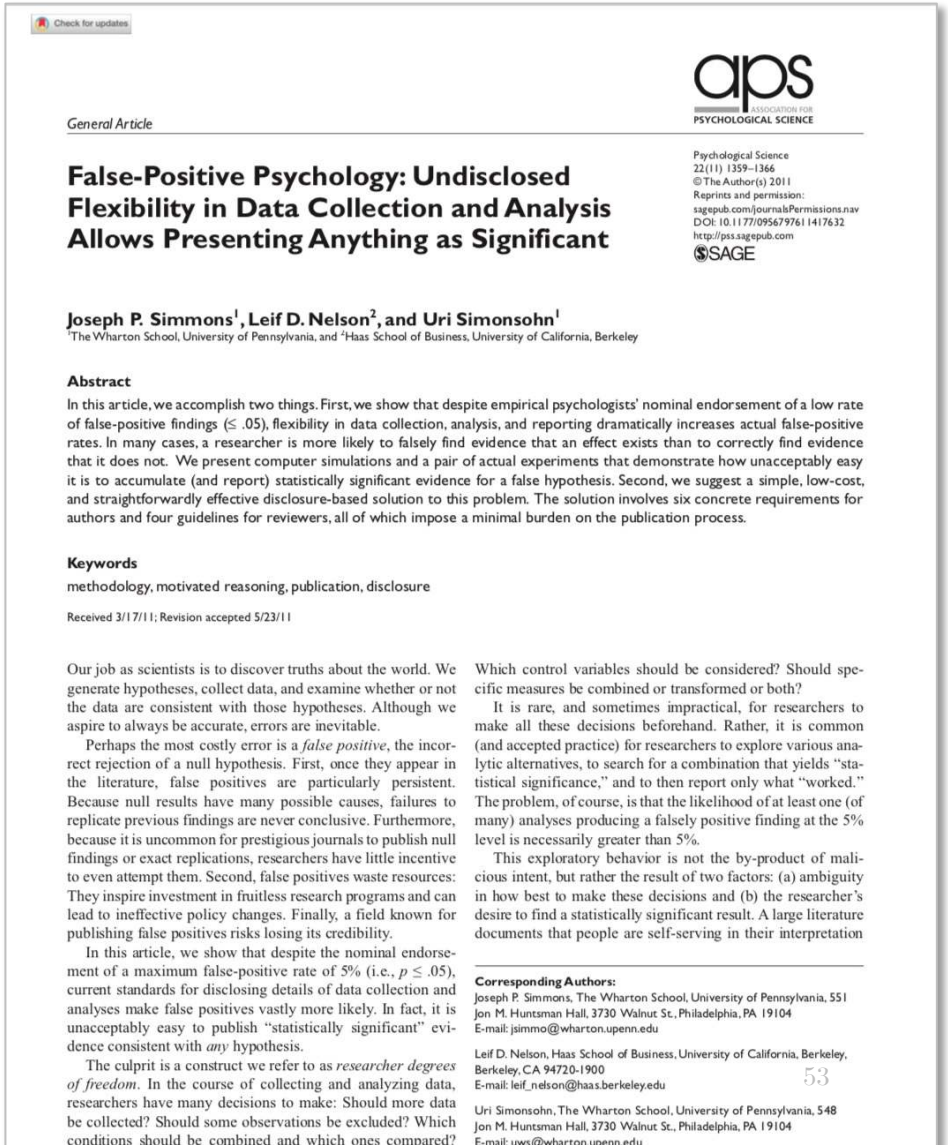
Intervention. Lowering significance threshold

Background problem 1. Variation in study soundness

Background problem 2. Publication bias

EXAMPLE A

Observation. Studies employing more QRPs can attain significance more effectively than those that do not



Check for updates

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

General Article

Psychological Science
22(11) 1359–1366
©The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
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

Abstract

In this article, we accomplish two things. First, we show that despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤ .05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. We present computer simulations and a pair of actual experiments that demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis. Second, we suggest a simple, low-cost, and straightforwardly effective disclosure-based solution to this problem. The solution involves six concrete requirements for authors and four guidelines for reviewers, all of which impose a minimal burden on the publication process.

Keywords

methodology, motivated reasoning, publication, disclosure

Received 3/17/11; Revision accepted 5/23/11

Our job as scientists is to discover truths about the world. We generate hypotheses, collect data, and examine whether or not the data are consistent with those hypotheses. Although we aspire to always be accurate, errors are inevitable.

Perhaps the most costly error is a *false positive*, the incorrect rejection of a null hypothesis. First, once they appear in the literature, false positives are particularly persistent. Because null results have many possible causes, failures to replicate previous findings are never conclusive. Furthermore, because it is uncommon for prestigious journals to publish null findings or exact replications, researchers have little incentive to even attempt them. Second, false positives waste resources: They inspire investment in fruitless research programs and can lead to ineffective policy changes. Finally, a field known for publishing false positives risks losing its credibility.

In this article, we show that despite the nominal endorsement of a maximum false-positive rate of 5% (i.e., $p \leq .05$), current standards for disclosing details of data collection and analyses make false positives vastly more likely. In fact, it is unacceptably easy to publish "statistically significant" evidence consistent with *any* hypothesis.

The culprit is a construct we refer to as *researcher degrees of freedom*. In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?

It is rare, and sometimes impractical, for researchers to make all these decisions beforehand. Rather, it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields "statistical significance," and to then report only what "worked." The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.

This exploratory behavior is not the by-product of malicious intent, but rather the result of two factors: (a) ambiguity in how best to make these decisions and (b) the researcher's desire to find a statistically significant result. A large literature documents that people are self-serving in their interpretation

Corresponding Authors:
Joseph P. Simmons, The Wharton School, University of Pennsylvania, 551 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104
E-mail: jsimmo@wharton.upenn.edu

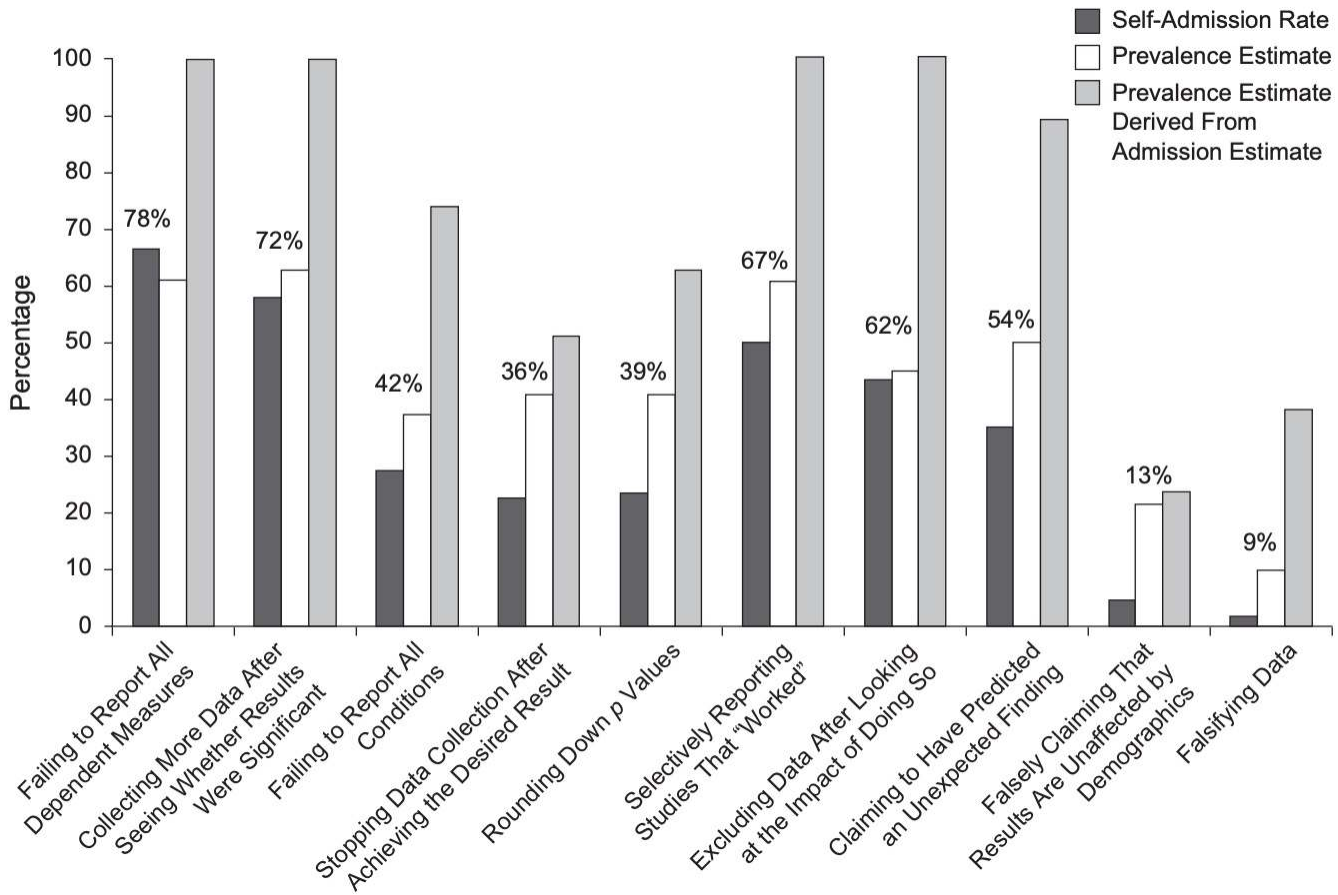
Leif D. Nelson, Haas School of Business, University of California, Berkeley, Berkeley, CA 94720-1900
E-mail: leif_nelson@haas.berkeley.edu

Uri Simonsohn, The Wharton School, University of Pennsylvania, 548 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104
E-mail: uws@wharton.upenn.edu

Effects of QRPs

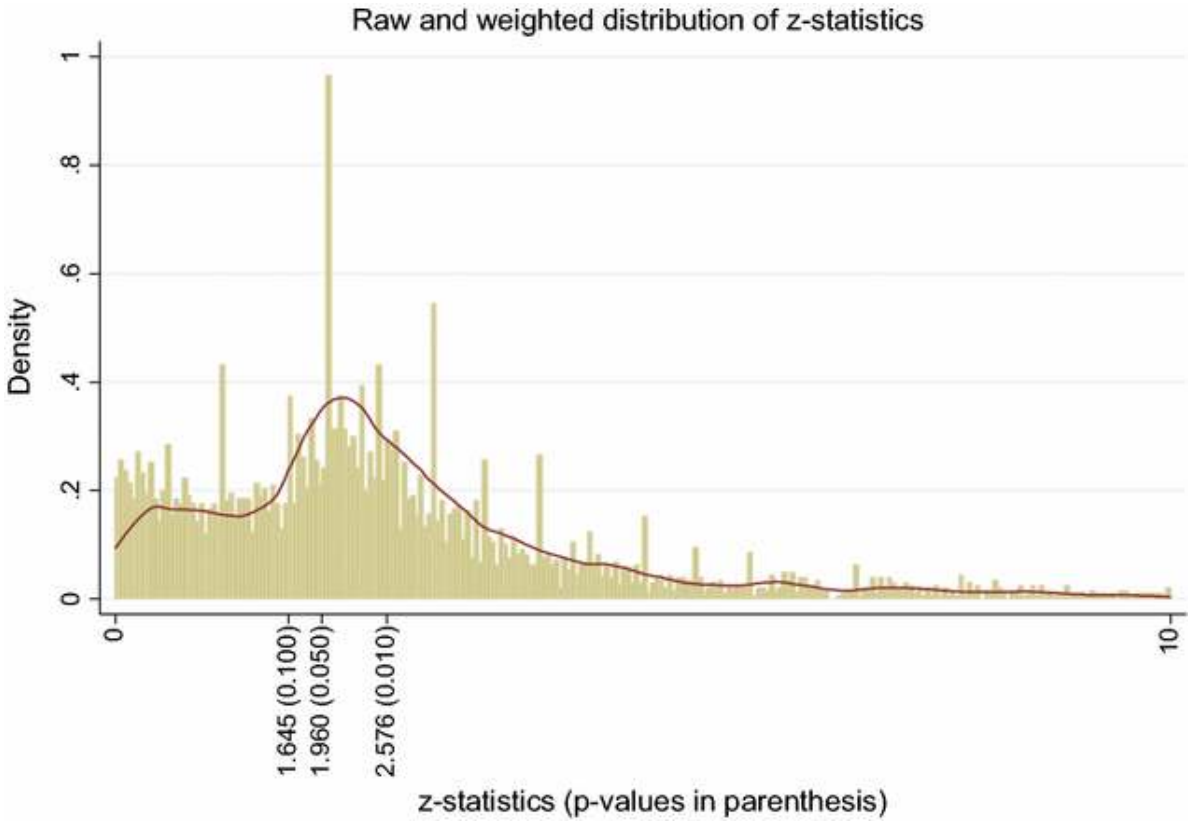
With simple optional stopping and selectively controlling for 3 covariates, Simmons et al (2011) obtain a false positive rate of 61%.

Estimates of QRPs



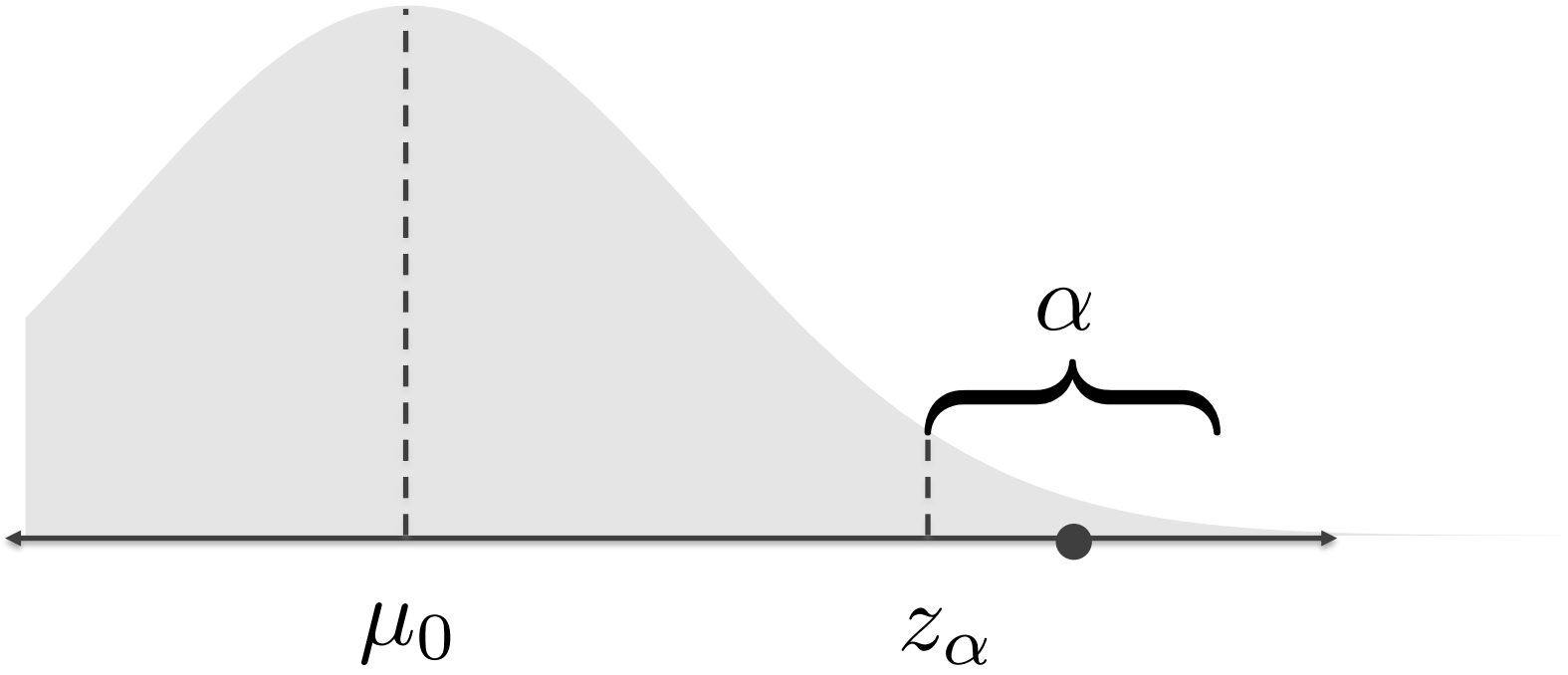
(John et al, 2012)

P-value distributions of published findings



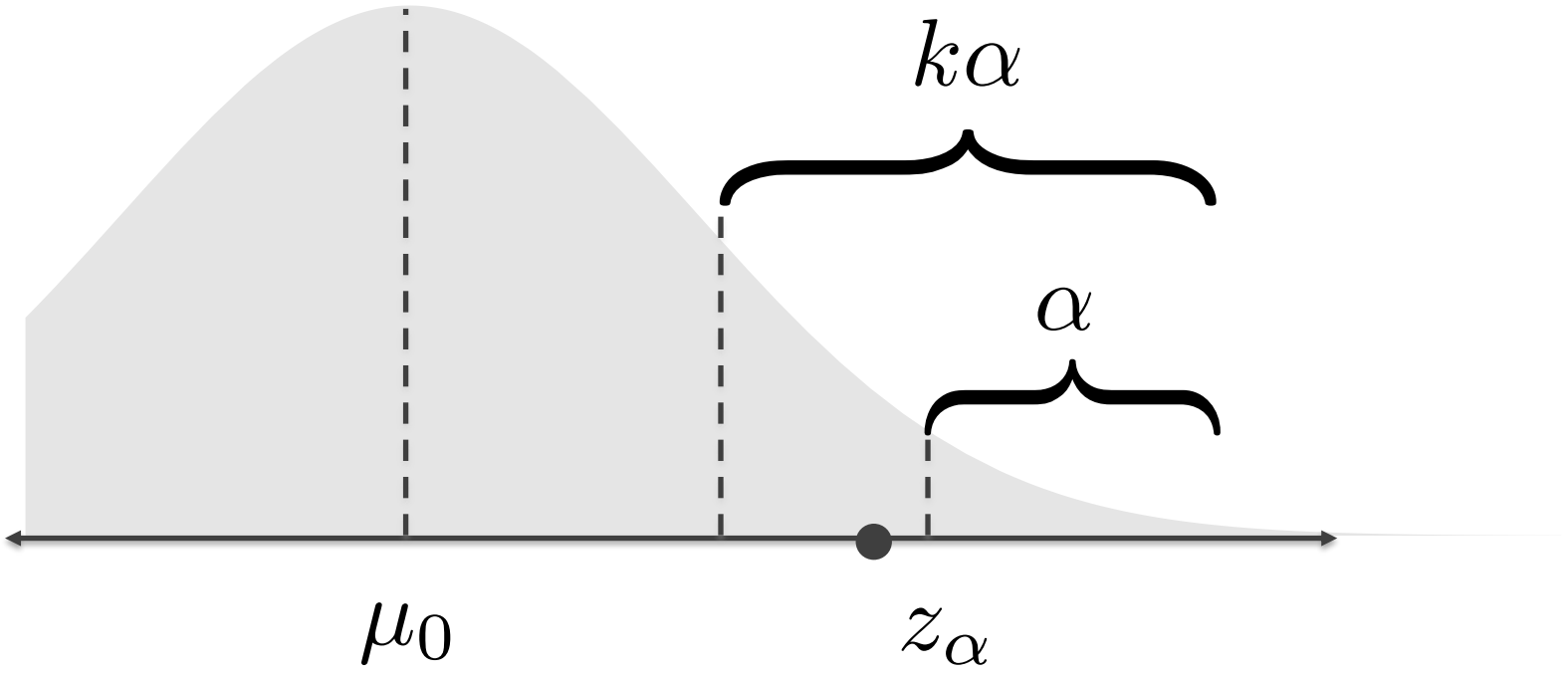
Methodological soundness

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

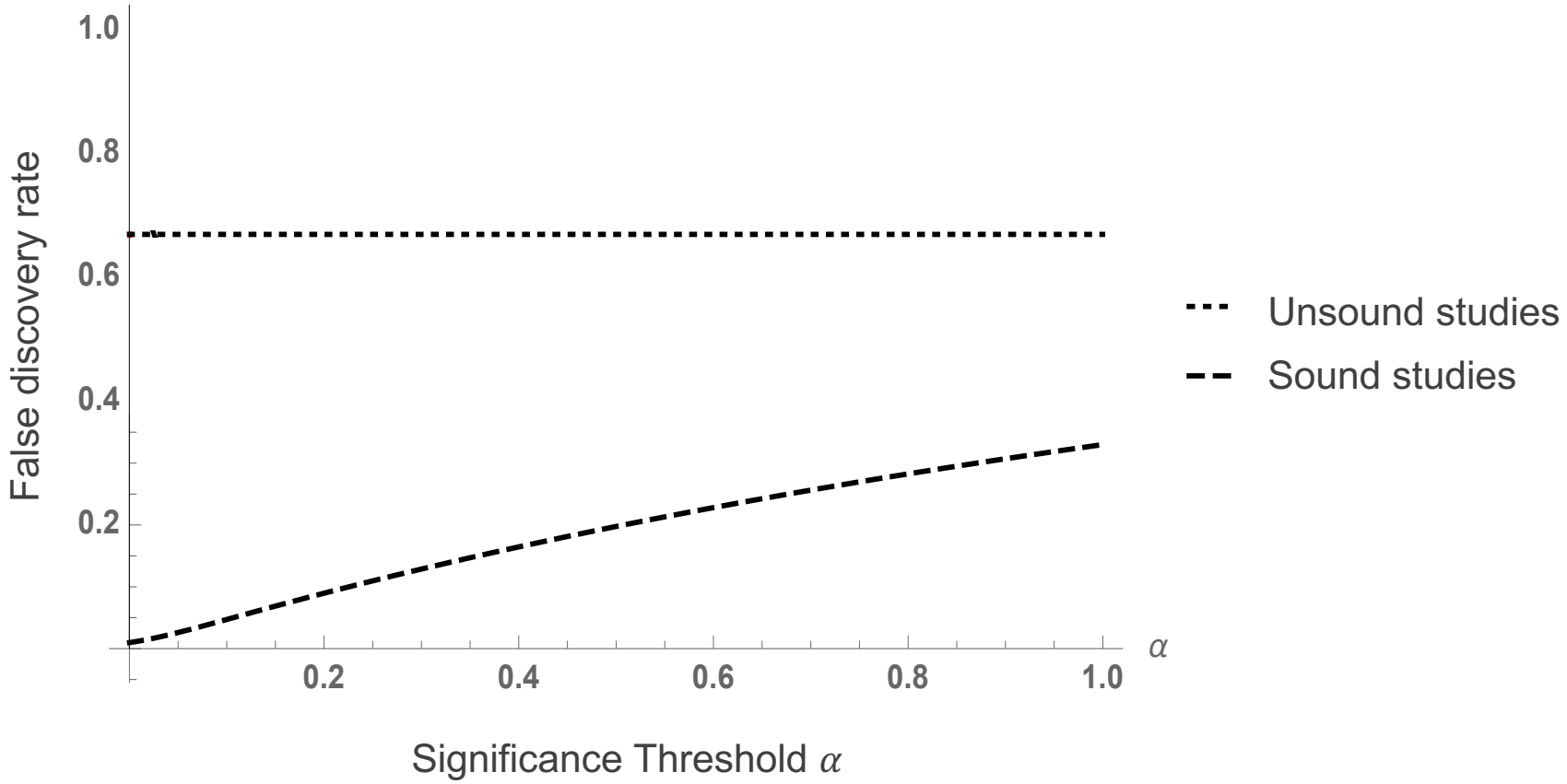


Methodological soundness

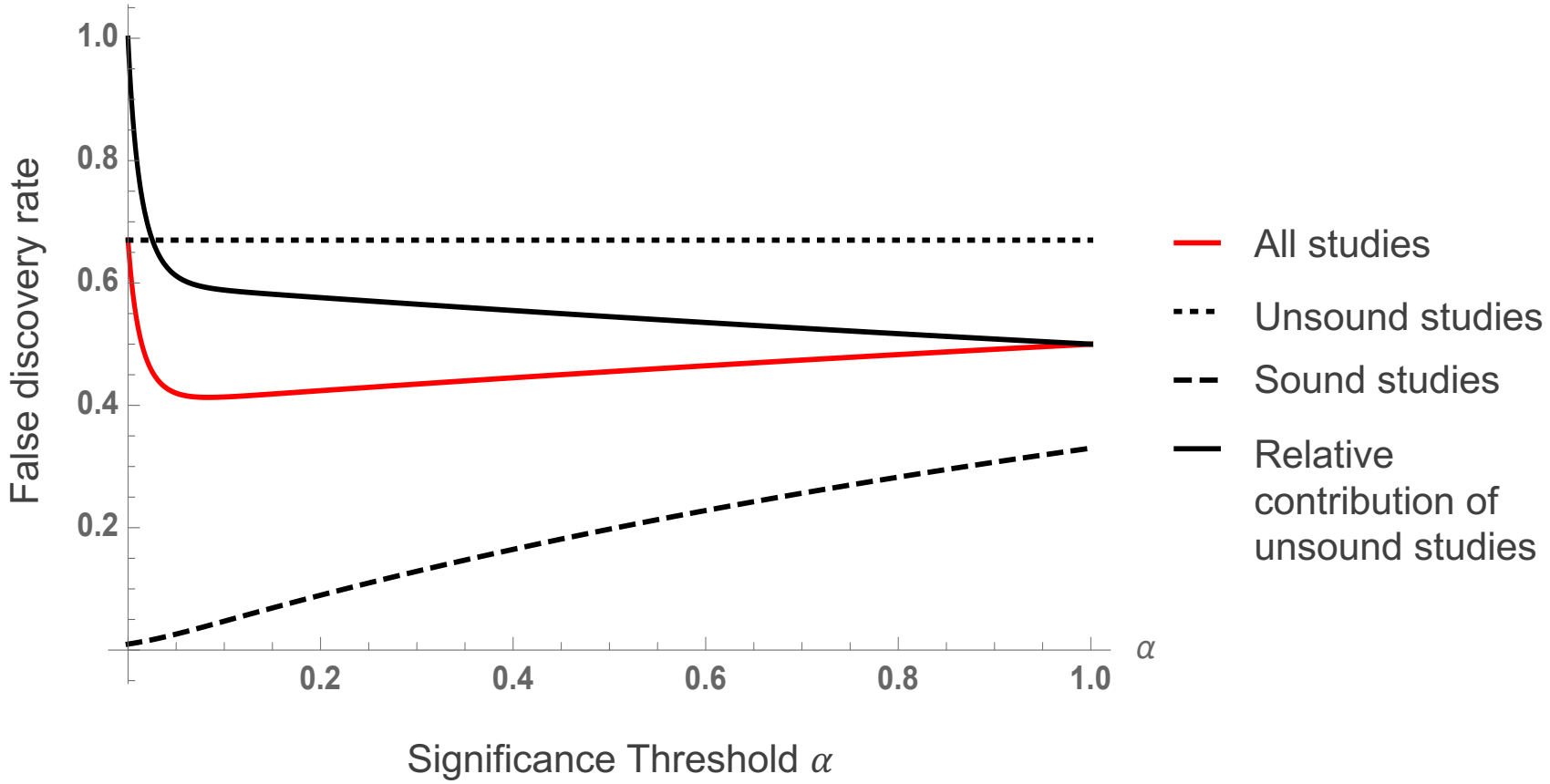
A *less-than-perfectly-sound* study will attain significance when the null hypothesis is true with a frequency *greater* than the significance threshold by some factor $k > 1$.

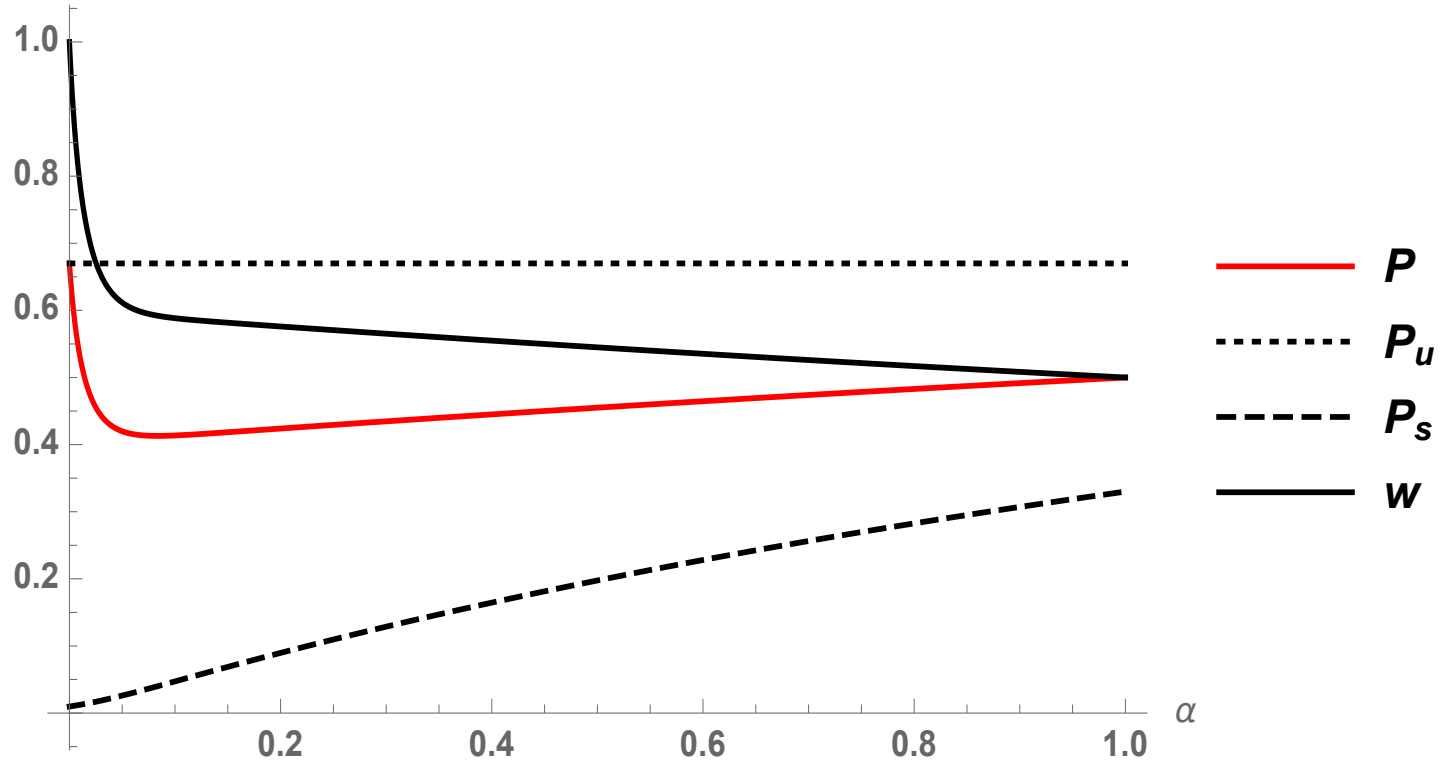


Effect of lowering the significance threshold on sound and unsound studies



Effect of lowering the significance threshold on a community of studies



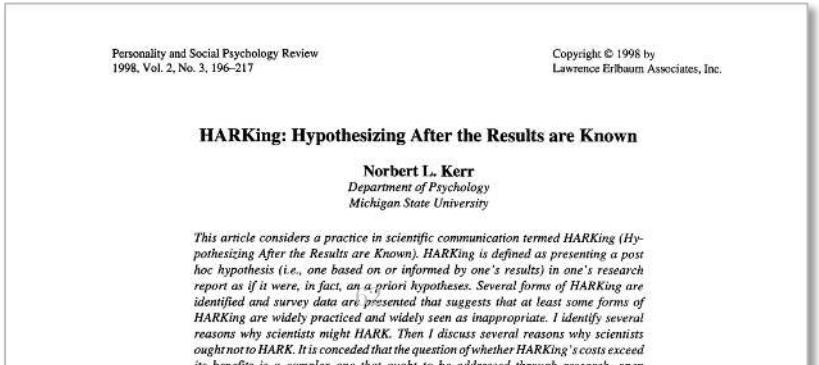

1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0
α
P
P_u
P_s
w

Simulation Result 1. If studies vary in soundness then lowering the significance threshold can increase the false discovery rate of the published literature.

HARKing

HARKing [hypothesizing after results are known] is the presentation of a hypothesis selected ex post to observing the data as if it were one selected ex ante to observing the data.

—Norbert Kerr, *HARKing*, 1998, p. 197



Personality and Social Psychology Review
1998, Vol. 2, No. 3, 196–217

Copyright © 1998 by
Lawrence Erlbaum Associates, Inc.

HARKing: Hypothesizing After the Results are Known

Norbert L. Kerr
Department of Psychology
Michigan State University

This article considers a practice in scientific communication termed HARKing (Hypothesizing After the Results are Known). HARKing is defined as presenting a post hoc hypothesis (i.e., one based on or informed by one's results) in one's research report as if it were, in fact, an a priori hypotheses. Several forms of HARKing are identified and survey data are presented that suggests that at least some forms of HARKing are widely practiced and widely seen as inappropriate. I identify several reasons why scientists might HARK. Then I discuss several reasons why scientists ought not to HARK. It is conceded that the question of whether HARKing's costs exceed

HARKing

HARKing is one the four horsemen of the replicability apocalypse [along with "publication bias, low statistical power, and *p*-hacking."]

—Dorothy Bishop, *Nature*, 2019, p. 435

EXAMPLE B

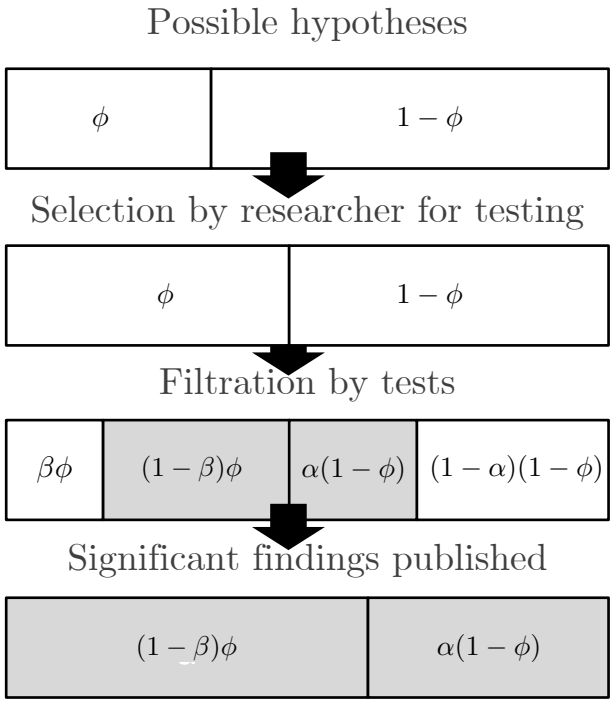
Intervention. Lowering the significance threshold

Background problem 1. fallback HARKing and informed researcher judgement

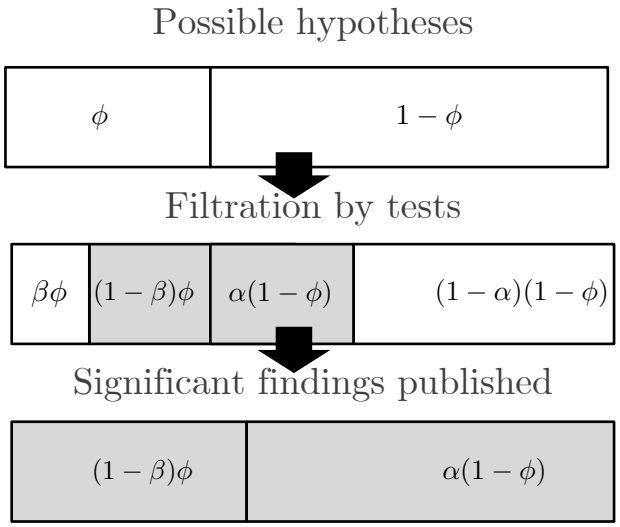
Background problem 2. Publication bias

Modeling fallback HARKing

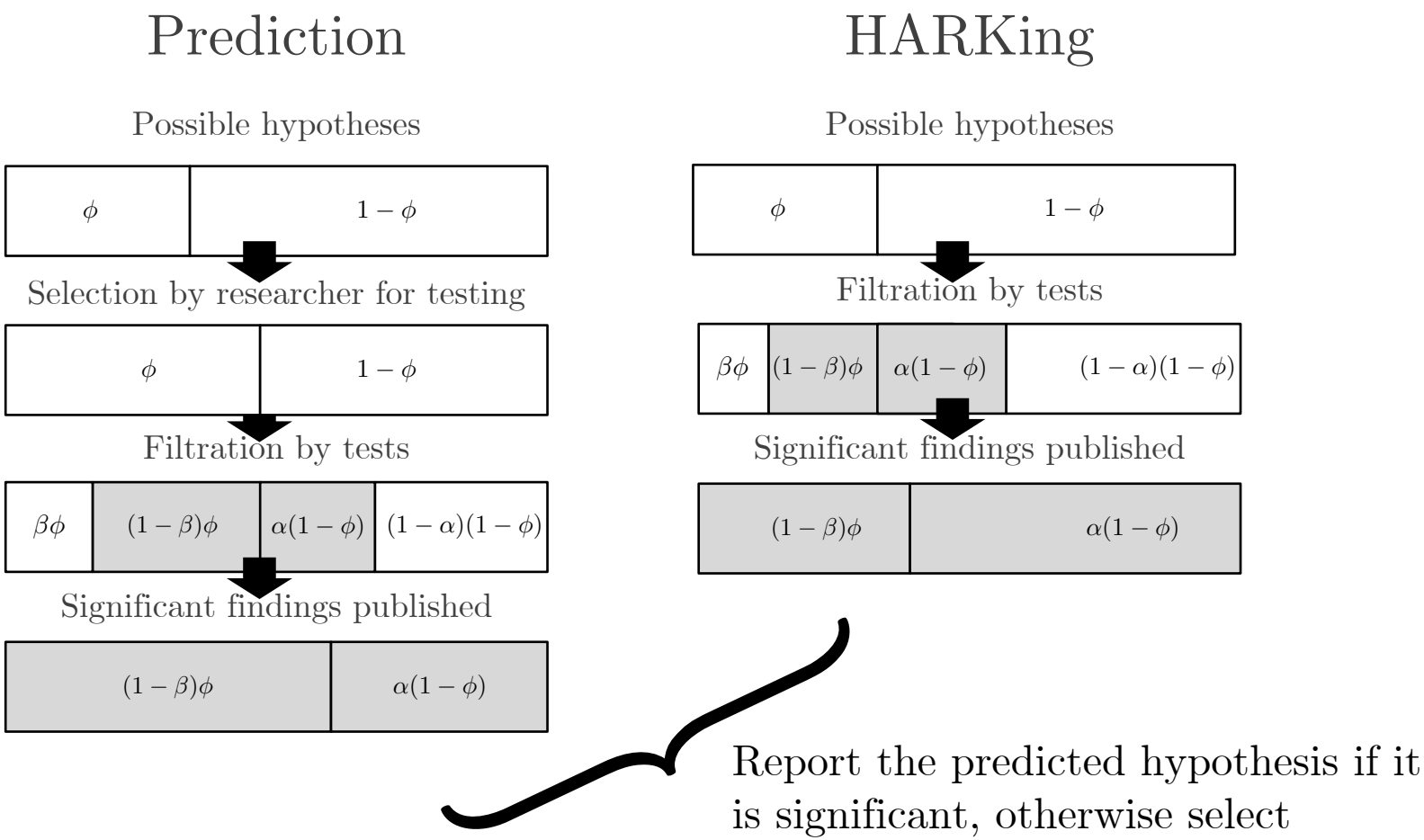
Prediction



HARKing



Modeling fallback HARKing

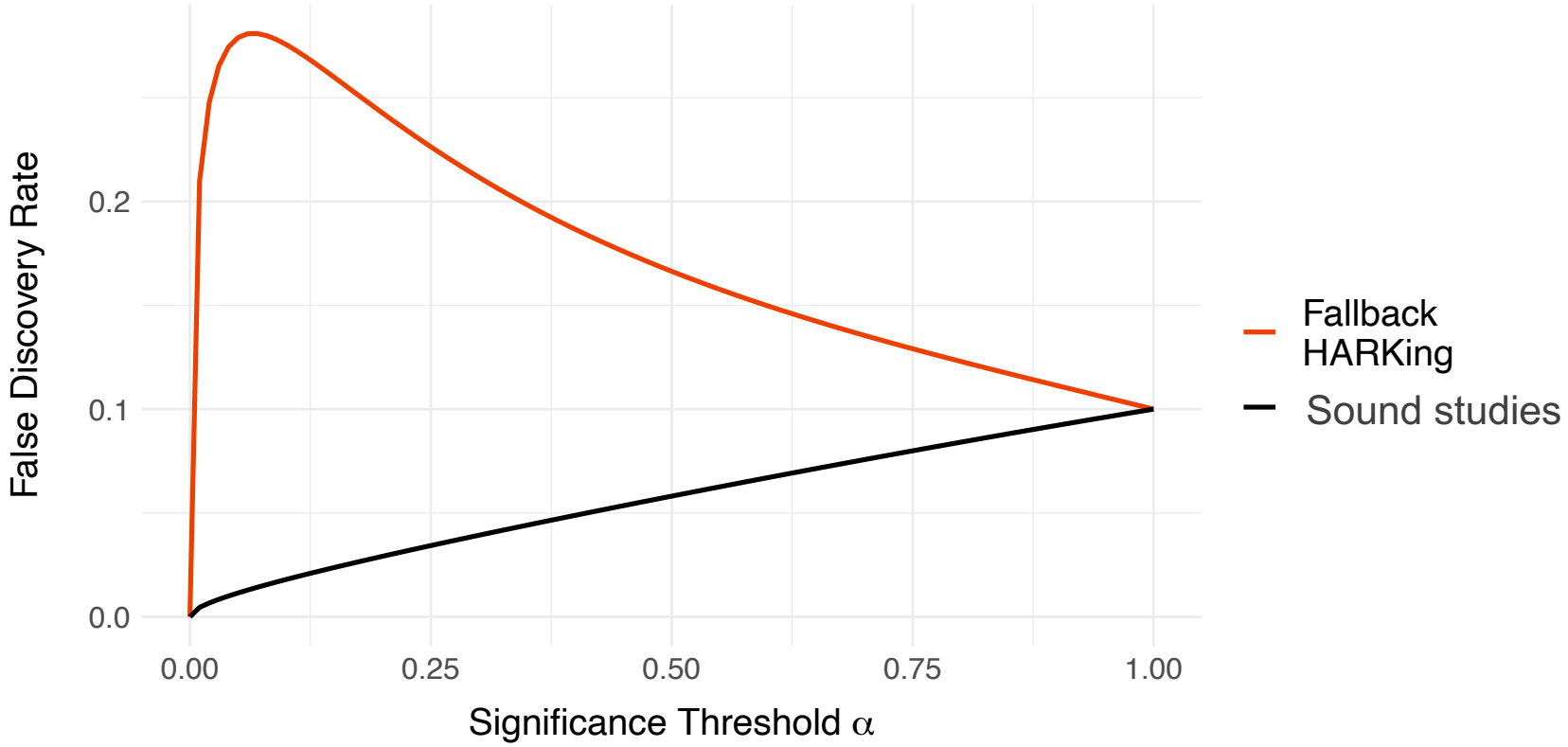


Report the predicted hypothesis if it is significant, otherwise select (HARK) from the set of statistically significant auxiliary hypotheses.

Modeling fallback HARKing

Researcher predict when they can and HARK when they must.

Effects of lowering the significance threshold when studies involve fallback HARKing



Cf. Mohseni, 'HARKing: from Misdiagnosis to Misprescription'

Simulation Result 2. If researcher choice of hypotheses is informed, and researchers are engaged in fallback HARKing, then lowering the significance threshold can increase the false discovery rate of the literature.

Preregistration

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



PNAS

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.

Published under the PNAS license.

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

Published online March 12, 2018.

Registered reports

Study proposals are peer reviewed and pre-accepted by journals before research is undertaken.



nature human behaviour

REVIEW ARTICLE

https://doi.org/10.1038/s41562-021-01193-7

Check for updates

The past, present and future of Registered Reports

Christopher D. Chambers and Loukia Tzavella

Registered Reports are a form of empirical publication in which study proposals are peer reviewed and pre-accepted before research is undertaken. By deciding which articles are published based on the question, theory and methods, Registered Reports offer a remedy for a range of reporting and publication biases. Here, we reflect on the history, progress and future prospects of the Registered Reports initiative and offer practical guidance for authors, reviewers and editors. We review early evidence that Registered Reports are working as intended, while at the same time acknowledging that they are not a universal solution for irreproducibility. We also consider how the policies and practices surrounding Registered Reports are changing, or must change in the future, to address limitations and adapt to new challenges. We conclude that Registered Reports are promoting reproducibility, transparency and self-correction across disciplines and may help reshape how society evaluates research and researchers.

After more than a decade of meta-research and debate, the life and social sciences are well in the midst of a credibility revolution[1–3]. Faced with evidence of publication bias[4–7], hindsight bias and selective reporting[8–10], insufficient sample sizes[10], inadequate data sharing[11] and suboptimal rates of both attempted[10,12] and successful replication[13–20], researchers from across a broad range of fields are unifying around a core mission to improve reproducibility and transparency. In doing so, the deeper aim of the openness agenda is to stimulate cultural reform, aligning what is beneficial for individual scientists with what is beneficial for science[21].

As scientists and policymakers grapple with the causes of irreproducibility, it has become clear that one of its main drivers is the so-called 'results paradox'. On the one hand, scientists are taught from their earliest years that the one part of the research process that they must keep at arm's length is the results of their research. The objective investigator—the detective—follows the data with discipline and restraint, never pressuring it to bend to their will, lest they fall prey to Richard Feynman's famous warning that "the first principle is that you must not fool yourself and you are the easiest person to fool" (p. 12)[22]. On the other hand, the very same researcher is sent a message from prestigious journals, funding agencies and evaluation committees that if you want to succeed in science, be sure to publish a lot of clear, novel, positive findings. Researchers are therefore presented with conflicting goals: be a good detective who never indulges in data massaging or cherry picking, but also be a good lawyer who wins arguments and produces a continual supply of beautiful results[23].

Many observed problems with reproducibility stem from researchers attempting to resolve this paradox while protecting their careers. When publishing in prestigious journals requires confirming one's hypotheses but the results defy expectations, researchers can resolve this conflict by following the advice of Bem[24,25] and rewriting their hypotheses to 'predict' those results—a form of hindsight bias known as "hypothesizing after results are known" (HARKing)[9]. When academic leaders tell authors to "go with strongest studies" because "weak data dilute strong data" (p. 79)[26] and that "what you don't have to do is tell the whole truth … you can select the results you present"[27], the responsive researcher answers by reporting the analyses that tell the best story, diverting negative or inconvenient results to the file drawer or converting them into publishable (probably false) positives. And when journal editors tell researchers that, all else being equal, some results are simply more deserving of publication than others, the strategic researcher responds by conducting a large number of small studies and reporting only the most persuasive findings (even if unreliable) rather than gambling on the outcome of larger, more definitive projects that may yield inconclusive data[28].

Registered Reports (RRs) were proposed in 2012 as a way to free researchers from the pressure to engage in these counterproductive practices, thereby breaking the cycle that perpetuates bias and irreproducibility (Fig. 1). The RR model originates from the simple philosophy that to defeat the distorting effects of outcome bias on science, we must focus on the process and blind the evaluation of science to research outcomes[21]. This blinding is achieved by splitting peer review into two stages. In the first stage, authors submit their research question(s), theory, hypotheses, detailed methods and analysis plans and any preliminary data as needed. Following detailed review and revision—usually according to specific criteria—proposals that are favourably assessed receive in principle acceptance (IPA), which commits the journal to publishing the final paper regardless of whether the hypotheses are supported, provided that the authors adhere to their approved protocol and interpret the results in line with the evidence. Following IPA, authors then typically register their approved protocol in a repository, either publicly or under a temporary embargo. Then, after completing the research, they submit a stage 2 manuscript that includes the approved protocol plus the results and discussion, which may include clearly labelled post hoc analyses in addition to the preregistered outcomes (that is, findings from both confirmatory and exploratory analyses). The reviewers from stage 1 and/or newly invited reviewers then assess the completed stage 2 manuscript, focusing on compliance with the protocol and whether the conclusions are justified by the evidence. Crucially, reviewers do not relitigate the theory, hypotheses or methods, thereby preventing knowledge of the results from influencing recommendations. RR guidelines specify that editors similarly cannot reject a manuscript on the basis of any new concerns about the methodology or rationale or on the basis of the results themselves.

The aim of this modified review process is to reduce as much as possible the potential for biased research practices such as

Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University, Cardiff, UK. e-mail: chambersc1@cardiff.ac.uk

EXAMPLE C

Intervention. Voluntary preregistration

Background problem 1. Variation in study soundness

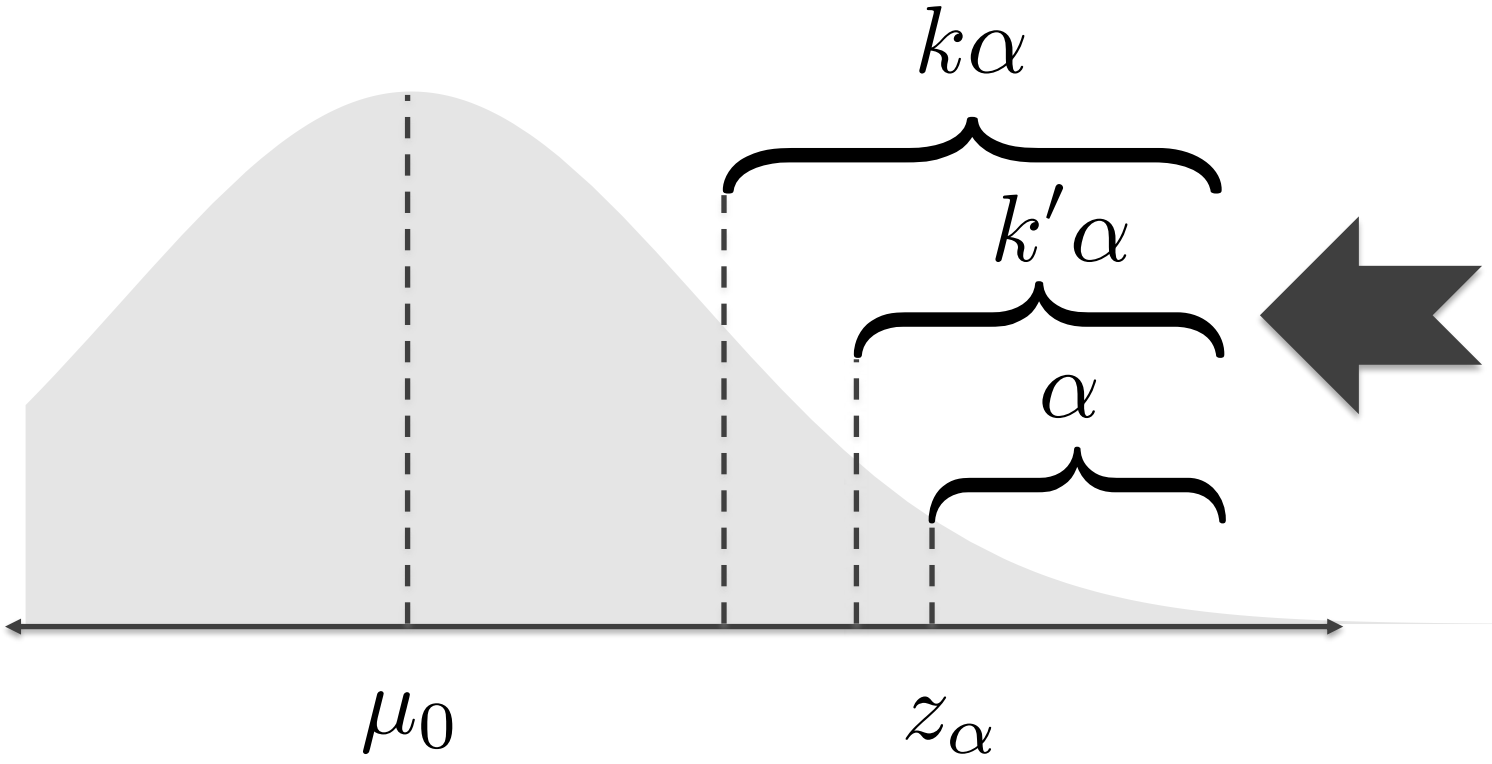
Background problem 2. Publication bias

EXAMPLE C

Observation. Voluntary preregistration and study soundness can be positively correlated

Modeling preregistration

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



Effectiveness of preregistration

Preregistration improves but does not completely mitigate methodological bias in the form of QRPs.



ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Claesen A, Gomes S, Tuerlinckx F, Vanpaemel W. 2021 Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies. *R. Soc. Open Sci.* **8**: 211037.
https://doi.org/10.1098/rsos.211037

Received: 23 June 2021
Accepted: 29 September 2021

Subject Category:
Psychology and cognitive neuroscience

Subject Areas:
psychology

Keywords:
preregistration, researcher degrees of freedom, open science, psychological science, transparency

Author for correspondence:
Aline Claesen
e-mail: aline.claesen@kuleuven.be

THE ROYAL SOCIETY PUBLISHING

Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies

Aline Claesen, Sara Gomes, Francis Tuerlinckx and Wolf Vanpaemel

Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, Leuven 3000, Belgium

AC, 0000-0002-0303-2441; SG, 0000-0002-2099-6314; FT, 0000-0002-1775-7654; WV, 0000-0002-5855-3885

Preregistration is a method to increase research transparency by documenting research decisions on a public, third-party repository prior to any influence by data. It is becoming increasingly popular in all subfields of psychology and beyond. Adherence to the preregistration plan may not always be feasible and even is not necessarily desirable, but without disclosure of deviations, readers who do not carefully consult the preregistration plan might get the incorrect impression that the study was exactly conducted and reported as planned. In this paper, we have investigated adherence and disclosure of deviations for all articles published with the Preregistered badge in *Psychological Science* between February 2015 and November 2017 and shared our findings with the corresponding authors for feedback. Two out of 27 preregistered studies contained no deviations from the preregistration plan. In one study, all deviations were disclosed. Nine studies disclosed none of the deviations. We mainly observed (un)disclosed deviations from the plan regarding the reported sample size, exclusion criteria and statistical analysis. This closer look at preregistrations of the first generation reveals possible hurdles for reporting preregistered studies and provides input for future reporting guidelines. We discuss the results and possible explanations, and provide recommendations for preregistered research.

1. Introduction

During the research process, researchers inevitably make numerous decisions, collectively known as researcher degrees of freedom [1]. Among others, researchers need to decide on

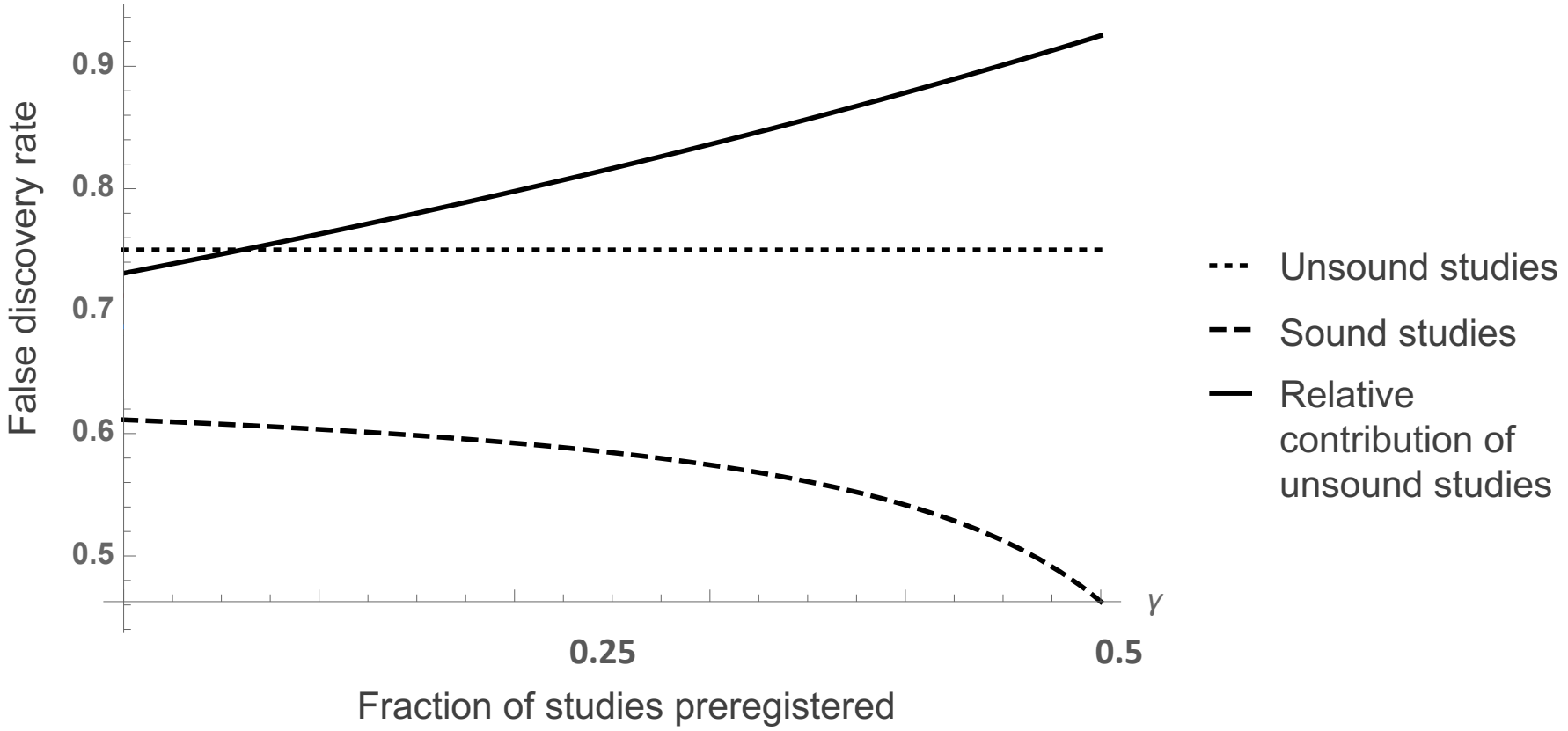
© 2021 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

A population-level problem

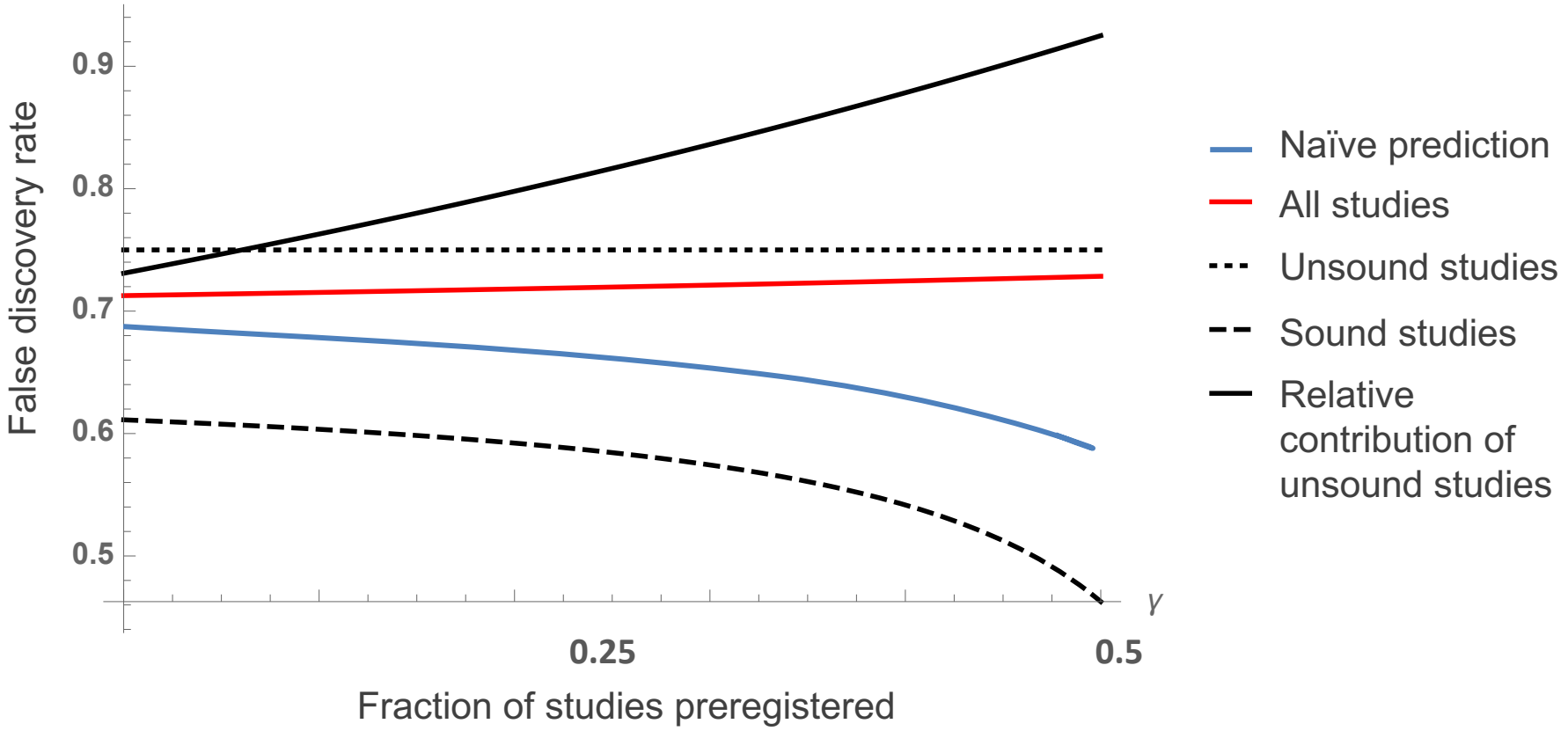
Voluntary preregistration can make it so:

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

Effect of voluntary preregistration on false discovery under *heterogeneity*



Effect of voluntary preregistration on false discovery under *heterogeneity*



Simulation Result 3. If preregistration and study soundness are positively correlated (e.g., via researcher disposition toward each), then *voluntary preregistration* can increase the false discovery rate of the literature.

EXAMPLE D

Intervention. Voluntary preregistration

Background problem 1. Publication bias

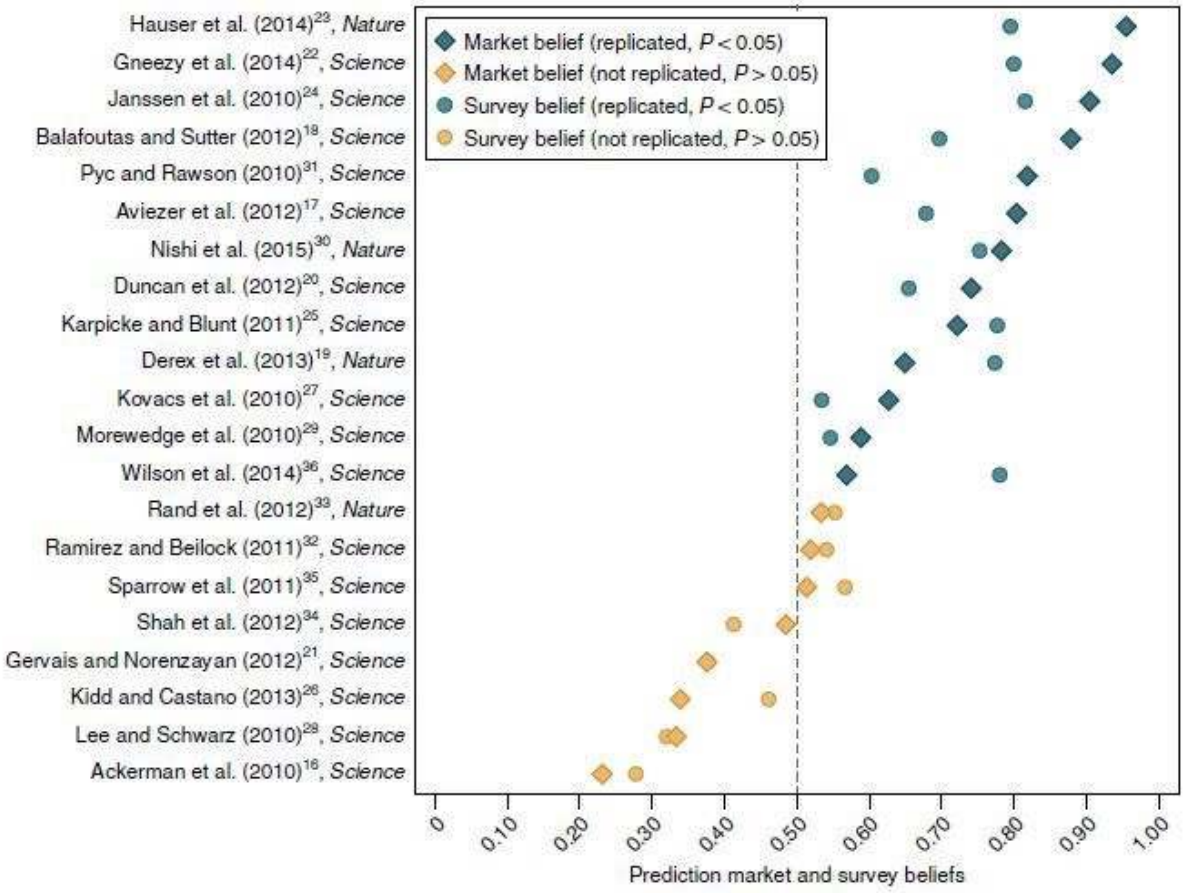
Background fact 2. Variation in prior odds of hypotheses between studies

EXAMPLE D

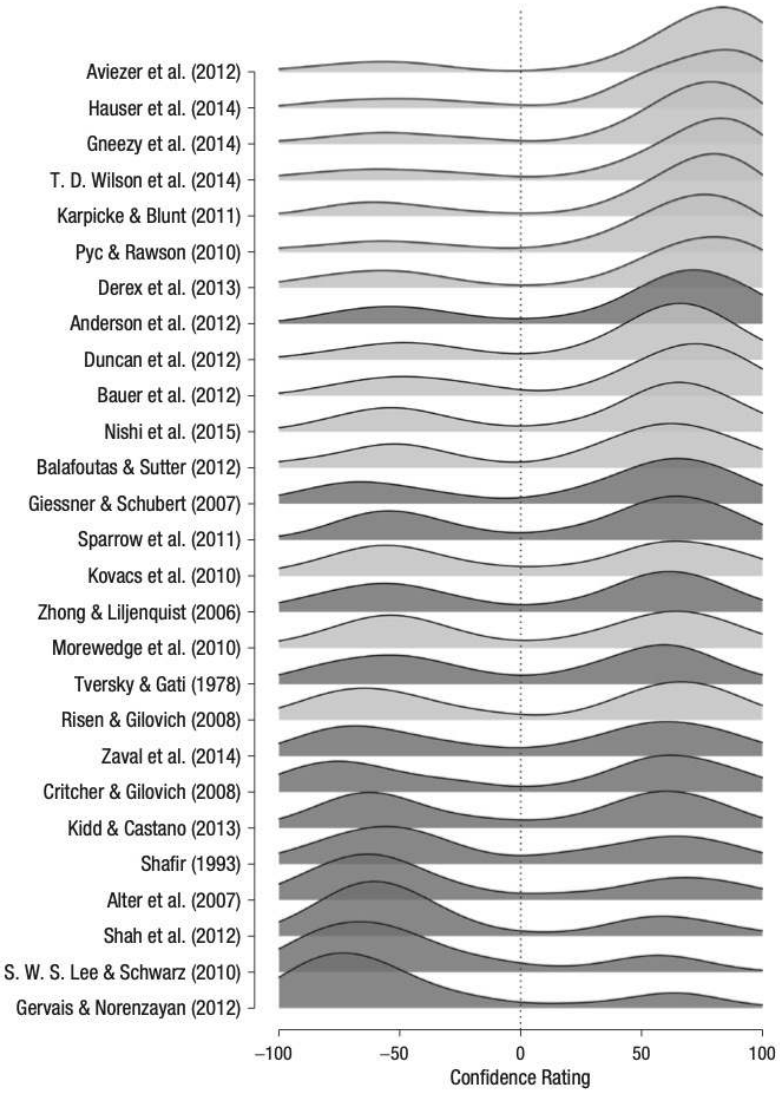
Observation. Voluntary preregistration and study hypothesis prior odds can be positively correlated

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

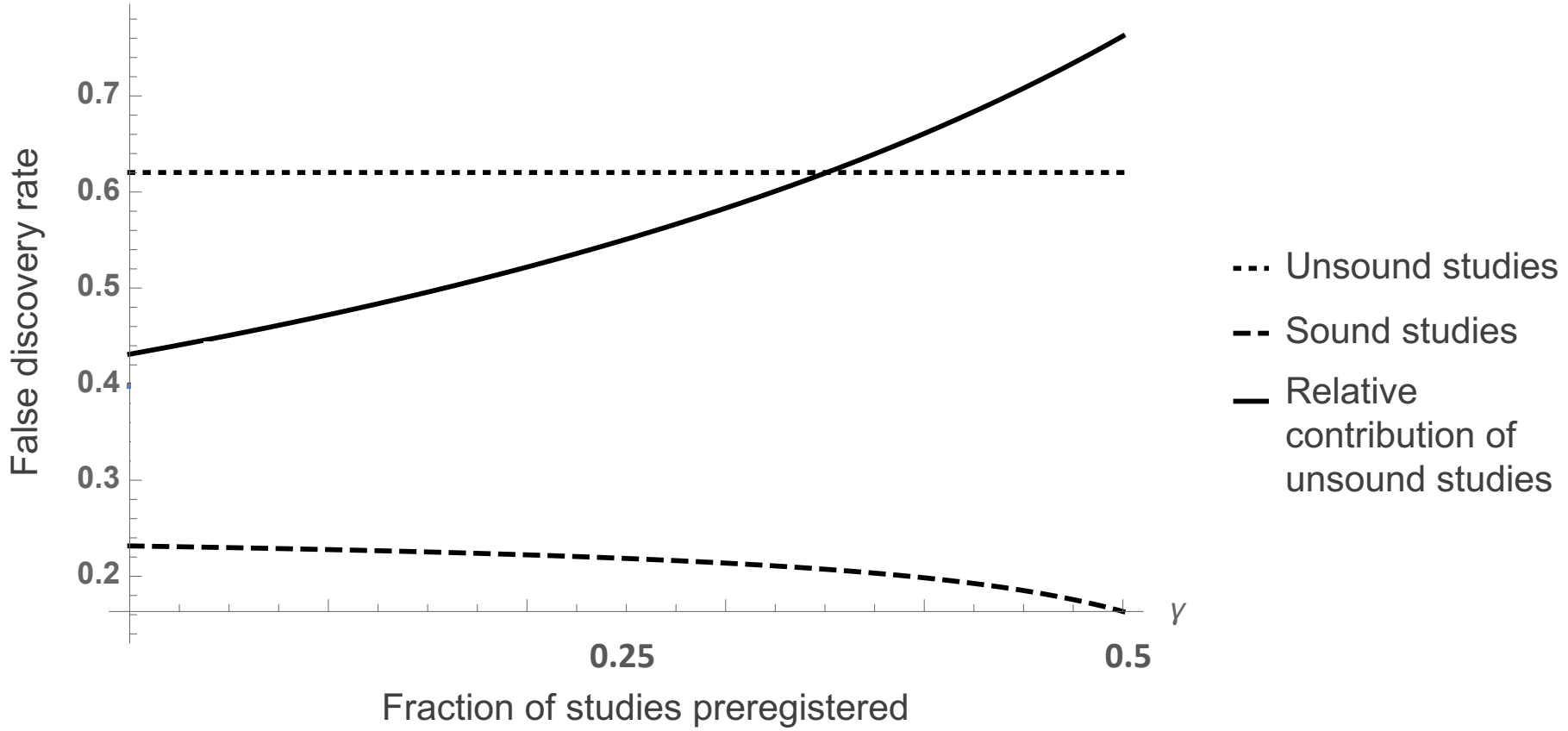


(Camerer et al, 2018)

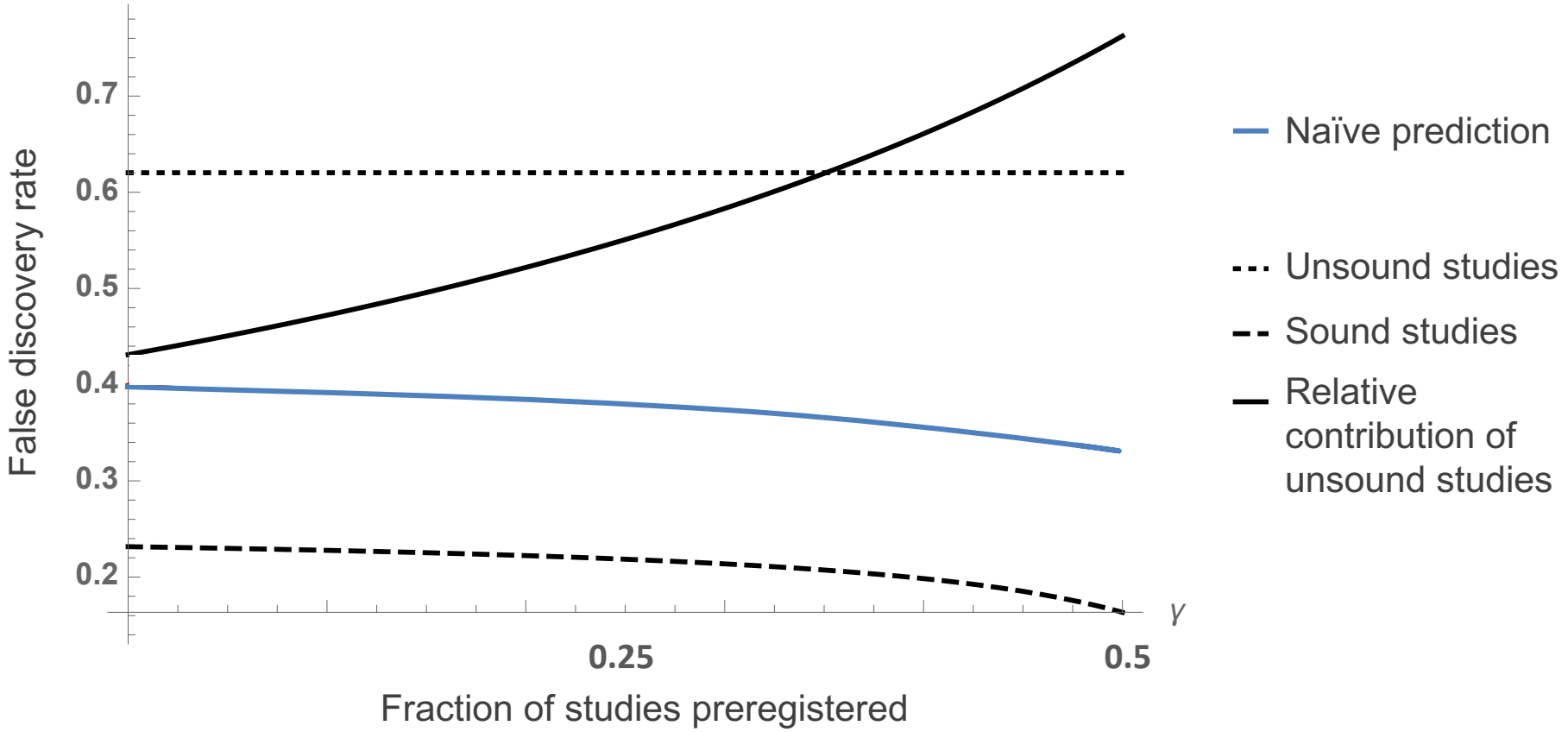


(Hoogeveen et al, 2020)

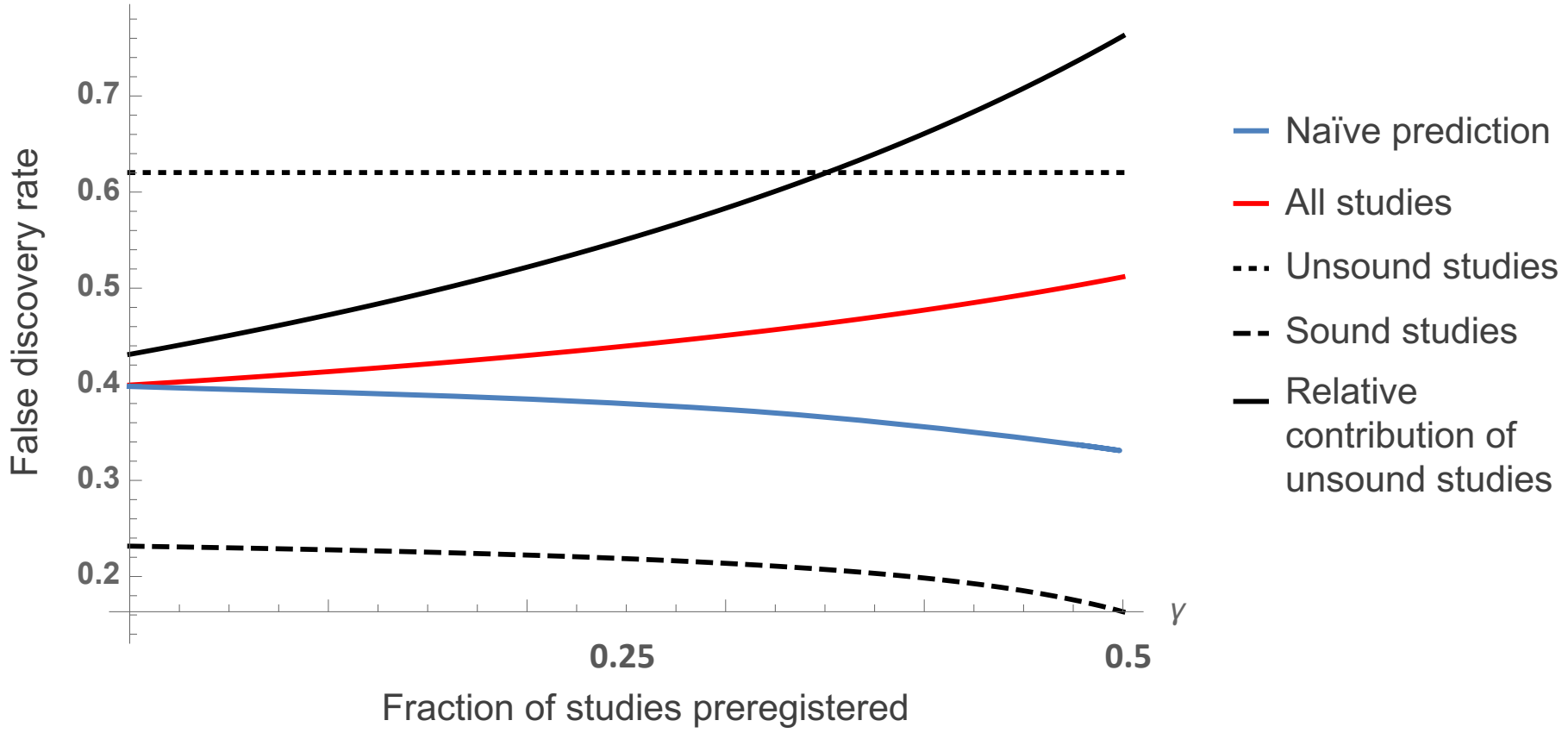
Effect of preregistration on false discovery with correlation via researcher judgment



Effect of preregistration on false discovery with correlation via researcher judgment



Effect of preregistration on false discovery with correlation via researcher judgment



Simulation Result 4. If preregistration and study prior odds are positively correlated—e.g., via researcher judgment of prior odds affecting researcher disposition to preregister—then *voluntary preregistration* can increase the false discovery rate of the literature.

We can say something general about the conditions under which interventions in can produce such effects.

Theorem 5 (Soundness-dependent selection). If the magnitude of the effect of the intervention on reducing the type I error rate of studies is positively correlated with study unsoundness, then that intervention *can* increase the proportion of significant outcomes that are of types with the highest false positive rates, thereby increasing the false discovery rate of the literature.

Indeed, in the the presence of maximally unsound studies (e.g., produced by fraud) we have:

Theorem 6 (Backfire). If the unsoundness of studies has full support, then for any intervention that can arbitrarily decrease the type I error rate of studies[†] at a cost to their power, there will be critical point beyond which the mean false discovery of the published literature will increase.

[†]Assuming studies are non-degenerate in that unsoundness is non-extremal: $0 < k\alpha < 1$.

Backfire is the logical extreme of what's possible.
What do we actually expect?

Discussion

- Soundness-dependent selection has *always* been at play.
- This has implications for the natural selection of bad science (Cf. Smaldino 2016; Grimes et al 2017).
- *Publication bias* exacerbates this selection

Morals

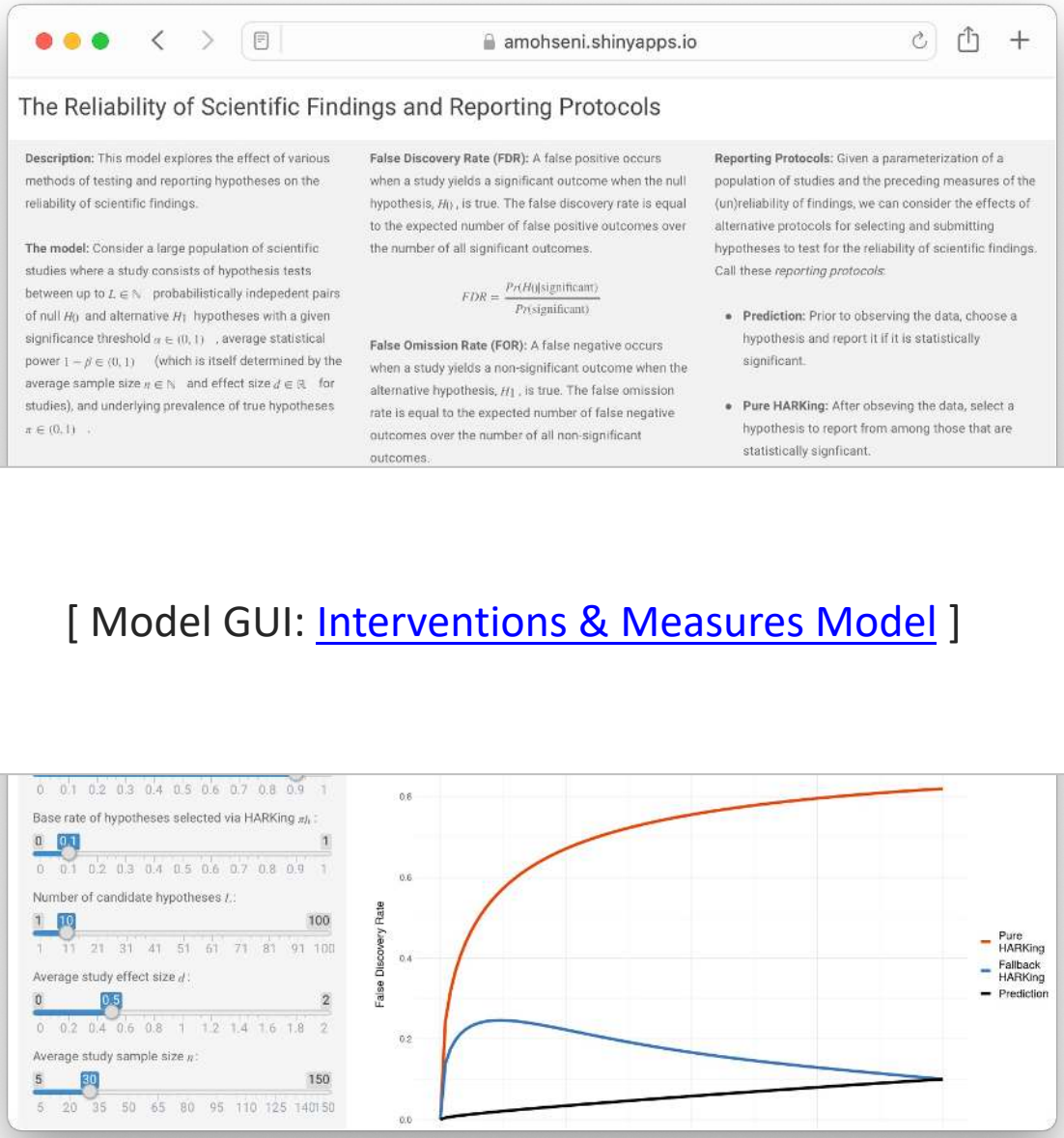
Interventions in evidentiary standards and research practices *tend not to affect all studies equally*.

Whenever the effects of interventions are *not entirely uncorrelated* with study soundness, we should be alert to their interaction.

Publication bias makes this all worse: Another argument for interventions like *registered reports*.

Measures that matter

- **False discovery rate** — this has dominated the conversation, but it is not the only, or necessarily the most important statistical property of published studies.
- **False omission rate**
- **Absolute discovery rate**
- **Magnitude exaggeration ratio**
- **Decision-relevant informativeness**



[Model GUI: Interventions & Measures Model]

An answer





ϕ	$1-\phi$

$\beta\phi$	$(1-\beta)\phi$	$\alpha(1-\phi)$	$(1-\alpha)(1-\phi)$

$(1-\beta)\phi$	$\alpha(1-\phi)$

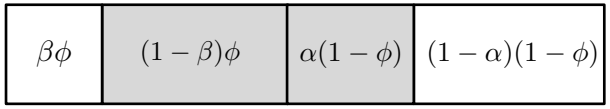
An answer







Better theory





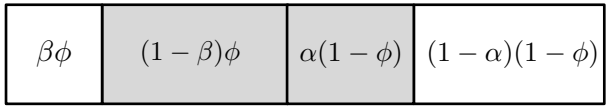
An answer







Better theory



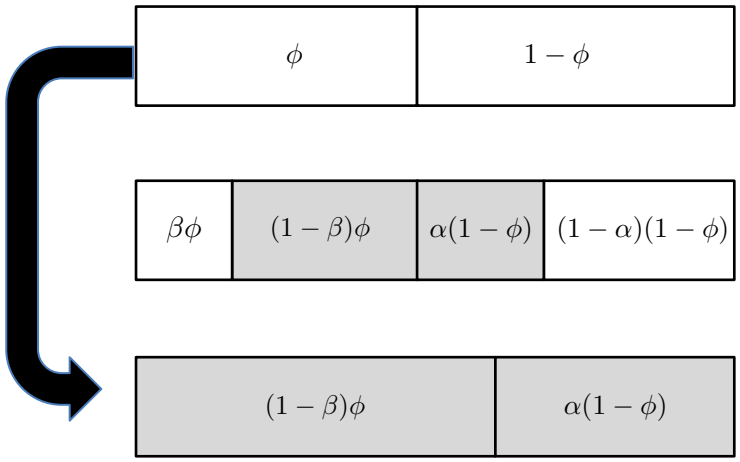
Better tests



An answer







Better theory

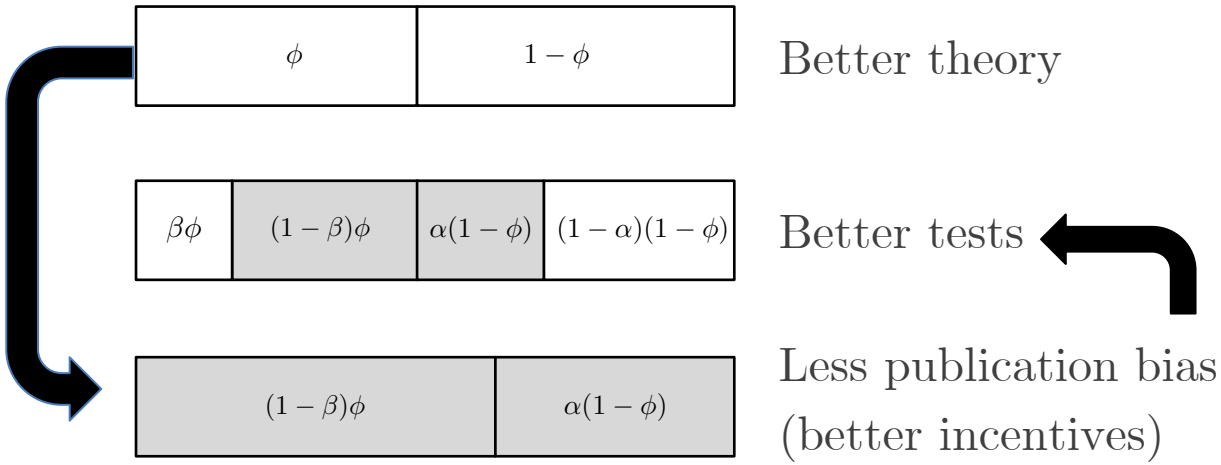
Better tests

Less publication bias
(better incentives)

An answer


Top scientist: We will find the 'God particle'
The "father" of an elusive subatomic particle said Monday he is almost sure it will be discovered in the next year in a race between powerful research equipment in the United States and Europe.

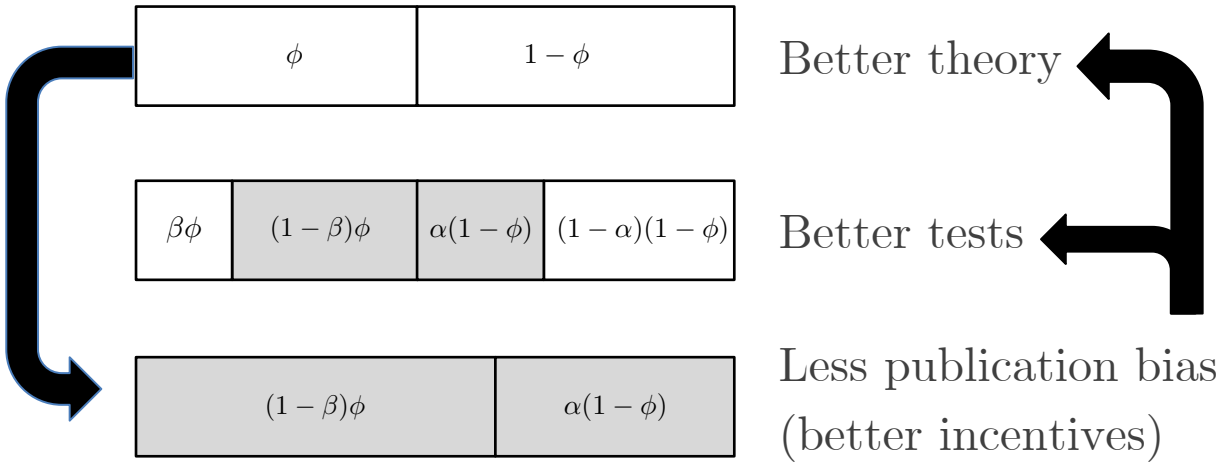

Cancer-busting broccoli sprout pills? It's a thing.
Translational scientist's 'green chemoprevention' research extends from eastern China to Seattle and now, into space


ϕ
$1-\phi$
Better theory
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
Better tests
$(1-\beta)\phi$
$\alpha(1-\phi)$
Less publication bias
(better incentives)

An answer


Top scientist: We will find the 'God particle'
The "father" of an elusive subatomic particle said Monday he is almost sure it will be discovered in the next year in a race between powerful research equipment in the United States and Europe.


Cancer-busting broccoli sprout pills? It's a thing.
Translational scientist's 'green chemoprevention' research extends from eastern China to Seattle and now, into space


ϕ
$1-\phi$
$\beta\phi$
$(1-\beta)\phi$
$\alpha(1-\phi)$
$(1-\alpha)(1-\phi)$
$(1-\beta)\phi$
$\alpha(1-\phi)$
Better theory
Better tests
Less publication bias
(better incentives)

Thank you.

Code and data for this project are available at:

https://osf.io/7h89w/

