HARKing: From Misdiagnosis to Misprescription

Aydin Mohseni

University of Pennsylvania

Department of Philosophy





My research

Primary research question:

How do we best understand and improve science?



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 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



How can we assess proposals for interventions in the norms of scientific practice?

How can we assess proposals for interventions in the norms of scientific practice?

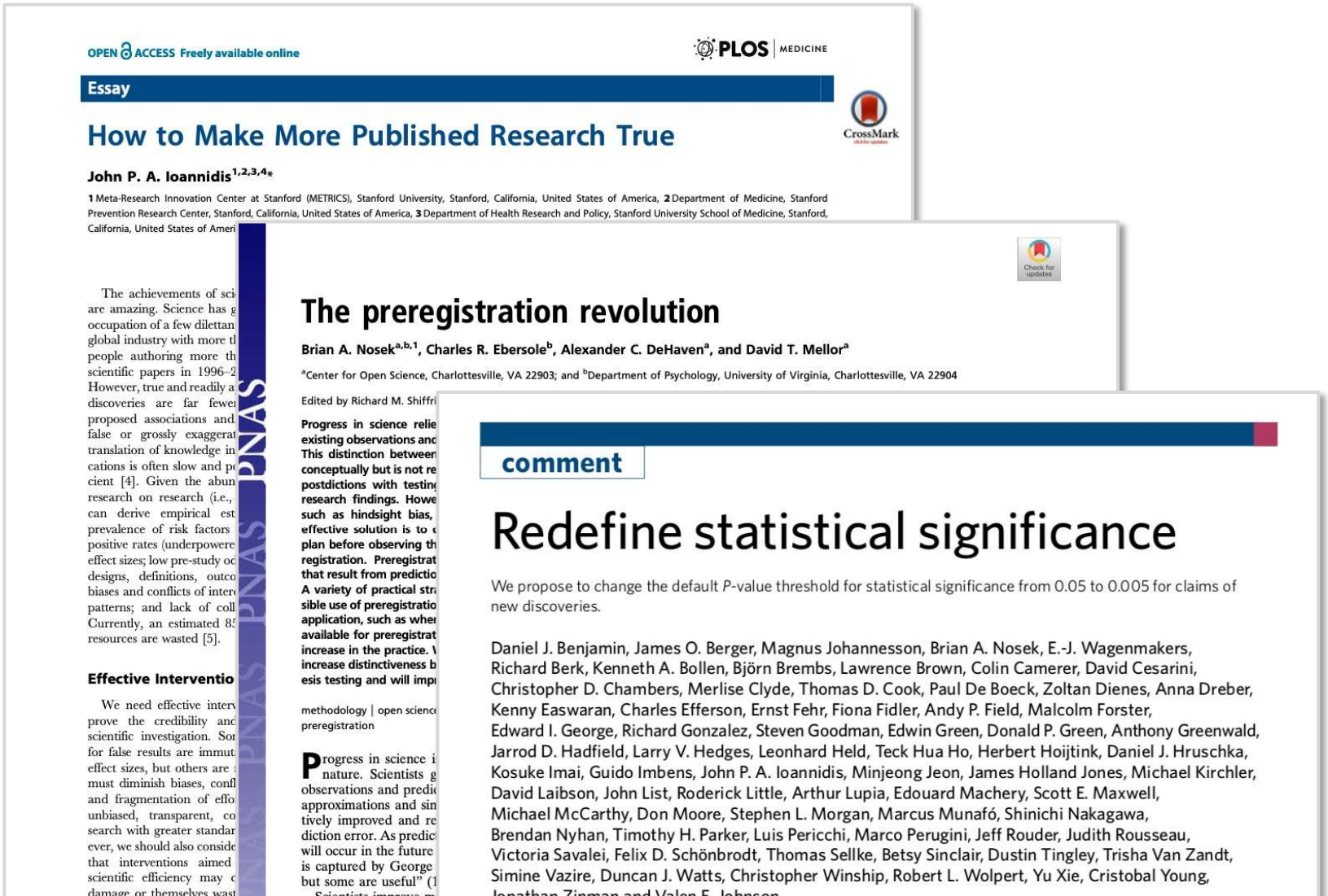
What is the difference between prediction and accommodation?

How can we assess proposals for interventions in the norms of scientific practice?

What is the difference between prediction and accommodation?

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

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,

Backdrop

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

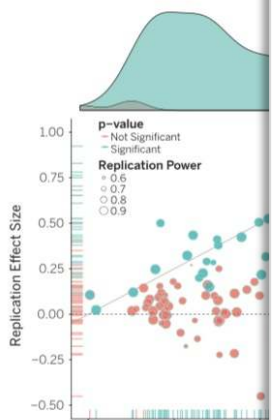
INTRODUCTION: Reproducibility is a defi
ing feature of science, but the extent to whi
it characterizes current research is unknow
Scientific claims should not gain creden
because of the status or authority of th
originator but by the replicability of th
supporting evidence. Even research of exe
plary quality may have irreproducible emp
ical findings because of random or systema
error.

RATIONALE: There is concern about the r
and predictors of reproducibility, but limit
evidence. Potentially problematic practices i
clude selective reporting, selective analysis, a
insufficient specification of the conditions n
essary or sufficient to obtain the results. Dir
replication is the attempt to recreate the cc
ditions believed sufficient for obtaining a p

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 as-

ON OUR WEB SITE

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



nature human behaviour

LETTERS

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

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer[1,16], Anna Dreber[2,16],
Magnus Johannesson[2,16], Michael
Thomas Pfeiffer[9,16], Adam Altmejd
Anup Gampa[7,8], Emma Heikensten[2],
Julia Rose[3], Eric-Jan Wagenmakers[14]

**Being able to replicate scientific findings i
entific progress[1–15]. We replicate 21 system
experimental studies in the social sciences pu
and *Science* between 2010 and 2015[16–36].
follow analysis plans reviewed by the origi
pre-registered prior to the replications. The
high powered, with sample sizes on average
higher than in the original studies. We find a
in the same direction as the original study fo
ies, and the effect size of the replications is
50% of the original effect size. Replicability v
(57%) and 14 (67%) studies for complemen
indicators. Consistent with these results, the
positive rate is 67% in a Bayesian analysis. T
size of true positives is estimated to be 71%,
both false positives and inflated effect sizes
contribute to imperfect reproducibility. Furth**

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

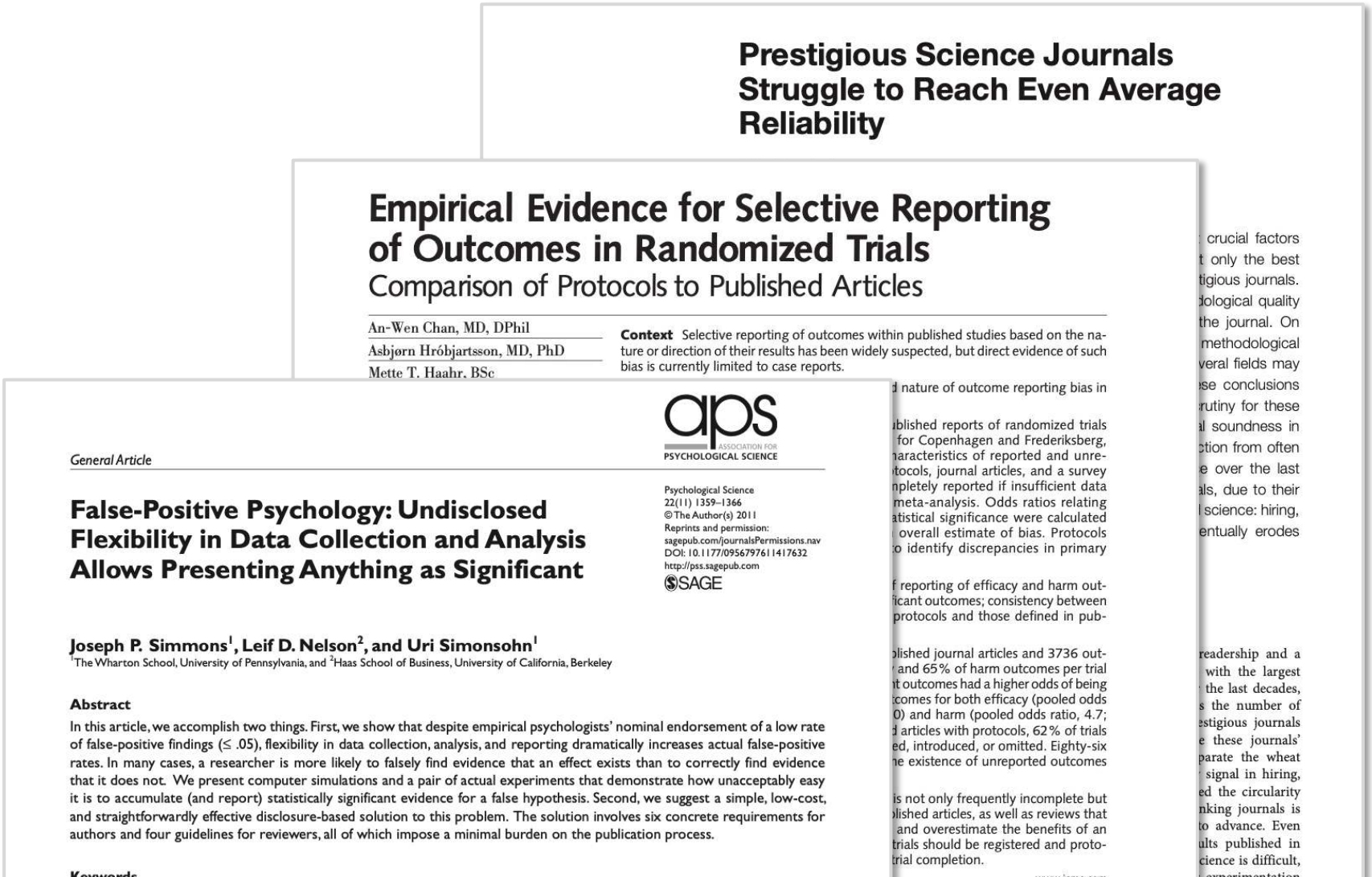
to 'feasible/marketable', and the financial costs of pursuing a full-blown drug discovery and development programme for a particular target could ultimately be hundreds of millions of Euros. Even in the earlier stages, investments in activities such as high-throughput screening programmes are substantial, 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 invest-

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

Causal factors



Prestigious Science Journals Struggle to Reach Even Average Reliability

crucial factors
t only the best
tigious journals.
dological quality
the journal. On
methodological
veral fields may
se conclusions
rutiny for these
l soundness in
ction from often
e over the last
als, due to their
l science: hiring,
entually erodes

readership and a
with the largest
the last decades,
s the number of
estigious journals
e these journals'
parate the wheat
signal in hiring,
ed the circularity
nking journals is
to advance. Even
ults published in
cience is difficult,

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.

d nature of outcome reporting bias in

blished reports of randomized trials
for Copenhagen and Frederiksberg,
haracteristics of reported and unre-
tocols, journal articles, and a survey
npletely reported if insufficient data
meta-analysis. Odds ratios relating
atistical significance were calculated
overall estimate of bias. Protocols
o identify discrepancies in primary

f reporting of efficacy and harm out-
ficant outcomes; consistency between
protocols and those defined in pub-

blished journal articles and 3736 out-
and 65% of harm outcomes per trial
t outcomes had a higher odds of being
comes for both efficacy (pooled odds
0) and harm (pooled odds ratio, 4.7;
d articles with protocols, 62% of trials
ed, introduced, or omitted. Eighty-six
he existence of unreported outcomes

is not only frequently incomplete but
blished articles, as well as reviews that
and overestimate the benefits of an
trials should be registered and proto-
trial completion.

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

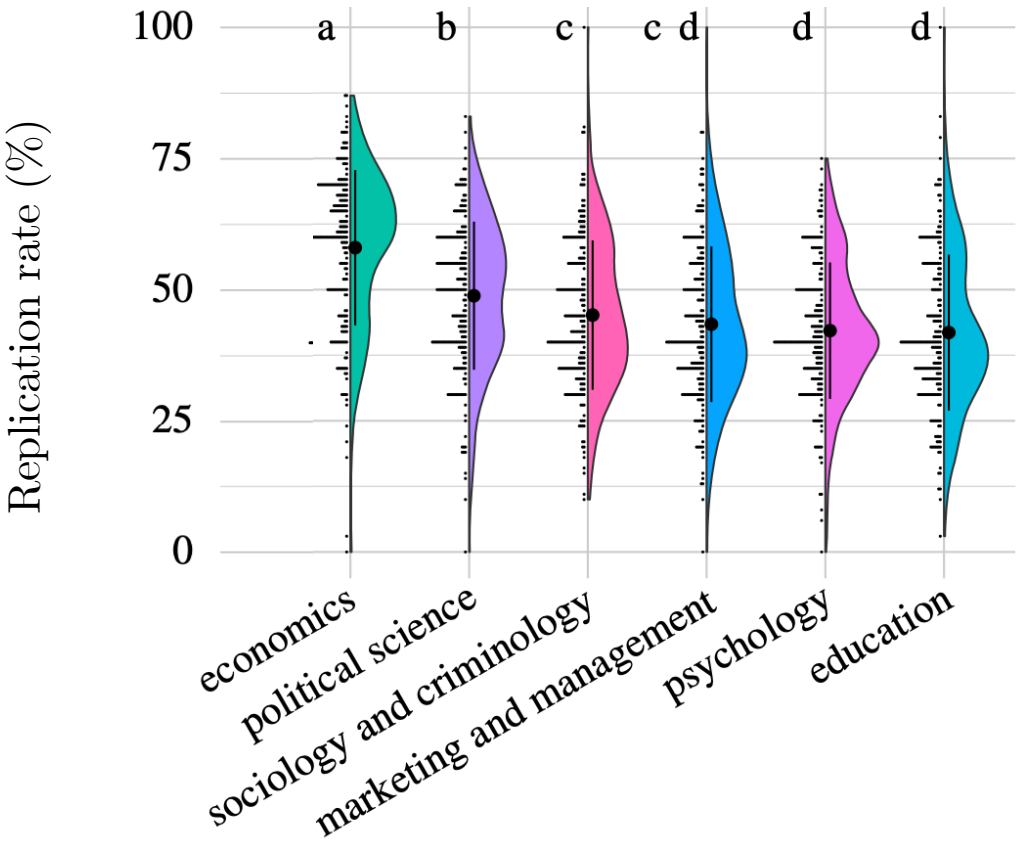
False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

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

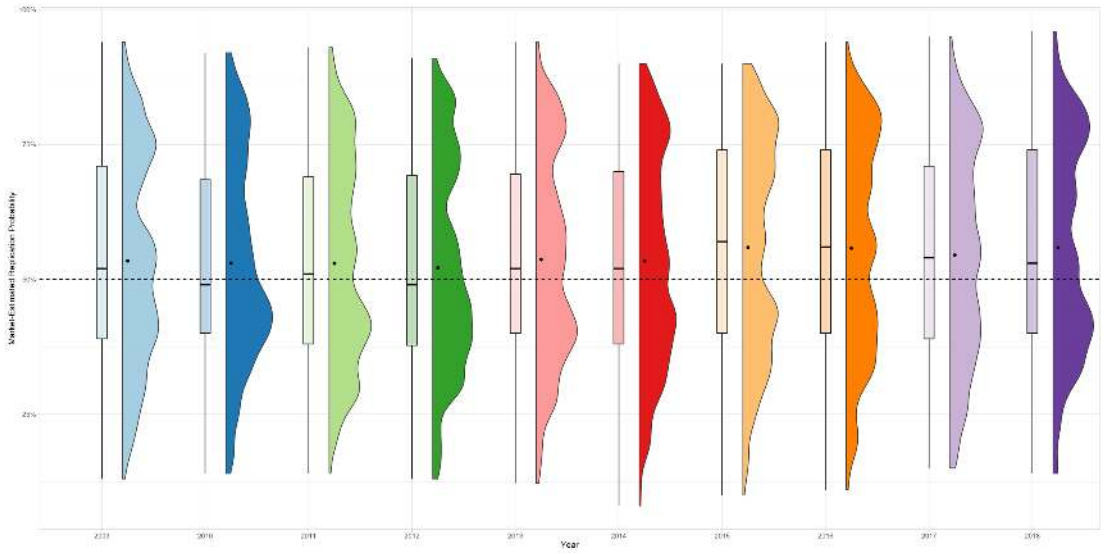
In this article, we accomplish two things. First, we show that despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤ .05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher 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.

Variation in replication rates across the sciences



—Gordon et al, 2020, "Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE program"

Very modest improvement in replicability



53.4% in 2009
to
55.8% in 2018

—Gordon et al, 2020, "Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE program"

Aims of this talk

Meta level. Present a framework for modeling a range of interventions in the replication crisis.

Object level. To demonstrate the potential value of the framework by using it to elucidate the dynamics of a specific research practice: HARKing.

The filter model

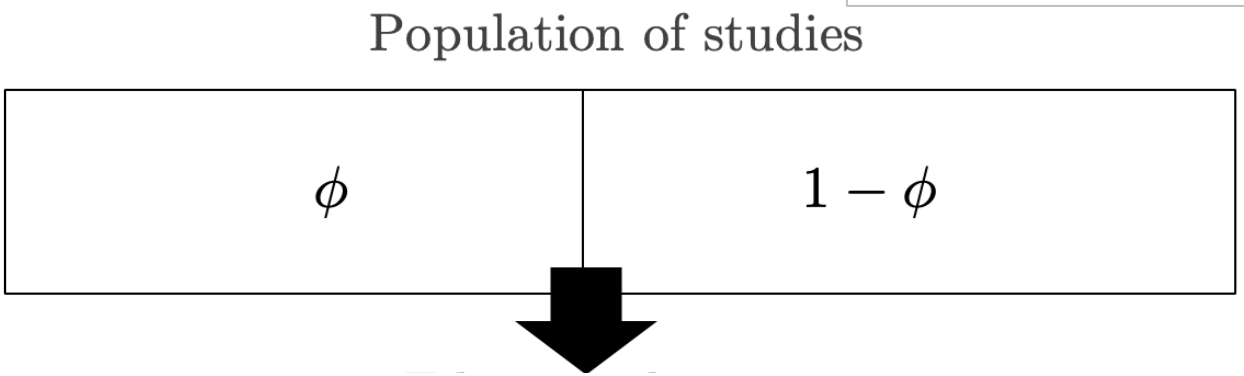
A shareable piece of conceptual machinery

The filter model

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

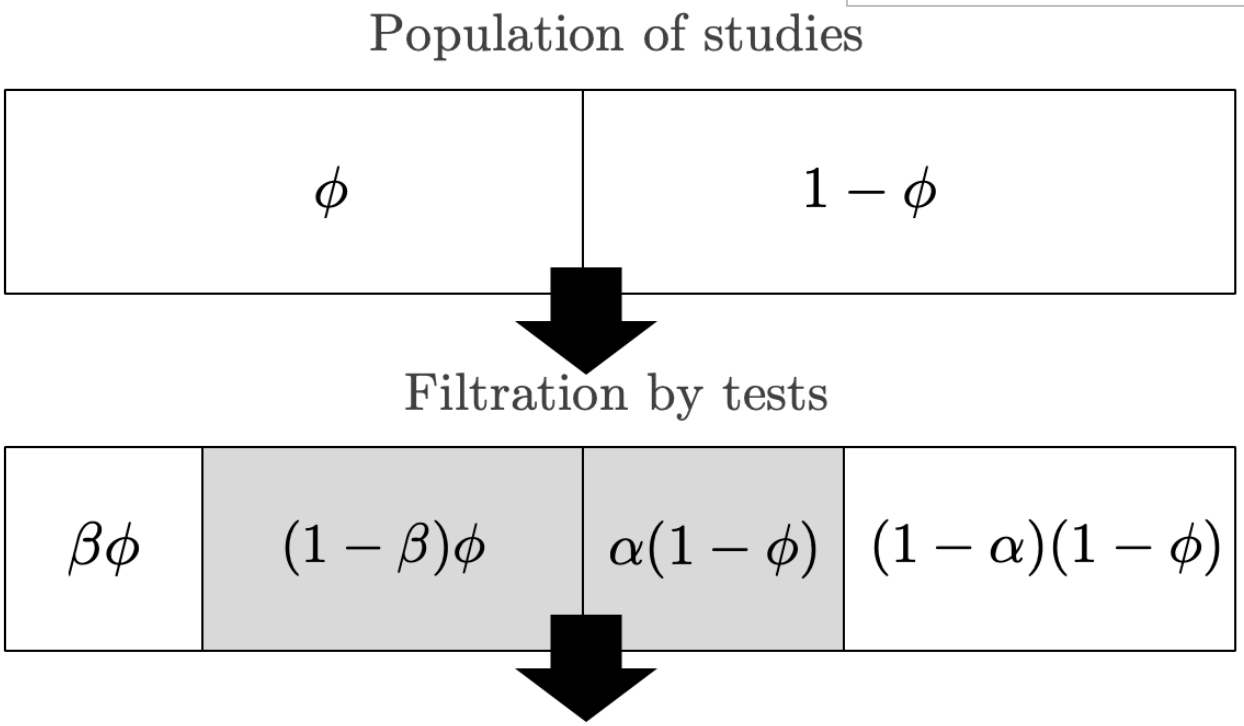
The filter model

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



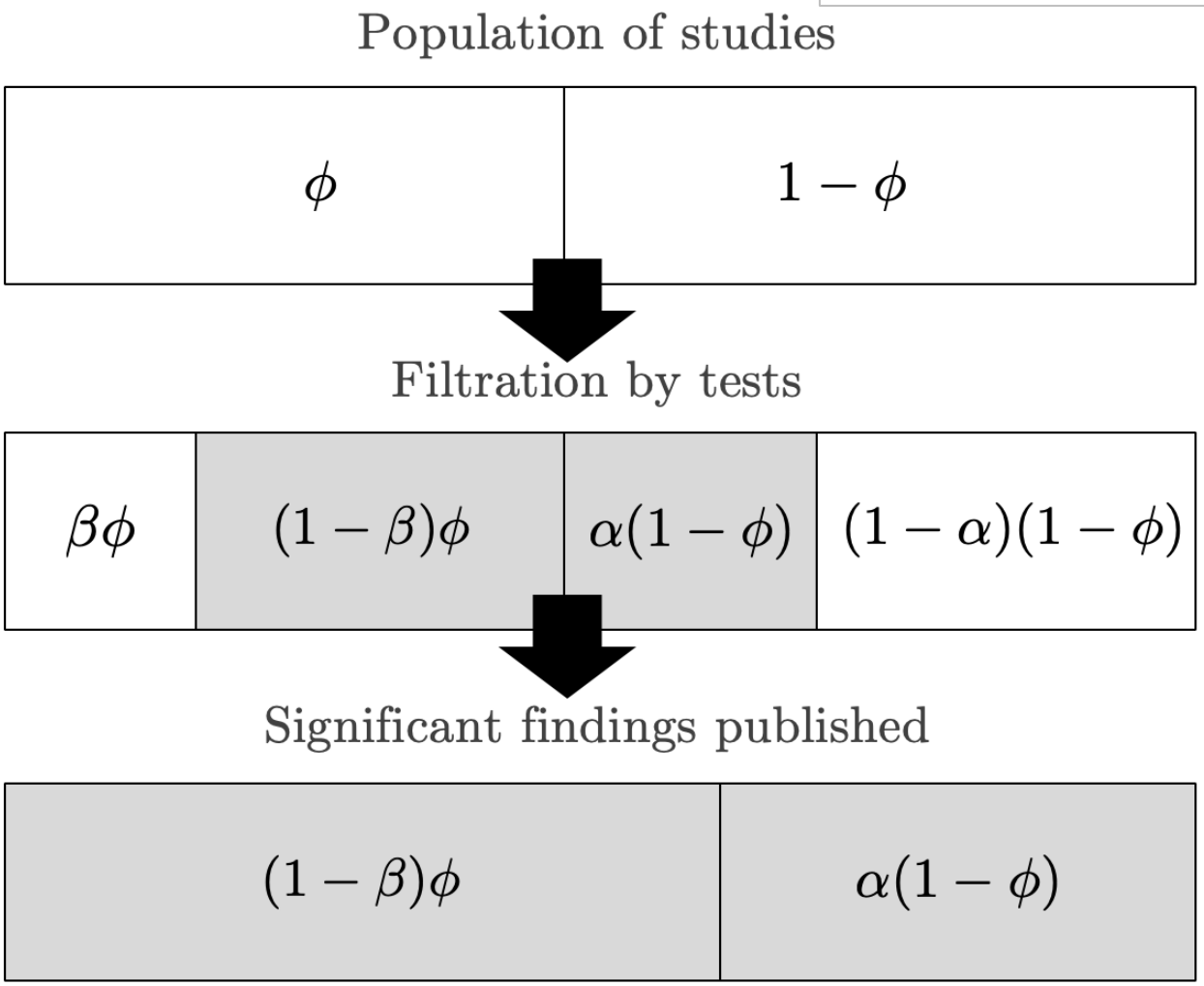
The filter model

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



The filter model

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



Significant findings published



False discovery rate of published findings

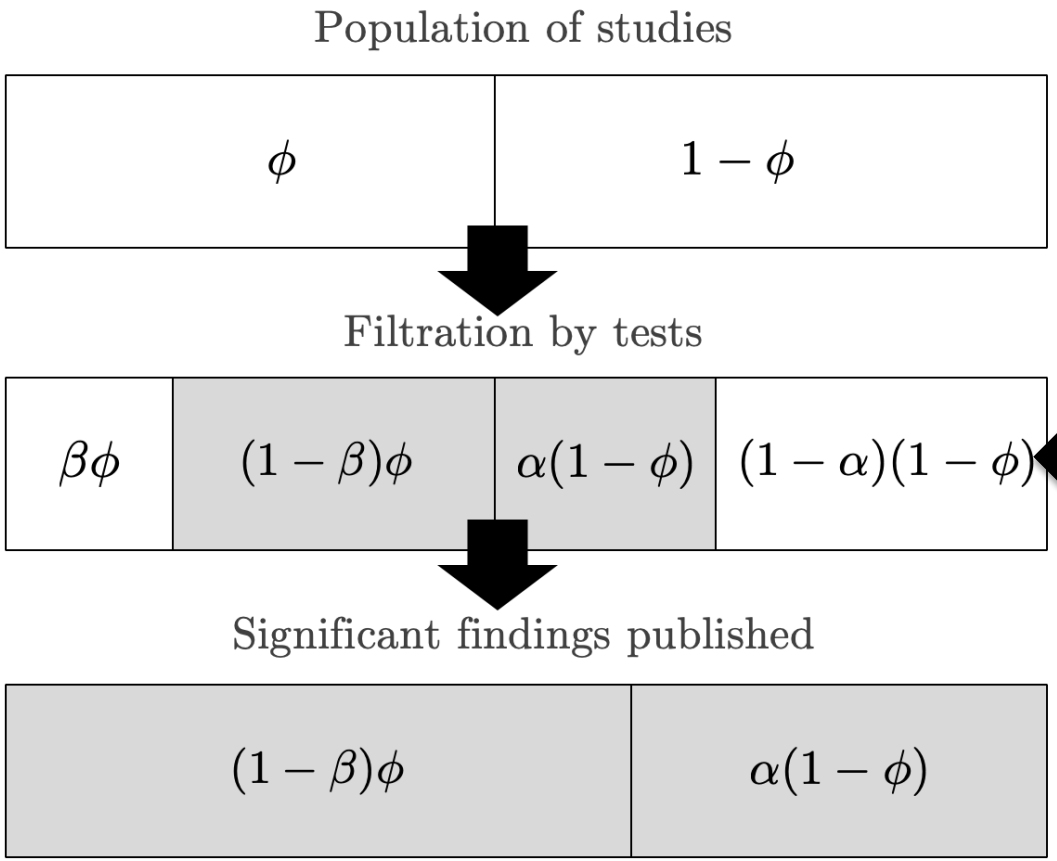
$$\frac{\alpha(1-\phi)}{(1-\beta)\phi+\alpha(1-\phi)}$$

$\alpha(1-\phi)$ ← significant & false

$(1-\beta)\phi+\alpha(1-\phi)$ ← significant

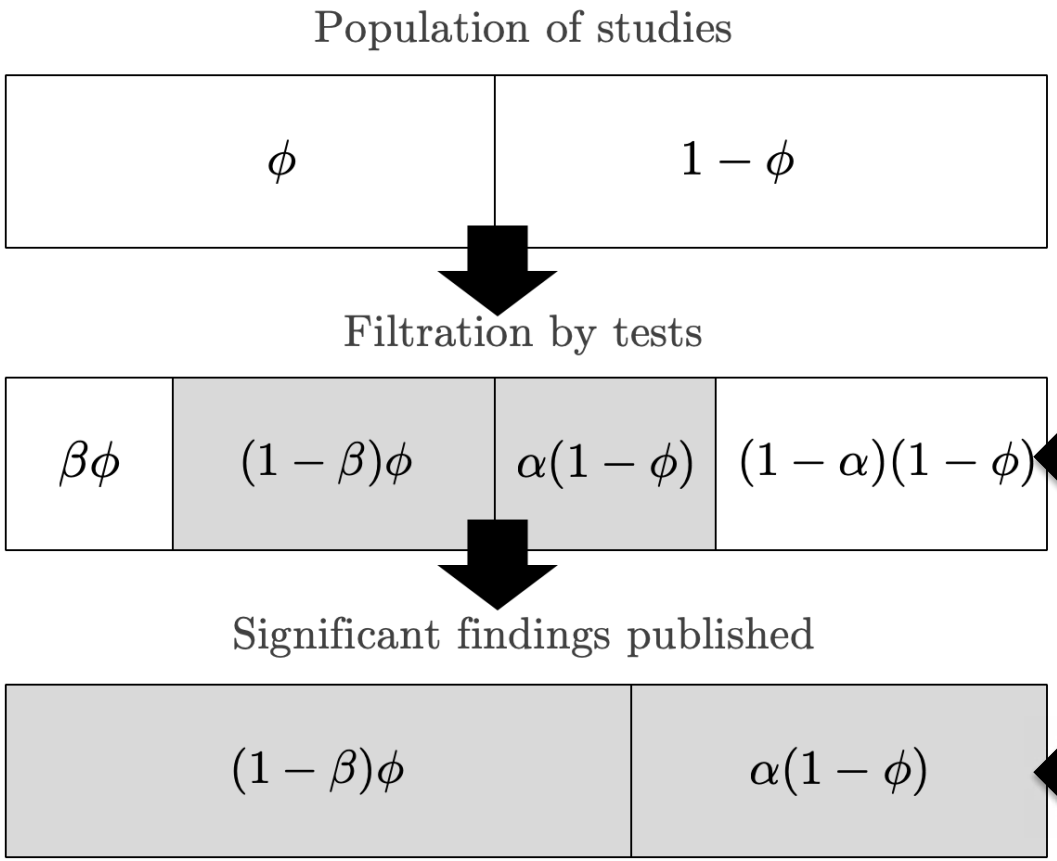
(The *rate of failed replications* is an estimate of this quantity.)

Interventions



- Preregister
- Lower significance threshold
- Increase study power

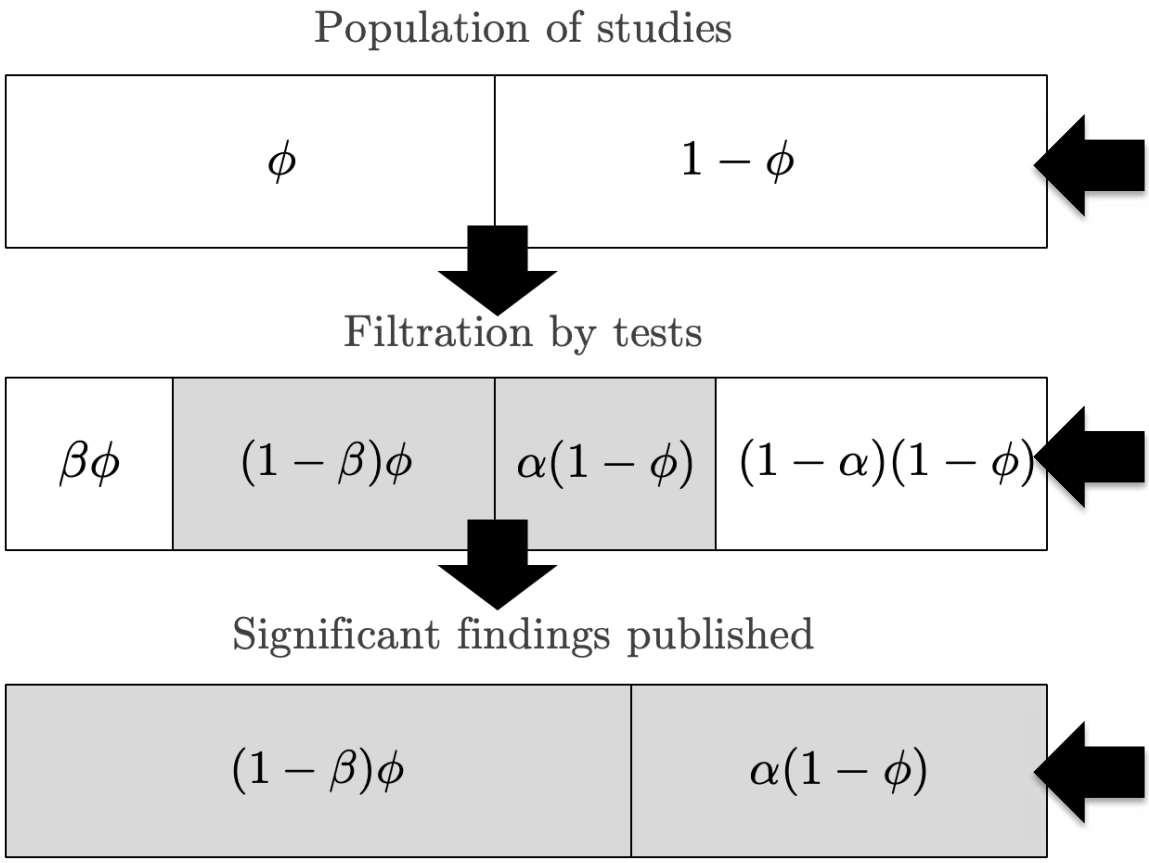
Interventions



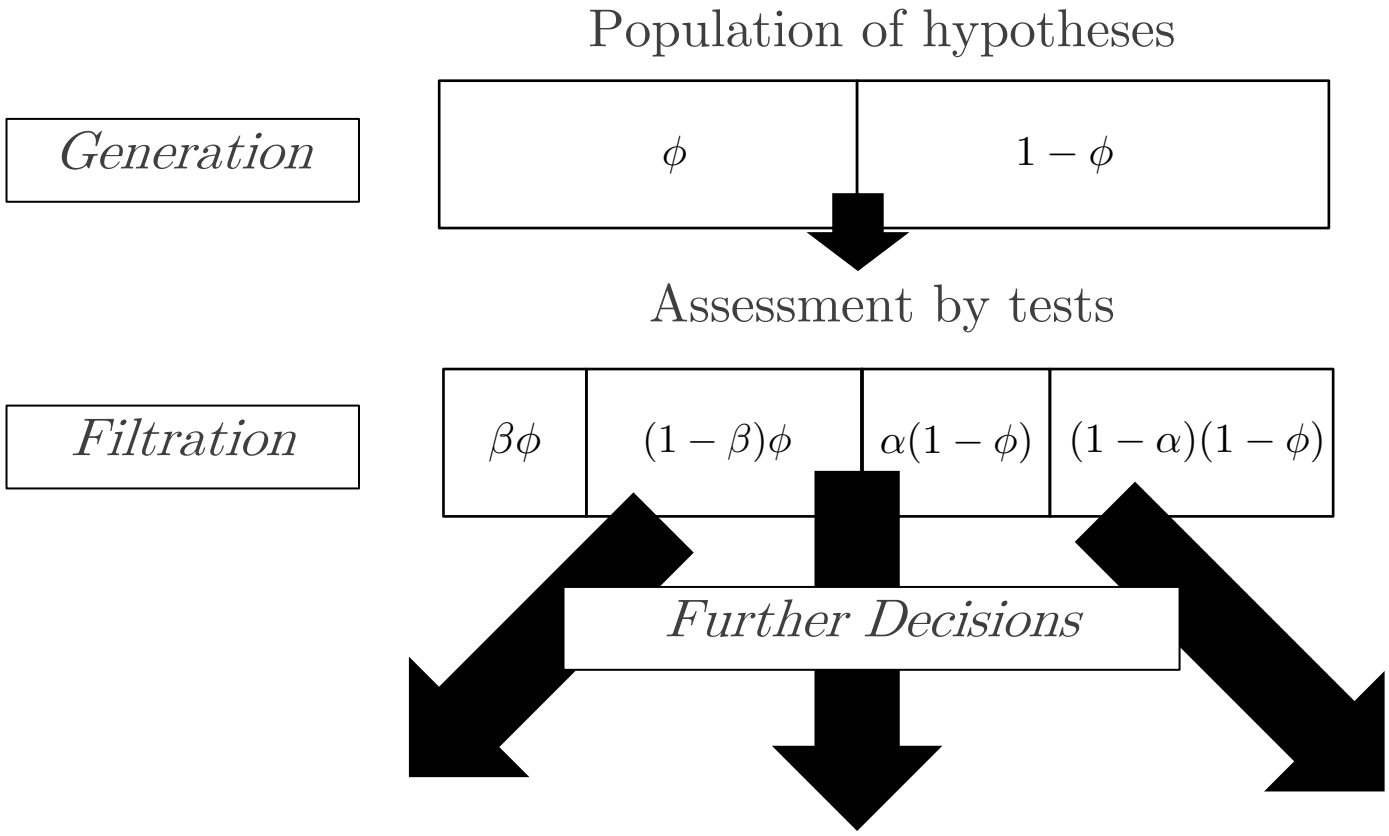
- Preregister
- Lower significance threshold
- Increase study power

- Publish null results

Interventions



- Improve theory
- Preregister
- Lower significance threshold
- Increase study power
- Publish null results

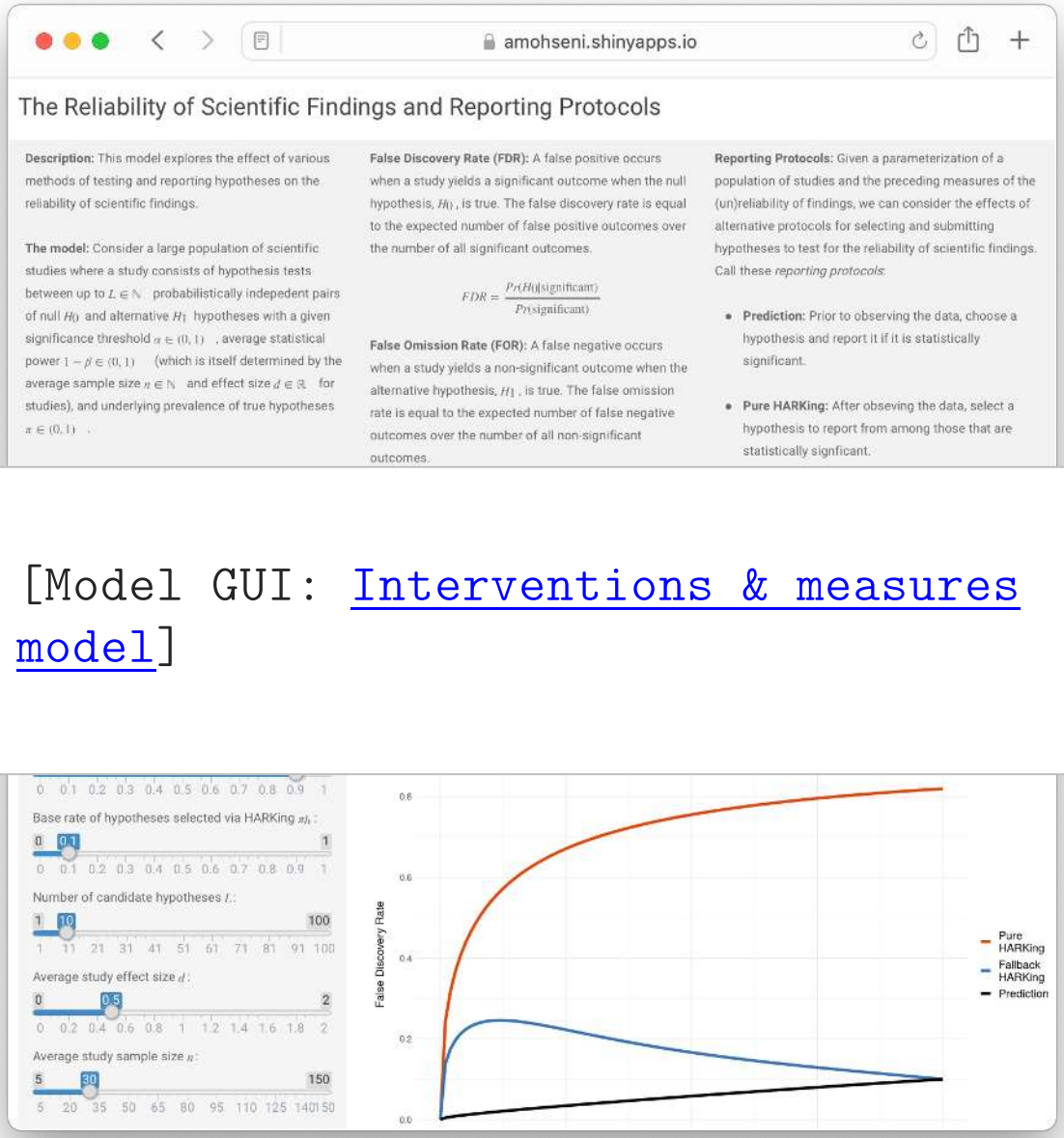


Further test
(e.g., confirmatory studies)

Publication

- False discovery rate
- False omission rate
- Absolute discovery rate
- Exaggeration ratio
- Informativeness

Canonization
(e.g., in textbooks)



[Model GUI: Interventions & measures model]

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

HARKing

HARKing 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

Contrast this with the picture of science in which a scientist makes a prediction and then gathers data to test the prediction.

Why does HARKing undermine findings?

Account 1. (Kerr, 1998) HARKing undermines frequentist guarantees of reliability of statistical tests.

Why does HARKing undermine findings?

Account 1. (Kerr, 1998) HARKing undermines frequentist guarantees of reliability of statistical tests.

Account 2. (Mayo, 1996, 2018) HARKing violates the requirements of severe testing.

Why does HARKing undermine findings?

Account 1. (Kerr, 1998) HARKing undermines frequentist guarantees of reliability of statistical tests.

Account 2. (Mayo, 1996, 2018) HARKing violates the requirements of severe testing.

Account 3. (Rubin, 2017) HARking misrepresents hypothesis formulated ex post as formulated ex ante.

Claim 1. None of these accounts is right.

Claim 2. The right account is a Bayesian one.

Claim 3. Misdiagnosis of HARKing can lead to misprescription in the context of the replication crisis.

Objections to HARKing I'm NOT addressing

- Intentional deception.
- Disclosing less than complete information.
- Failing to communicate an accurate model of science.
- And so on...

That is, accounts that don't have to do with the reliability of scientific findings.

Rather, I'm interested in the following question:

How, precisely, does HARKing interact with the reliability of scientific findings?

Claim 1. None of the standard accounts is right.

Structure of the argument

Structure of the argument

P1 Each account i claims that HARKing undermines the reliability of scientific findings because it exhibits a particular property φ_i.

Structure of the argument

P1 Each account i claims that HARKing undermines the reliability of scientific findings because it exhibits a particular property φ_i.

P2 HARKing can increase the reliability of findings while still satisfying property φ_i.

Structure of the argument

P1 Each account i claims that HARKing undermines the reliability of scientific findings because it exhibits a particular property φ_i.

P2 HARKing can increase the reliability of findings while still satisfying property φ_i.

$\therefore$ Hence, φ_i cannot explain why HARKing undermines the reliability of scientific findings.

HARK! Who Goes There? Three accounts of HARKing

Account 1. HARKing undermines frequentist guarantees of reliability of statistical tests.

Account 1. HARKing undermines frequentist guarantees of reliability of statistical tests.

The frequentist locates reliability in terms of guarantees of error-control in hypothetical, infinite repetitions of her tests.

Account 1. HARKing undermines frequentist guarantees of reliability of statistical tests.

Example. Let $P(\text{significant} \mid H_0) = \alpha = .05$.

Assume, under the HARKing protocol, that there are 20 independent candidate hypotheses, *all false*.

For such a HARKed test, we get $P(\text{significant} \mid H_0) = 1 - .95^{20} = 0.64$.

Account 1. HARKing undermines frequentist guarantees of reliability of statistical tests.

"*P*-values are meaningless when taken out of context of all the analyses performed to get them."

—Bishop, (2017, p. 435)

Account 2. HARKing violates a the requirements of severe testing.

"HARKing is problematic for scientific progress because it results in hypotheses that are always confirmed and never falsified by the results."

—Kerr, (1998, p. 2)

Account 2. HARKing violates a the requirements of severe testing.

Under the severe testing account, a statistical test provides us with warrant for/corroboration of a hypothesis in so far as it submits it to a severe test, where a severe test is one such that "it would detect an error in the study hypothesis if one were present" (Mayo, 2016, p. 13).

Account 2. HARKing violates a the requirements of severe testing.

In so far as HARKing involves failing to report non-significant results for a given hypotheses, it fails to provide a severe test of that hypothesis.

Account 3. HARKing misrepresents hypothesis formulated ex post as formulated ex ante.

"...it is deceptive for researchers to using current results to construct post hoc hypotheses that are then reported as if they were a priori hypotheses and to misrepresent accommodation as prediction."

—Rubin, 2017, p. 11

Account 3. HARKing misrepresents hypothesis formulated ex post as formulated ex ante.

This is typically framed as a sort of ethical objection.

We will see that this account is most nearly right; not on ethical grounds, but rather on *epistemological* ones.

The Reliability of Scientific Findings

Any argument that a particular method (HARKing) undermines the *reliability of scientific findings* needs to define that notion.

Classically, we want to have our epistemic methods produce fewer false beliefs and more true ones.

In the context of null hypothesis significance testing, we can ask that fewer of the results we declare significant be false, and fewer of the results we declare non-significant actually be true.

These correspond to the requirement that our methods exhibit lower rates of *false discovery* and *false omission* respectively.

False discovery rate

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

The *false discovery rate* is equal to the expected number of false positive outcomes over the number of all significant outcomes.

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

False omission rate

A *false negative* occurs when a study yields a non-significant outcome when the null hypothesis is false.

The *false omission rate* is equal to the expected number of false negative outcomes over the number of all non-significant outcomes.

$$Pr(H_1|\text{not significant}) = \frac{Pr(H_1, \text{not significant})}{Pr(\text{not significant})}$$

Definition (Reliability). Let a population of studies be specified by the fixed false positive rate $\alpha \in (0,1)$ and mean false negative rate $\beta \in (0,1)$ of its tests, along with the fraction of true hypothesis $\pi \in (0,1)$ from which candidate hypothesis are selected for testing.

Definition (Reliability). Let a population of studies be specified by the fixed false positive rate $\alpha \in (0,1)$ and mean false negative rate $\beta \in (0,1)$ of its tests, along with the fraction of true hypothesis $\pi \in (0,1)$ from which candidate hypothesis are selected for testing.

We say that a given a protocol for reporting findings P is *more reliable* than another P' if, for any such population of studies, findings produced employing P produce a lower false discovery rate and false omission rate than those produced using P'.

What of classical error-control guarantees? (Type I and II errors)

Consider a domain of scientific study in which no true hypothesis are available—that is, where the null hypothesis is always true.

Regardless of the Type I and II error rates of tests, all significant findings will be false discoveries.

What of classical error-control guarantees? (Type I and II errors)

More generally, one and the same frequentist guarantee—Type I and II error rates—is compatible with any reliability of scientific findings since reliability is critically a function of the prevalence of true hypotheses put to test.

What of data-dependent measures? (p-values and test severity)

The same shortcoming applies to such measures.

Insofar as a measure ignores the prevalence of true hypotheses submitted to testing, it will assign the same measure of reliability to a set of findings assured to be false as to a set of findings assured to be true.

When HARKing Cannot Be Bad

Let researchers select their hypotheses from a single set at random.

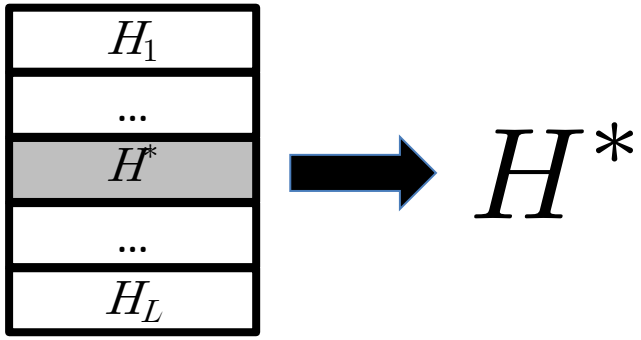
This is equivalent to setting the prior odds of hypotheses being true to be equal.

Suppose a researcher observes some data and obtains the p-values for L independent and ex ante identical hypothesis tests. Consider the reporting protocols:

Suppose a researcher observes some data and obtains the p-values for L independent and ex ante identical hypothesis tests. Consider the reporting protocols:

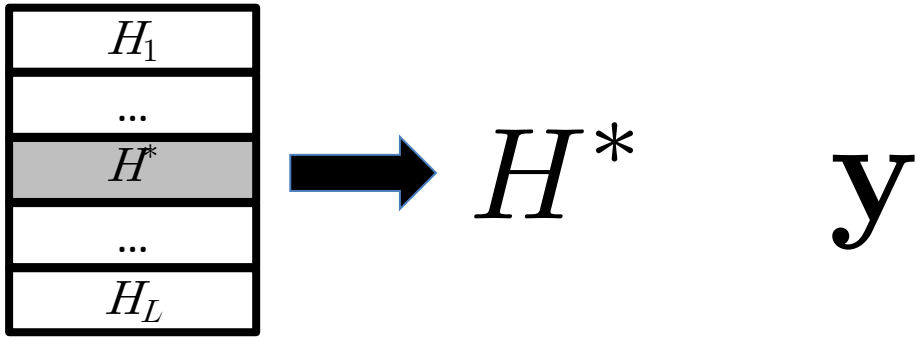
1. *Prediction,* $\boldsymbol{P_p}$: **Test single hypothesis.**

Prediction



1. Select a hypothesis H^* ex ante to observing the data.

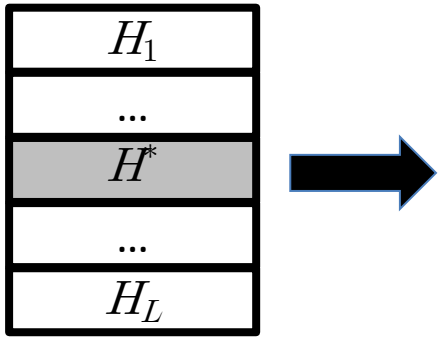
Prediction



1. Select a hypothesis H^* ex ante to observing the data.

2. Observe the data

Prediction



H^* $\mathbf{y}$ Report H^* if $p_{H^*} < \alpha$

1. Select a hypothesis H^* ex ante to observing the data.

2. Observe the data

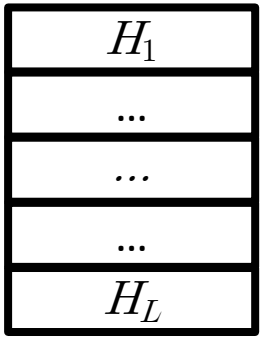
3. Report H^* if the data show that it is significant.

Suppose a researcher observes some data and obtains the p-values for L independent and ex ante identical hypothesis tests. Consider the reporting protocols:

1. *Prediction,* P_p: Test single hypothesis.

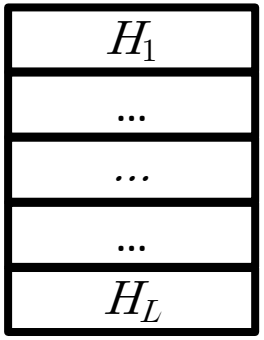
2. ***HARKing,*** $\boldsymbol{P_h}$**: Randomly select from significant hypotheses.**

HARKing



1. Don't commit to any hypothesis ex ante to observing the data.

HARKing

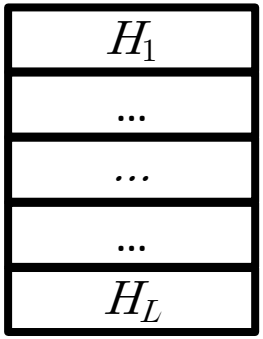


2. Observe the data

y

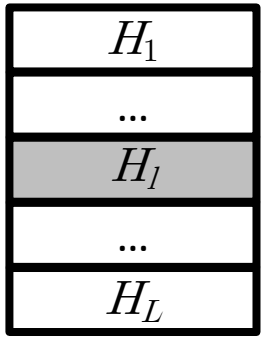
1. Don't commit to any hypothesis ex ante to observing the data.

HARKing



2. Observe the data

y



H_l

s.t. $p_{H_l} < \alpha$

1. Don't commit to any hypothesis ex ante to observing the data.

3. Report an H_l, if one exists, such that the data show that it is significant.

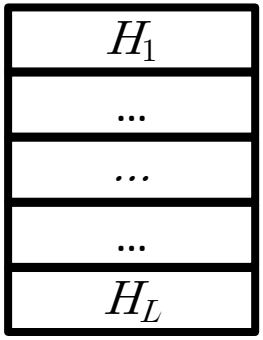
Suppose a researcher observes some data and obtains the p-values for L independent and ex ante identical hypothesis tests. Consider the reporting protocols:

1. *Prediction,* P_p: Test single hypothesis.

2. *HARKing,* P_h: Randomly select from significant hypotheses.

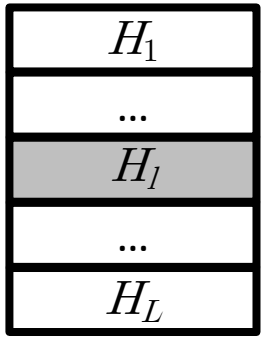
3. ***Selective HARKing,*** $\boldsymbol{P_{sh}}$**: Report hypothesis with the lowest** p**-value.**

Selective HARKing



2. Observe the data

y



H_l

$$\text{s.t. } \min\{p_{H_l}\}_{l=1}^{L} < \alpha$$

1. Don't commit to any hypothesis ex ante to observing the data.

3. Report an H_l, if one exists, such that the data show that it is significant and it has the lowest p-value.

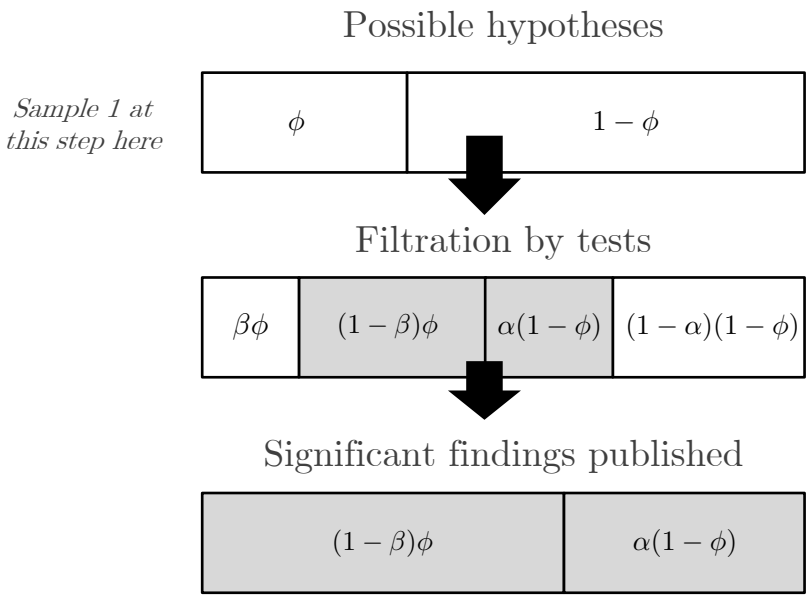
Proposition 1. When the base rates of true hypotheses selected vis prediction and HARKing are equal,

1. Prediction yields the same false discovery rate as HARKing, both exceeding the *false discovery rate* of selective HARKing

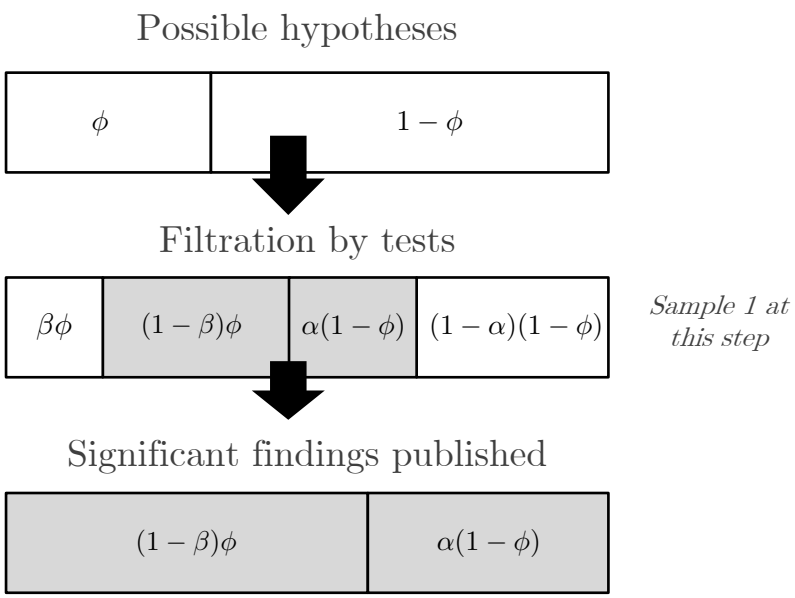
$$FDR(P_p) = FDR(P_h) \leq FDR(P_{sh})$$

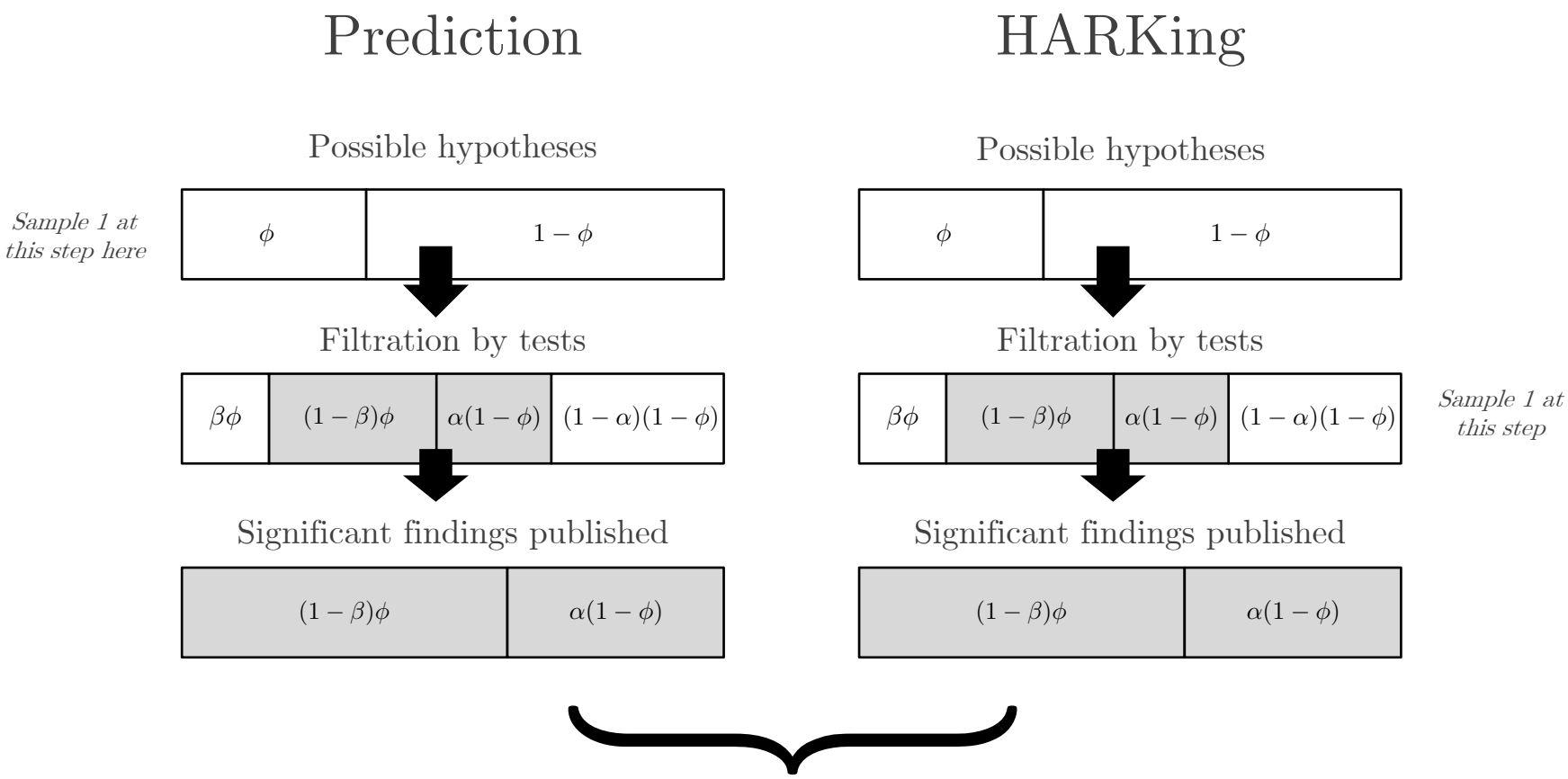
2. The *false discovery rate* for selective HARKing is decreasing in the number of hypotheses.

Prediction



HARKing





The fraction of true and false hypotheses must be the same.

Proposition 2. When the base rates of true hypotheses selected vis prediction and HARKing are equal, Prediction yields the same *false omission rate* as HARKing

$$FOR(P_p) = FOR(P_h).$$

Notice that HARKing satisfies each:

$\boldsymbol{\varphi_1}$: Undermines frequentist guarantees of error control,

Notice that HARKing satisfies each:

$\boldsymbol{\varphi_1}$: Undermines frequentist guarantees of error control,

$\boldsymbol{\varphi_2}$: Violates requirement of providing a 'severe test',

Notice that HARKing satisfies each:

$\boldsymbol{\varphi_1}$: Undermines frequentist guarantees of error control,

$\boldsymbol{\varphi_2}$: Violates requirement of providing a 'severe test',

$\boldsymbol{\varphi_3}$: Misrepresents when hypothesis was formulated,

Notice that HARKing satisfies each:

$\boldsymbol{\varphi_1}$: Undermines frequentist guarantees of error control,

$\boldsymbol{\varphi_2}$: Violates requirement of providing a 'severe test',

$\boldsymbol{\varphi_3}$: Misrepresents when hypothesis was formulated,

while being at least as reliable as prediction.

Notice that HARKing satisfies each:

$\boldsymbol{\varphi_1}$: Undermines frequentist guarantees of error control,

$\boldsymbol{\varphi_2}$: Violates requirement of providing a 'severe test',

$\boldsymbol{\varphi_3}$: Misrepresents when hypothesis was formulated,

while being at least as reliable as prediction.

Hence none of $\boldsymbol{\varphi_1}$, $\boldsymbol{\varphi_2}$, and $\boldsymbol{\varphi_3}$ explain why HARKing undermines the reliability of scientific findings.

When HARKing Must Be Bad

When is HARKing bad for the reliability of scientific findings?

Under the conditions we described—equal prior odds of hypotheses selected via prediction and HARKing—selecting hypotheses ex post to observing the data cannot undermine the reliability of one's findings.

Thus, for HARKing to be bad in the course of normal scientific practice, one of the assumptions of our model must not obtain.

The natural candidate is that scientists do not in fact choose their hypotheses at random.

The hypotheses the researcher chooses prior to observing the data may be informed by theory, previous findings in the literature, and common sense.

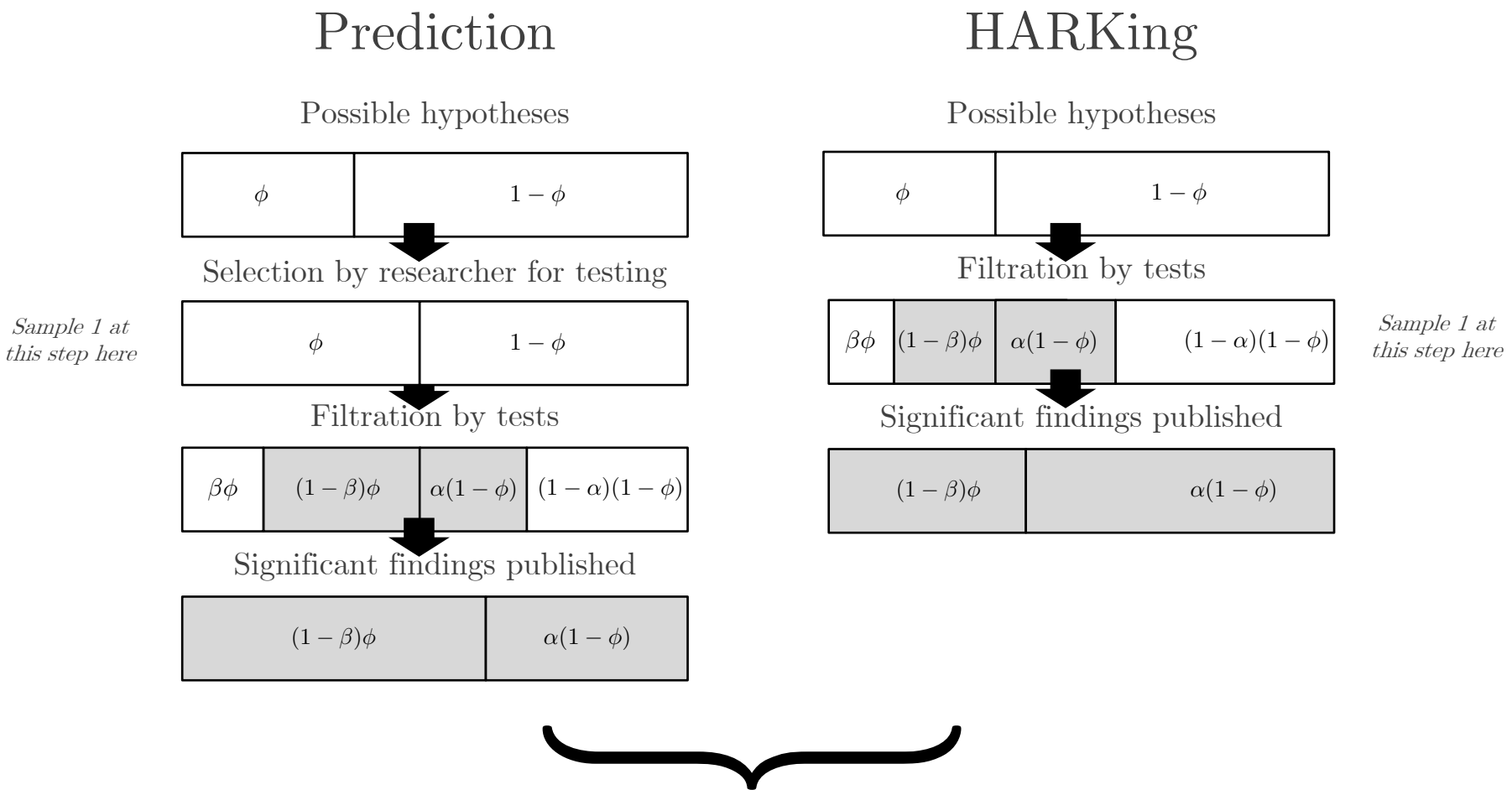
When a researcher is suitably informed, a hypothesis she selects is more likely to be true than a hypothesis selected at random from the set of hypotheses that are merely logically consistent with her data.

In such a case, we can expect the researcher to do better than chance prior to observing the data.

In contrast, when a researcher is engaged in HARKing, however, she is no longer filtering the set of possible hypotheses via her domain knowledge. Rather, the full set of hypotheses consistent with her data are submitted to the filter of statistical significance.

Formally, this corresponds to a model in which the researcher chooses from two sets of hypotheses with different fractions of true and false hypotheses.

When she chooses ex ante to observing her data, via prediction, she tests a subset of hypothesis with some prevalence of true hypotheses; and when she chooses ex post to observing her data, via HARKing, she chooses from a superset with a different prevalence of true hypotheses.



With the addition of research judgement in hypothesis selection, prediction exhibits a smaller fraction of false discoveries.

Proposition 3. Let hypotheses selected via prediction and HARKing exhibit base rates π_p and π of true hypotheses, respectively. Then prediction is more reliable than HARKing with respect to false discovery, $FDR(P_p) < FDR(P_h)$, just in case $\pi_p > \pi$.

This also provides an epistemic explication of the misrepresentation objection to HARKing.

If we expect the researcher to have substantial domain knowledge, we expect the hypotheses she selects ex ante to be more likely to be true.

If we are misled on this count, and the hypothesis reported does not boast such support, we will (intuitively) assign too great a credence to her findings.

When (not) to HARK?

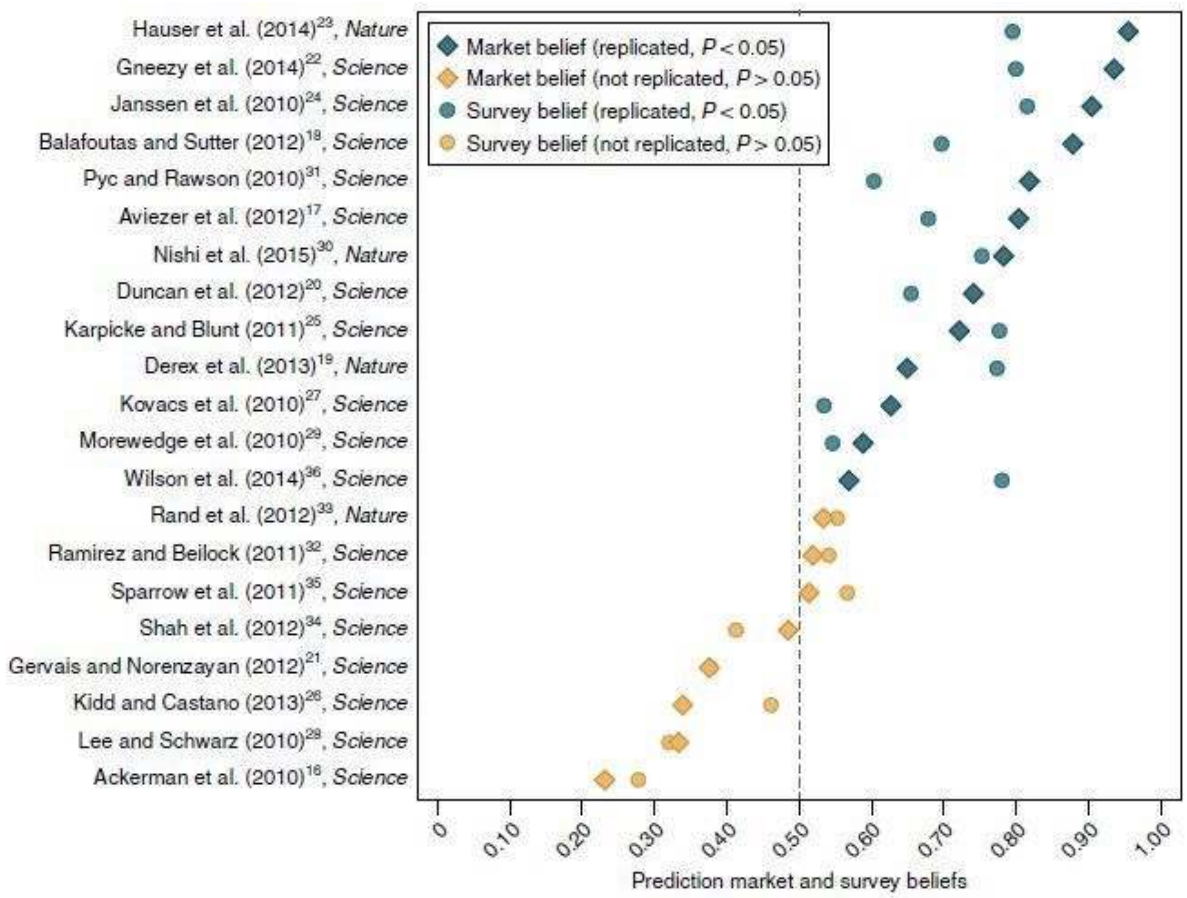
This also tells us when HARKing will produce to more reliable findings than prediction.

Consider an unlucky world in which the scientist's judgment is anti-correlated with the truth.

In this world, it is better to HARK than to predict.

However, even if the researcher's judgment is not much better than chance, it is important to notice that prediction is much less efficient at producing discoveries—which may be the priority in exploratory analyses.

We should limited researcher judgment when the set of relevant hypotheses is very large, highly complicated, or where researchers are known to be biased.



(Camerer et al, 2018)

Moral 1

When researcher judgment on a problem is well-informed, HARKing will tend to lower the reliability of scientific findings.

However, when researcher judgment is limited HARKing can actually outperform prediction.

Moral 2

Current accounts of HARKing that stem from a frequentist philosophy of statistics fail to explain the actual logic of its interaction with the reliability of scientific findings; and a Bayesian analysis of the problem makes this all clear.

Thank you. Looking forward to discussion.



All code and data for this project is available at: https://osf.io/7h89w/