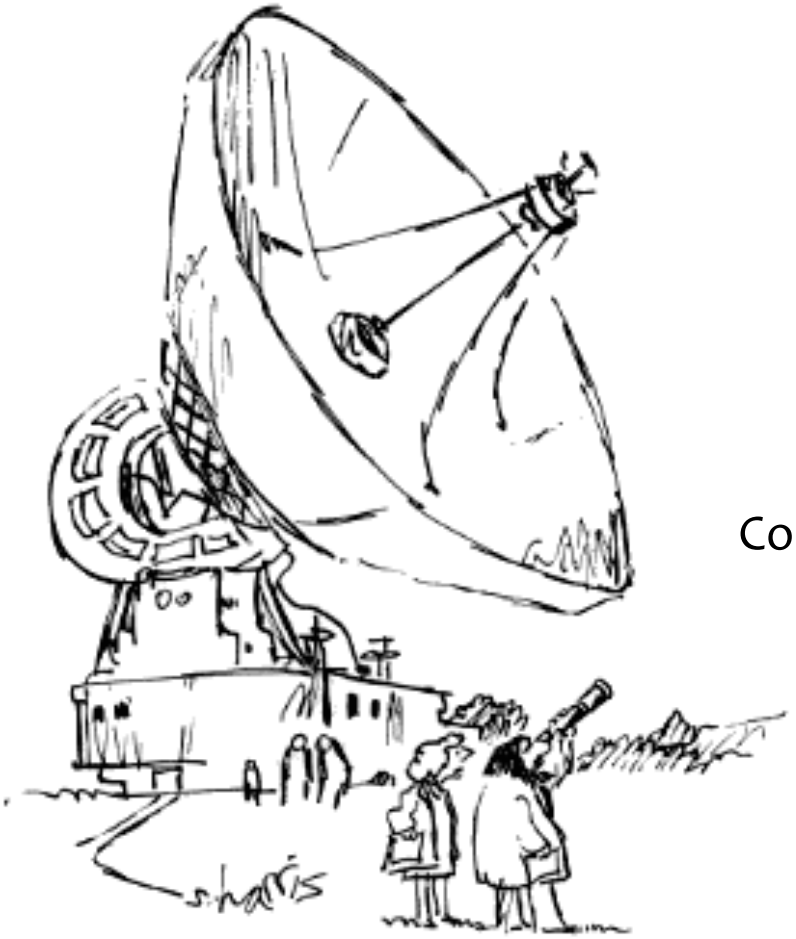


"Just checking."

Thoughts on the Replication Crisis

David Jensen

College of Information & Computer Sciences
Computational Social Science Institute
Center for Data Science

Summary

- Recent studies have reporting unexpectedly low rates of replication of important findings. This has been called *the replication crisis*.

Summary

- Recent studies have reporting unexpectedly low rates of replication of important findings. This has been called *the replication crisis*.
- However…
  - …low replication rates should not be *unexpected*…
  - …given the structure of the research enterprise.

Summary

- Recent studies have reporting unexpectedly low rates of replication of important findings. This has been called *the replication crisis*.
- However…
  - …low replication rates should not be *unexpected*…
  - …given the structure of the research enterprise.
- To improve things, we can either…
  - …improve the behavior of individuals…
  - …or change the structure of the system itself.

Low replication rates should not be unexpected

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summa

There is
current pu
false. The p
is true may
bias, the nu
same quest
of true to n
relationship
field. In this
is less likely
conducted i
effect sizes a
greater num
of tested rela
greater flexibi
outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may

"After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV."

and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined

nd can
ther the
tionships
few
usands
at may
sider,

her there
among
i) or
of the
ips. The
onship
probability

of a study finding a true relationship reflects the power 1 – β (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that *c* relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true

Table 4. PPV of Research Findings for Various Combinations of Power ($1 - \beta$), Ratio of True to Not-True Relationships *(R)*, and Bias *(u)*

$1 - \beta$	*R*	*u*	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study.

RCT, randomized controlled trial.

DOI: 10.1371/journal.pmed.0020124.t004

Table 4. PPV of Research Findings for Various Combinations of Power ($1 - \beta$), Ratio of True to Not-True Relationships

$1 - \beta$		PPV
0.80		0.85
0.95		0.85
0.80		0.41
0.20		0.23
0.20		0.17
0.80		0.20
0.20		0.12
0.20		0.010
0.20		0.015

Practical Example	PPV
Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
Confirmatory meta-analysis of good-quality RCTs	0.85
Meta-analysis of small inconclusive studies	0.41
Underpowered, but well-performed phase I/II RCT	0.23
Underpowered, poorly performed phase I/II RCT	0.17
Adequately powered exploratory epidemiological study	0.20
Underpowered exploratory epidemiological study	0.12

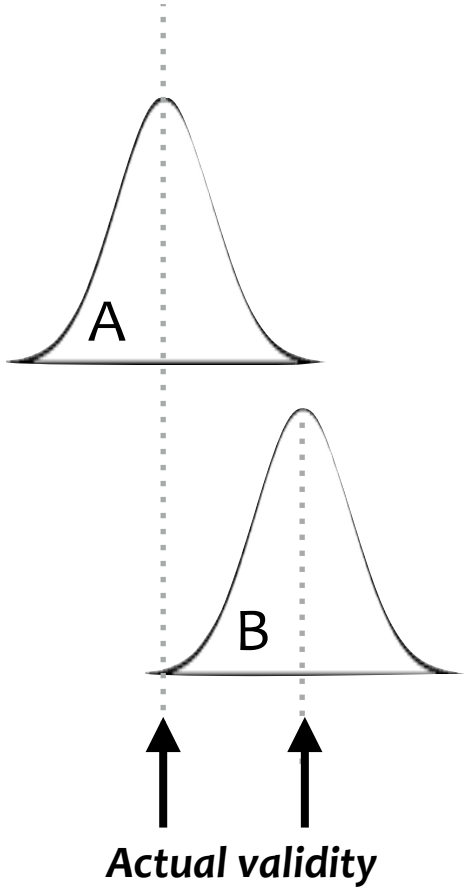
The estimated PPVs (positive pre
RCT, randomized controlled trial.
DOI: 10.1371/journal.pmed.0020124.t004

The structure of the scientific enterprise encourages this

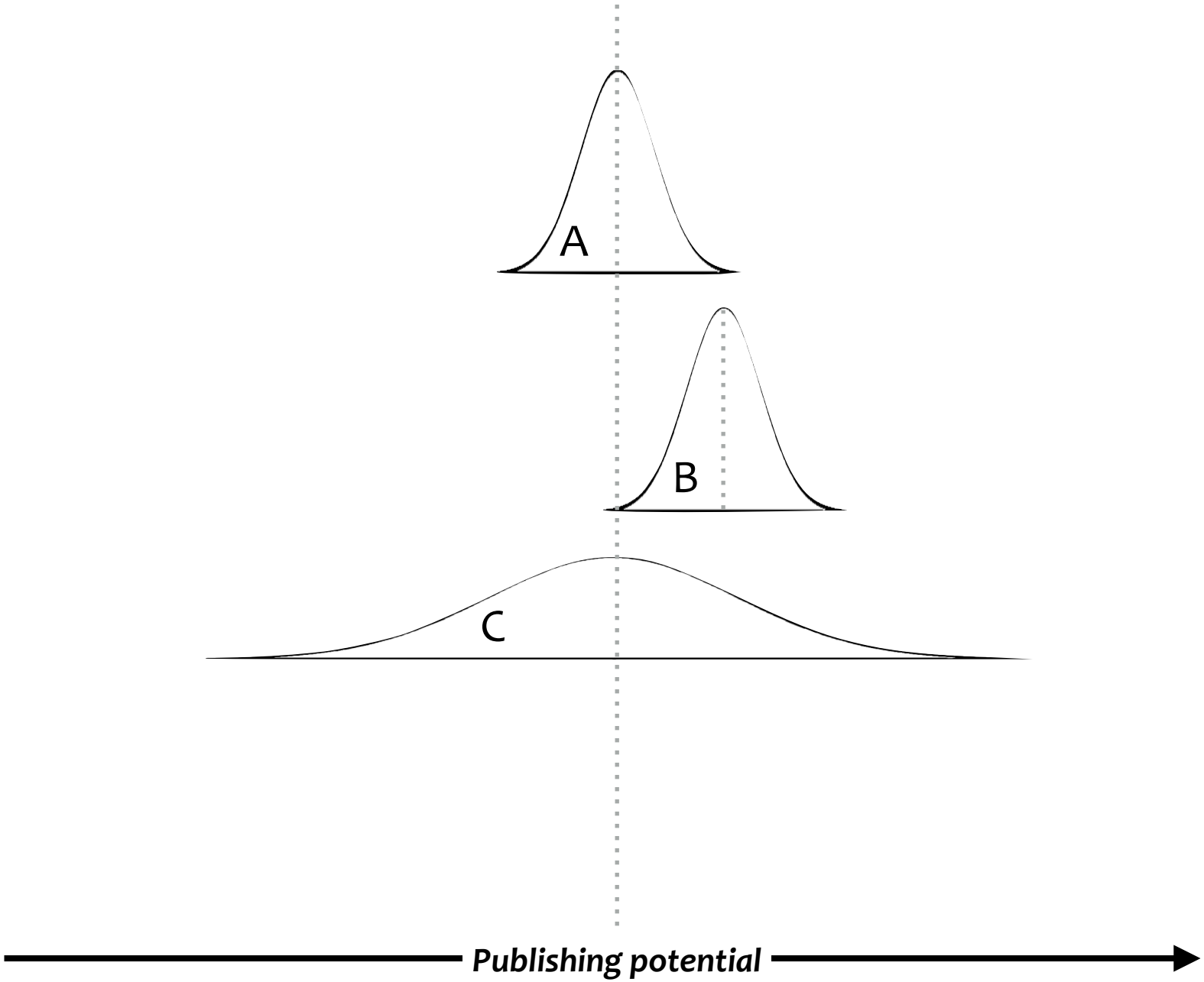
The scientific enterprise produces bias

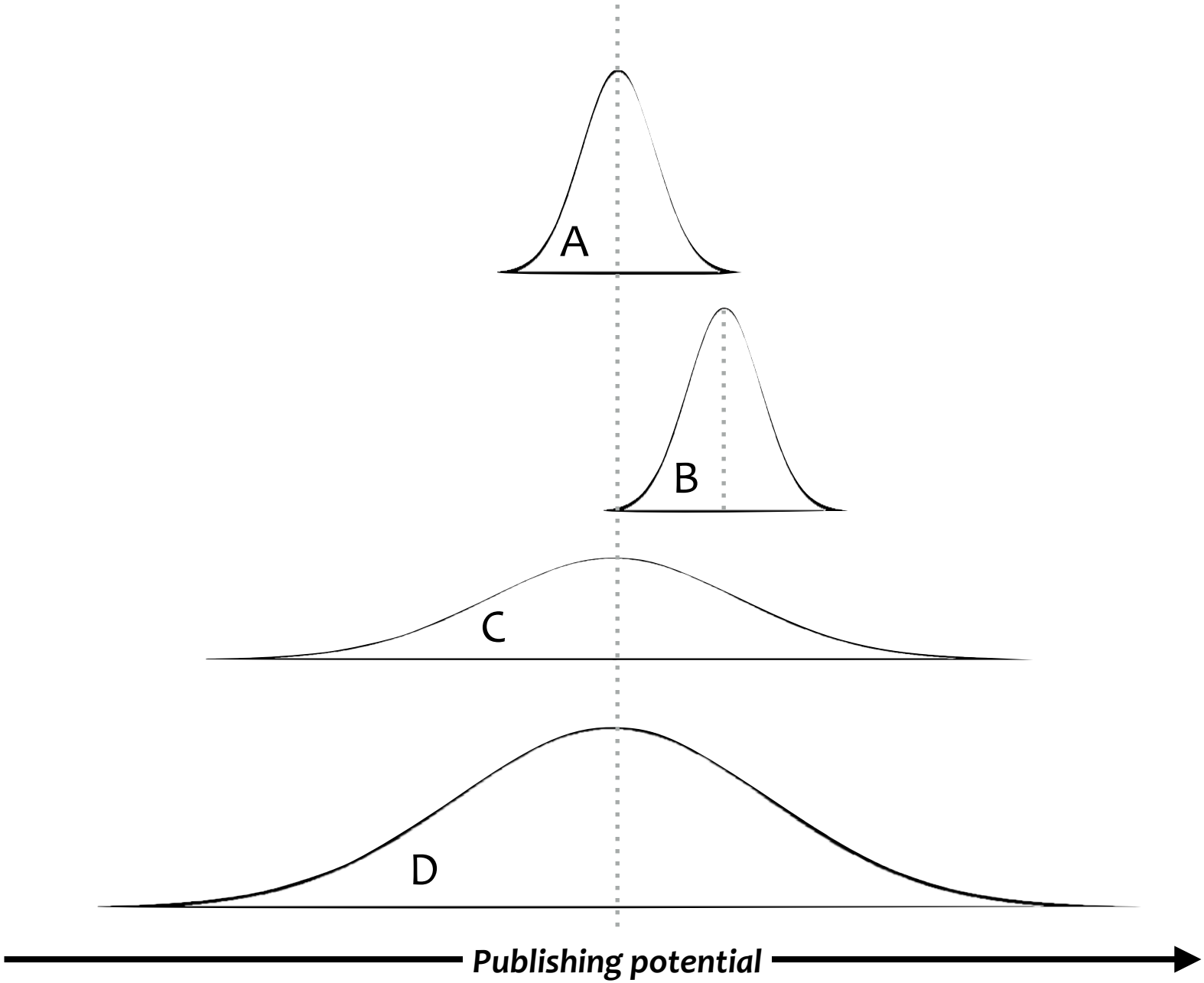
- Any system that…
  - Produces a large number of items
    (e.g., large numbers of potential findings)
  - Scores each item with some variance, and
    (e.g., estimates of significance or effect size)
  - Selects the item with the maximum score
    (e.g., publishes the most significant findings)
- …will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

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



Publishing potential


A
B
C
Publishing potential


A
B
C
D
Publishing potential

THE NEW YORKER

ANNALS OF SCIENCE

THE TRUTH WEARS OFF

Is there something wrong with the scientific method?

BY JONAH LEHRER

DECEMBER 13, 2010

On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in the early nineties. The drugs, sold under brand names such as Abilify, Seroquel, and Zyprexa, had been tested on schizophrenics in several large clinical trials, all of which had demonstrated a dramatic decrease in the subjects' psychiatric symptoms. As a result, second-generation antipsychotics had become one of the fastest-growing and most profitable pharmaceutical classes. By

Many results that are rigorously proved and accepted start shrinking in later studies.

THE NEW YORKER

ANNALS OF SCIENCE

THE TRUTH WEARS OFF

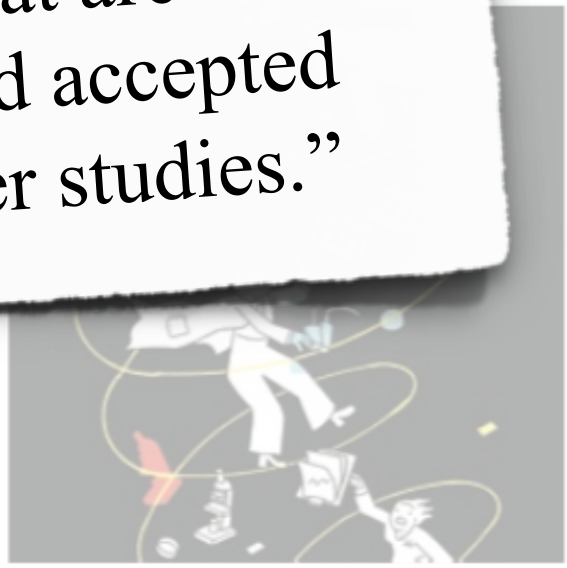
Is there something wrong with the scientific method?

BY JONAH LEHRER

DECEMBER

On Se
psyc
in a hotel
startling n
atypical or
on the mar
brand names such as Abilify, Seroquel, and Zyprexa, had been tested on schizophrenics in several large clinical trials, all of which had demonstrated a dramatic decrease in the subjects' psychiatric symptoms. As a result, second-generation antipsychotics had become one of the fastest-growing and most profitable pharmaceutical classes. By

"Many results that are rigorously proved and accepted start shrinking in later studies."



Many results that are rigorously proved and accepted start shrinking in later studies.

How can we do better?

- Improve individual behavior (reduce variability)
  - Encourage better methodology, more care in research conduct, and higher standards for evidence in reviewing.
  - However, the highest variance groups will still publish more often if other aspects of the system doesn't change.
  - because…

Current systems implicitly reward bias

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

How can we do better?

- Restructure the system to change incentives for individuals (reduce long-term variability)
  - Enable ongoing, rapid, and transparent revision of the scientific literature (far beyond *errata*)
  - Encourage reproduction and replication (e.g., high-profile publication only if easy-to-replicate)
  - Strongly reward long-standing, replicated results (e.g., "test of time" awards)
  - Clearly separate normal revision process from fraud and misconduct

jensen@cs.umass.edu