



Theme 4: Replication crisis



UNIVERSITY OF HELSINKI

Samuli Reijula

University of Helsinki

Example 1: Power posing

Research Report

Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance

Dana R. Carney¹, Amy J.C. Cuddy², and Andy J. Yap¹

¹Columbia University and ²Harvard University

Abstract

Humans and other animals express power through open, expansive postures, and they express powerlessness through closed, contractive postures. But can these postures actually cause power? The results of this study confirmed our prediction that posing in high-power nonverbal displays (as opposed to low-power nonverbal displays) would cause neuroendocrine and behavioral changes for both male and female participants: High-power posers experienced elevations in testosterone, decreases in cortisol, and increased feelings of power and tolerance for risk; low-power posers exhibited the opposite pattern. In short, posing in displays of power caused advantaged and adaptive psychological, physiological, and behavioral changes, and these findings suggest that embodiment extends beyond mere thinking and feeling, to physiology and subsequent behavioral choices. That a person can, by assuming two simple 1-min poses, embody power and instantly become more powerful has real-world, actionable implications.

Keywords

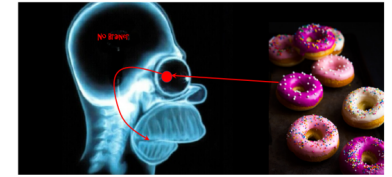
cortisol, embodiment, hormones, neuroendocrinology, nonverbal behavior, power, risk taking, testosterone

ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Psychological Science
21(10) 1363–1368
© The Author(s) 2010
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797610383437
<http://pss.sagepub.com>
SAGE



Example 2: Social priming



Journal of Personality and Social Psychology
1996, Vol. 71, No. 2, 230–244

Copyright 1996 by the American Psychological Association, Inc.
0022-3514/96/\$3.00

Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action

John A. Bargh, Mark Chen, and Lara Burrows
New York University

Previous research has shown that trait concepts and stereotypes become active automatically in the presence of relevant behavior or stereotyped-group features. Through the use of the same priming procedures as in previous impression formation research, Experiment 1 showed that participants whose concept of rudeness was primed interrupted the experimenter more quickly and frequently than did participants primed with polite-related stimuli. In Experiment 2, participants for whom an elderly stereotype was primed walked more slowly down the hallway when leaving the experiment than did control participants, consistent with the content of that stereotype. In Experiment 3, participants for whom the African American stereotype was primed subliminally reacted with more hostility to a vexatious request of the experimenter. Implications of this automatic behavior priming effect for self-fulfilling prophecies are discussed, as is whether social behavior is necessarily mediated by conscious choice processes.

Example 3: Ego depletion



PERSONALITY PROCESSES AND INDIVIDUAL DIFFERENCES

Ego Depletion: Is the Active Self a Limited Resource?

Roy F. Baumeister, Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice
Case Western Reserve University

Choice, active response, self-regulation, and other volition may all draw on a common inner resource. In Experiment 1, people who forced themselves to eat radishes instead of tempting chocolates subsequently quit faster on unsolvable puzzles than people who had not had to exert self-control over eating. In Experiment 2, making a meaningful personal choice to perform attitude-relevant behavior caused a similar decrement in persistence. In Experiment 3, suppressing emotion led to a subsequent drop in performance of solvable anagrams. In Experiment 4, an initial task requiring high self-regulation made people more passive (i.e., more prone to favor the passive-response option). These results suggest that the self's capacity for active volition is limited and that a range of seemingly different, unrelated acts share a common resource.

Reproducibility Project: Psychology (RPP)



RESEARCH

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration^{*†}

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically 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.

Many Labs experiments 1-5



Many labs 1 (2012): 10 of the 13 studies replicated their original findings

Many Labs 2 (2018): Successfully replicated 14 of 28

Many Labs 3 (2014): Did timing (semester) make a difference? no

Many Labs 4 (2022): Original author involved in replication (Terror Management Theory). Did not replicate

Many Labs 5: (2020): Replicating replications from RPP. No improvement in replicability

New Many Labs projects: Many Babies, Many Smiles, Many Dogs, Many Birds, Many EEGs,

Social-science replication project (SSRP)

nature
human behaviour

LETTERS

<https://doi.org/10.1038/s41562-018-0399->

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

Colin F. Camerer^{1,16}, Anna Dreber^{2,16}, Felix Holzmeister^{3,16}, Teck-Hua Ho^{4,16}, Jürgen Huber^{3,16}, Magnus Johannesson^{5,16}, Michael Kirchler^{3,5,16}, Gideon Nave^{6,16}, Brian A. Nosek^{7,8,16*}, Thomas Pfeiffer^{9,16}, Adam Altmejd^{10,2}, Nick Buttrick^{7,8}, Taizan Chan¹⁰, Yiling Chen¹¹, Eskil Forsell¹², Anup Gampa^{7,8}, Emma Heikensten², Lily Hummer⁸, Taisuke Imai¹³, Siri Isaksson², Dylan Manfredi¹⁴, Julia Rose³, Eric-Jan Wagenmakers¹⁴ and Hang Wu¹⁵

Being able to replicate scientific findings is crucial for scientific progress^{1–15}. We replicate 21 systematically selected experimental studies in the social sciences published in *Nature* and *Science* between 2010 and 2015^{16–36}. The replications follow analysis plans reviewed by the original authors and pre-registered prior to the replications. The replications are high powered, with sample sizes on average about five times higher than in the original studies. We find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size. Replicability varies between 12 (57%) and 14 (67%) studies for complementary replicability indicators. Consistent with these results, the estimated true-positive rate is 67% in a Bayesian analysis. The relative effect size of true positives is estimated to be 71%, suggesting that both false positives and inflated effect sizes of true positives contribute to imperfect reproducibility. Furthermore, we find that peer beliefs of replicability are strongly related to replicability, suggesting that the research community could predict which results would replicate and that failures to replicate were not the result of chance alone.

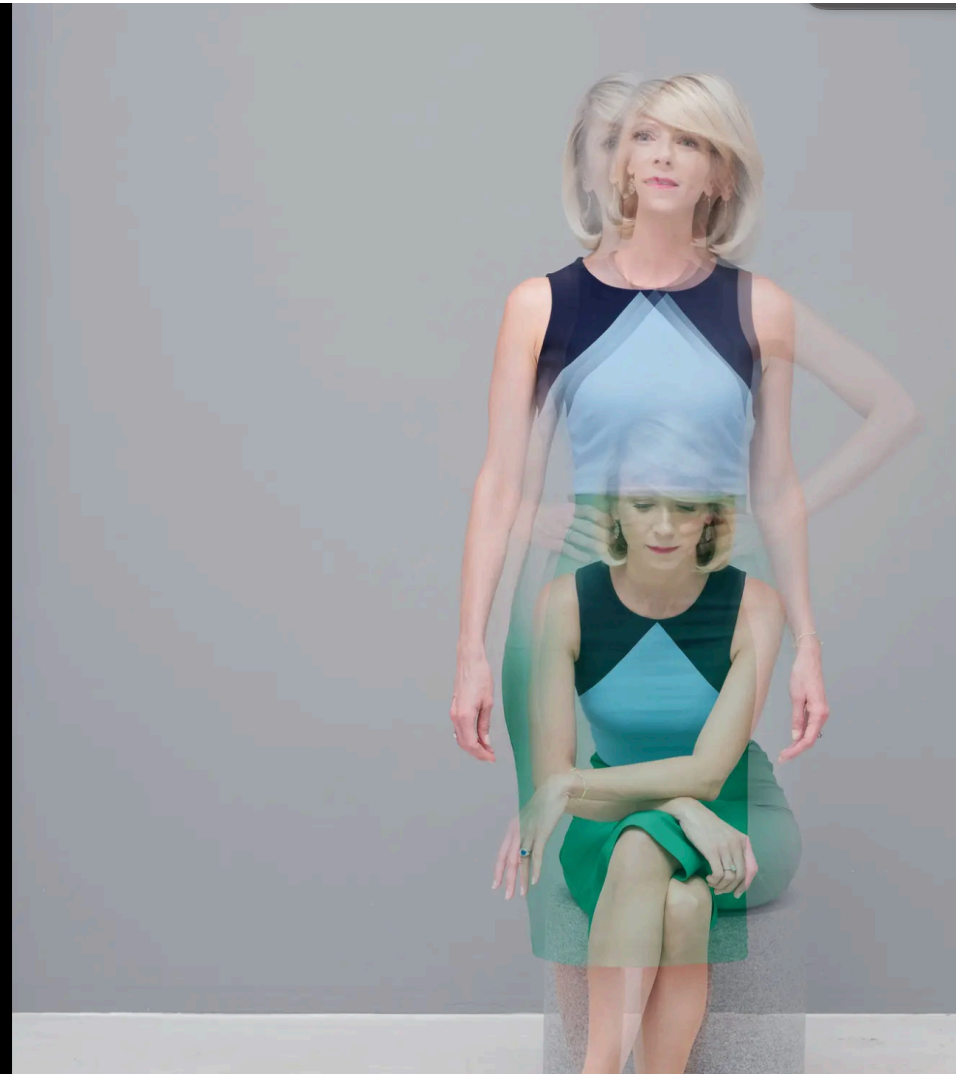
a significant effect in the same direction as the original studies 61% of replications¹³. Both the RPP and the EERP had high statistical power to detect the effect sizes observed in the original studies. However, the effect sizes of published studies may be inflated even for true-positive findings owing to publication or report biases^{40–42}. As a consequence, if replications were well powered to detect effect sizes smaller than those observed in the original studies, replication rates might be higher than those estimated in the RPP and the EERP.

We provide evidence about the replicability of experimental studies in the social sciences published in the two most prestigious general science journals, *Nature* and *Science* (the Social Sciences Replication Project (SSRP)). Articles published in these journals are considered exciting, innovative and important. We include all experimental studies published between 2010 and 2015 that (1) test for an experimental treatment effect between or within subjects, (2) test at least one clear hypothesis with a statistically significant finding, and (3) were performed on student or other accessible subject pools. Twenty-one studies were identified to meet these criteria. We used the following three criteria, in descending order to determine which treatment effect to replicate:

FEATURE

When the Revolution Came for Amy Cuddy

As a young social psychologist, she played by the rules and won big: an influential study, a viral TED talk, a prestigious job at Harvard. Then, suddenly, the rules changed.





DAN ARIELY

New York Times bestselling author of
Predictably Irrational and *The Upside of Irrationality*



THE (HONEST) TRUTH ABOUT DISHONESTY

How We Lie to Everyone—Especially Ourselves

ANNALS OF INQUIRY

THEY STUDIED DISHONESTY. WAS THEIR WORK A LIE?

Dan Ariely and Francesca Gino became famous for their research into why we bend the truth. Now they've both been accused of fabricating data.

By Gideon Lewis-Kraus

September 30, 2023

Replication controversies

Replication attempts often lead to conflict

In 2012, Doyen et al. published a failed replication of Bargh's social priming experiment (example 2)

Bargh's response, entitled "Nothing in their heads":

- Doyen's group "incompetent or ill-informed"
- Doyen's publication venue (PLOS One) questionable
- Tacit knowledge needed to perform successful replications
- Small but crucial disparities between the original and the replication attempt block successful replication



An instance of **experimenter's regress**: only a successful outcome is a certain indicator that the experiment is run properly - but the sides of the debate disagree about both issues

Researchers from both sides of the replication controversy questioned their adversaries' findings, methods and competence

A reaction to the crisis: metascience

Metascience/metaresearch:

A new research field / scientific/intellectual movement that arose as a reaction to the replication crisis (Peterson & Panofsky 2023)

Well-funded, institutionalization through Metascience conferences (2019→)

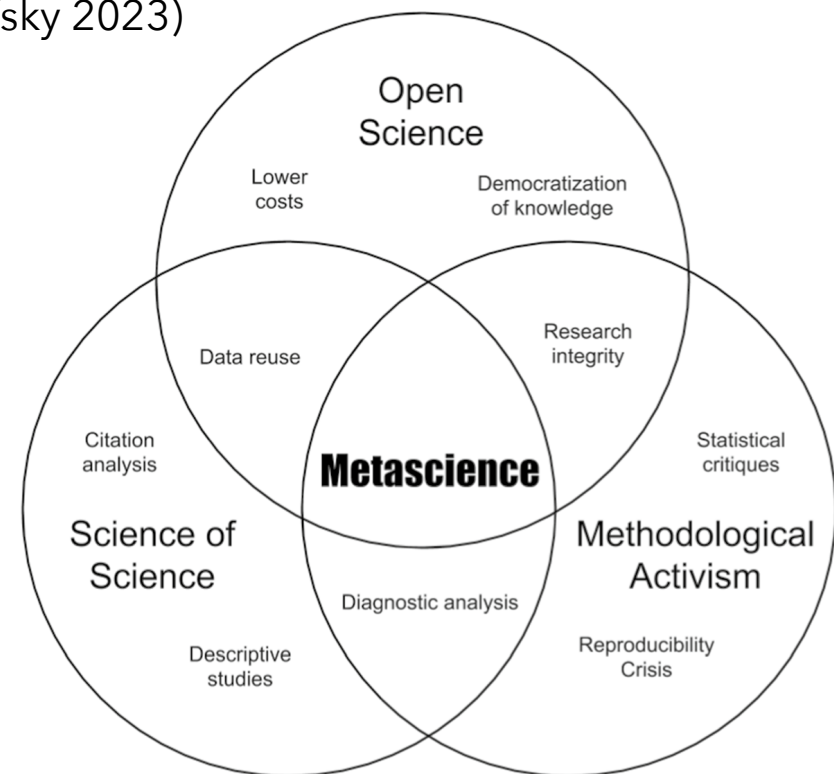
Practitioner (psychology) led

- methodological activism
- open science advocacy
- science of science methods

→ Statistical and institutional correctives

Little interaction with STS or philosophy of science

- social-science sensitivity missing (e.g. unintended consequences of institutional changes)
- reinventing the wheel (philosophy of science)



The crisis timeline

2005: Ioannidis: Why most published findings are false

2010 Retraction Watch website online

2011 Many Labs project started

2012 "I see a train wreck looming," -- Daniel Kahneman

2013 Data Colada blog online

2015 Replicability project: Psychology published

...

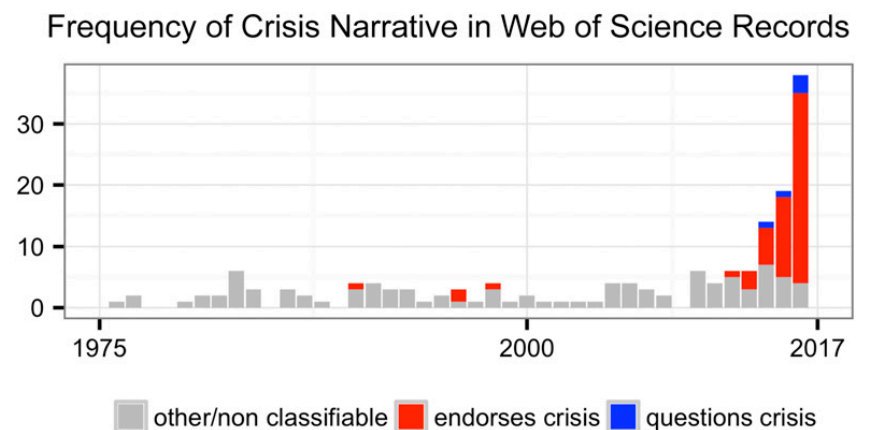
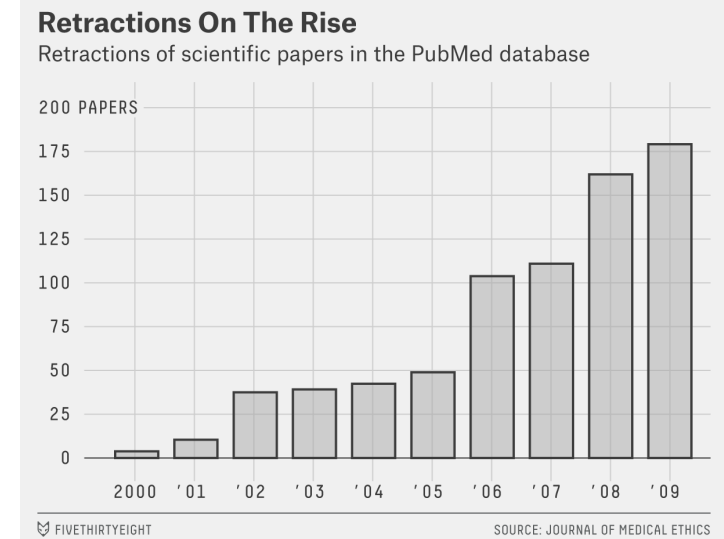


Fig. 1. Number of Web of Science records that in the title, abstract, or keywords contain one of the following phrases: "reproducibility crisis," "scientific crisis," "science in crisis," "crisis in science," "replication crisis," "replicability crisis."

Prediction market for psychologists

Dreber et al. 2015 set up a prediction market for psychologists: could researchers themselves predict which studies would replicate?

44 studies drawn from RPP (before the results came in)

Prediction markets correctly predict the outcome of 71% of the replications

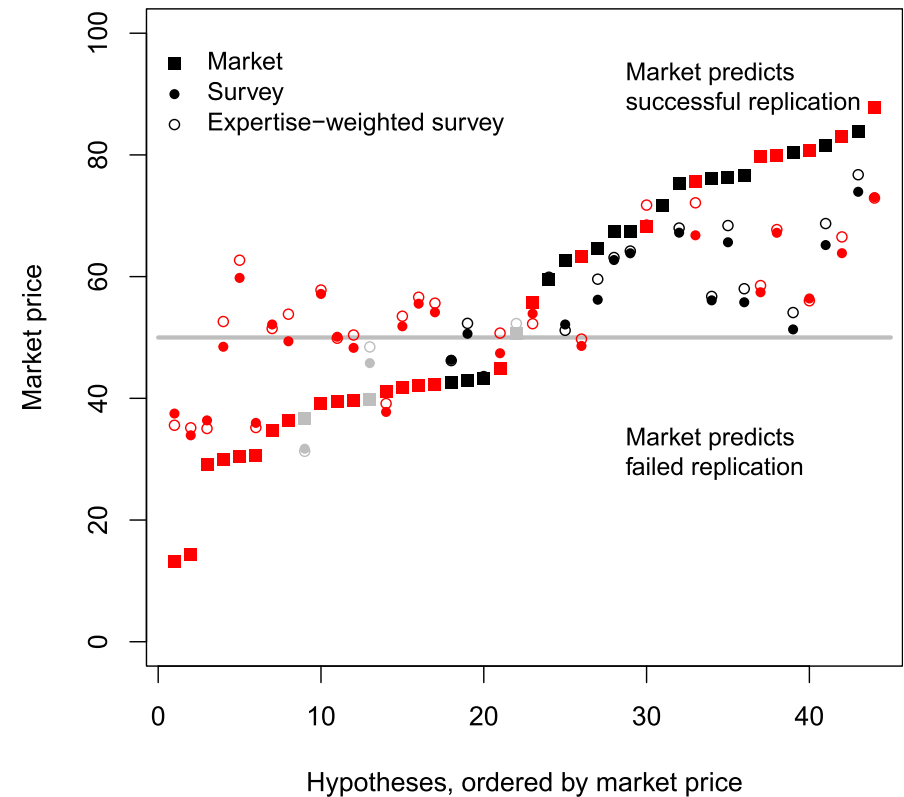


Fig. 1. Prediction market performance. Final market prices and survey predictions are shown for the replication of 44 publications from three top psychology journals. The prediction market predicts 29 out of 41 replications correctly, yielding better predictions than a survey carried out before the trading started. Successful replications (16 of 41 replications) are shown in black, and failed replications (25 of 41) are shown in red. Gray symbols are replications that remained unfinished (3 of 44).

groundwork

Why do we care about replication?

Why do we care about replication

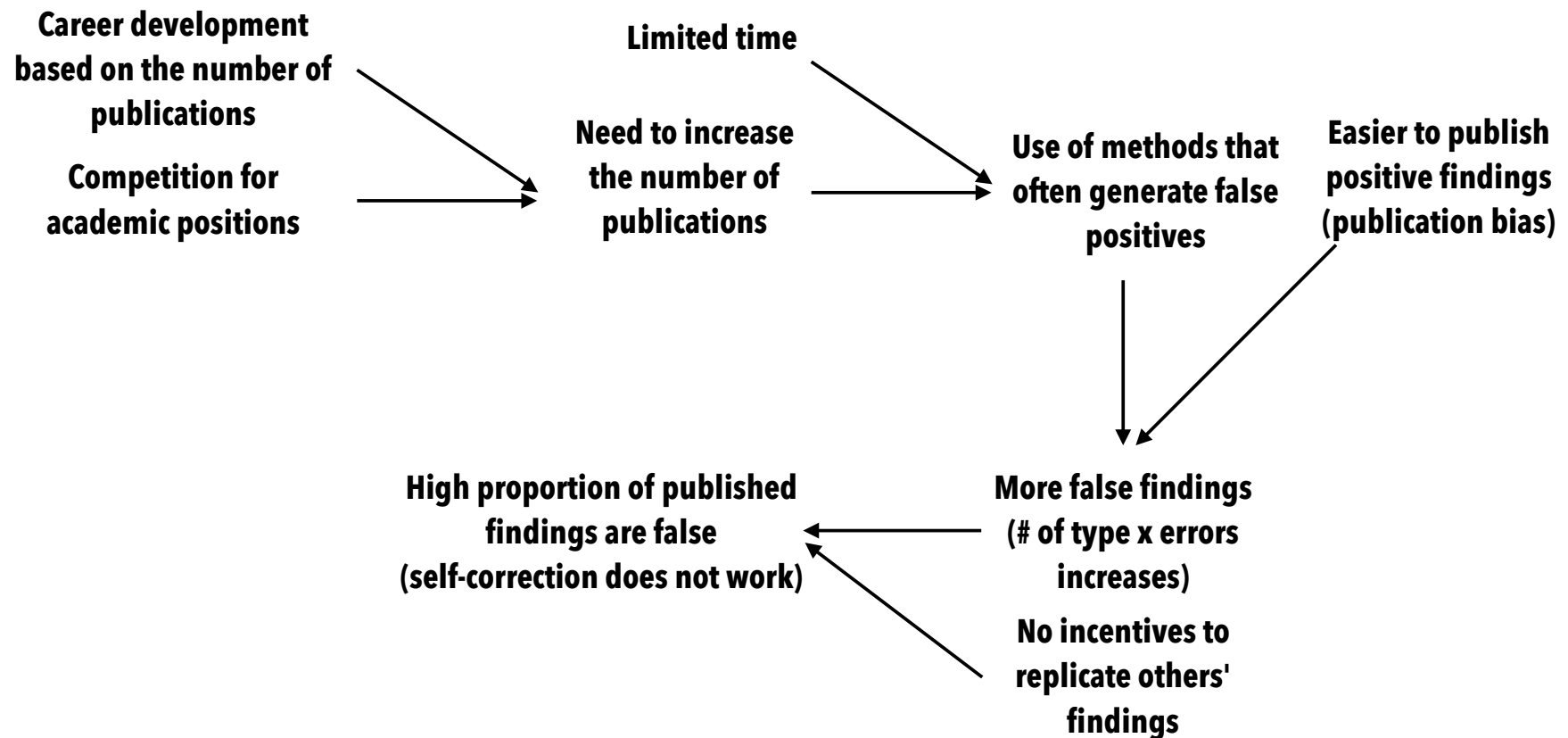
**Zwaan et al (2018):**

"The ability to systematically replicate research findings is a fundamental feature of the scientific process. Indeed, the idea that observations can be recreated and verified by independent sources is usually seen as a bright line of demarcation that separates science from non-science (Dunlap 1926). A defining feature of science is that researchers do not merely accept claims without being able to critically evaluate the evidence for them (e.g., Lupia & Elman 2014). Independent replication of research findings is an essential step in this evaluation process, and thus, replication studies should play a central role in science and in efforts to improve scientific practices."

Karl Popper (1959):

"Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested – in principle – by anyone. We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence', but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable."
(The Logic of Scientific Discovery, pp. 23-24)

Conditions for failure of scientific self-correction




Note: self-correction in science breaks down only when both:

1. there are lots of false positive findings in the literature
2. false positives not detected and removed

Replication types

(Schmidt 2009: Shall we really do it again? The powerful concept of replication is neglected in the social sciences.)



Direct replication

- The same experimental protocol applied to the same kind of materials (for instance, individuals taken from the population originally studied)
- The experiment should give an outcome that is the same or at least similar to that originally obtained
- Feest 2022: No identical situations → when is the situation similar enough?

Conceptual replication

- Attempt to see an effect in the same direction as that originally reported, using a different experimental protocol and/or materials
- Addresses the same theoretical claim but with different experimental materials
- Goal: generalizing a finding or testing its robustness

The resampling account of replication

(Machery 2020. What is a replication? Philosophy of Science)

"Experiment A replicates experiment B if and only if A consists of a sequence of events of the same type as B while resampling some of its experimental components in order to assess the reliability of the original experiment. "

Experimental components

- experimental units
- treatments, independent variables
- measurements, dependent variables
- settings

Each unit can be fixed or random

- if a factor is held fixed, the experimenter does not aim to generalize to other values of that factor

Implications:

- a new analysis of reliability and validity
- an account of the function of replications: checking reliability of prior experiments
- distinction between replication and extension
- notion of conceptual replication confused

**where do false positives come
from**

Ioannidis 2005: Why most published research findings are false

Ioannidis' argument focuses on hypothesis testing (e.g. in epidemiology)

The crucial factors (often ignored):

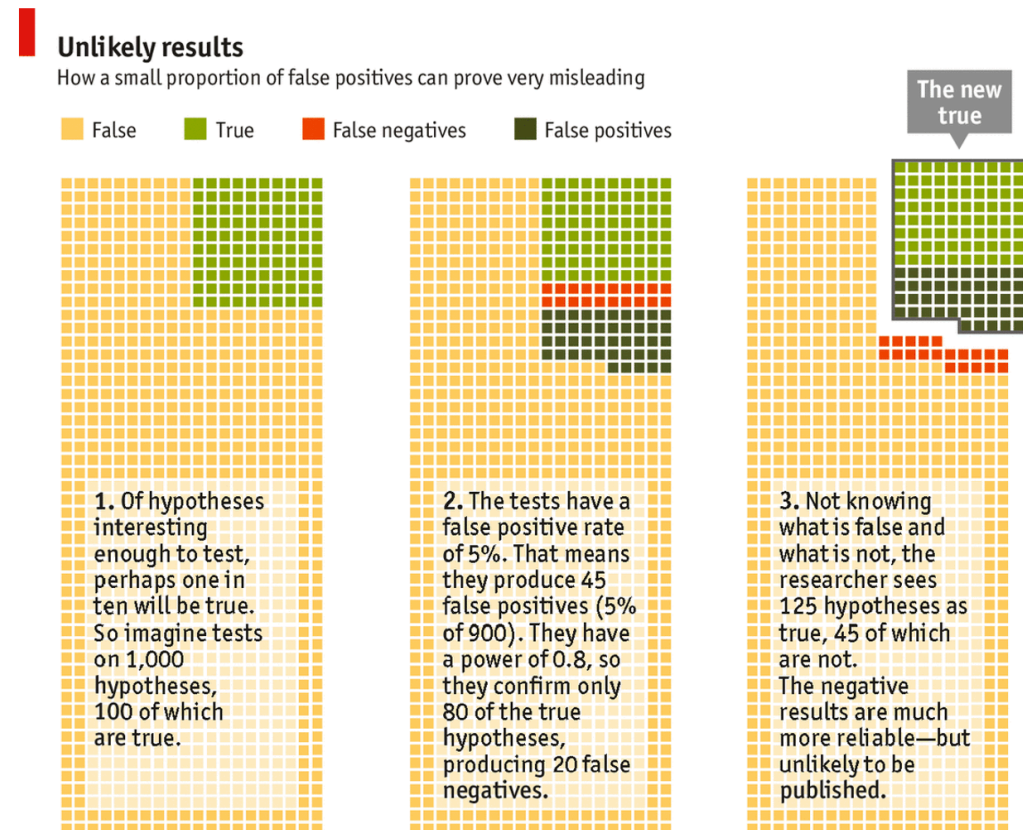
1. **Power of the study** (a measure of avoiding type II error)

- IF sample ↓ THEN power ↓
- IF effect size ↓ THEN power ↓
- IF power ↓ THEN error ↑

2. **Unlikeliness of the hypothesis** being tested

- IF prior probability of hypothesis ↑ THEN error ↓

"In the described framework, a positive predictive value exceeding 50% is quite difficult to get"



Source: *The Economist*

[Click here to watch an animation of this diagram](#)

A simple explanation here: <https://www.economist.com/briefing/2013/10/18/trouble-at-the-lab>

Research culture in psychology (observations)

At least in some parts of psychological research:

- Little overarching theory development
- Novelty strongly rewarded
- Status boost (in the psych department) from publishing bizarre findings ("environmental factor x influences behavioral property y"=))
- Not much attention to mechanisms underlying the phenomenon



Studied hypotheses from RPP

(Dreber et al 2015, supplementary information)

Table S3. Hypotheses for the 23 replication studies in the first set of prediction markets

Ref.	Hypothesis
33	White participants with high external motivation to respond without prejudice toward Blacks have an attentional bias toward neutral Black faces presented for 30 ms, but have an attentional bias away from neutral Black faces presented for 450 ms. These biases are eliminated when the faces display happy expressions.
34	Participants do not exhibit a delay in response when switching between pronouncing regular words and pronouncing nonwords.
35	Naive participants' judgments of the power and leadership of CEO faces are correlated positively with their companies' profits.
36	Repetition blindness (a reduction in reporting seeing an orthographically identical or similar word when it is presented in close temporal proximity amid a series of rapidly presented words or nonwords) will occur even for nonidentical orthographical neighbors (e.g., boss and bass) even when the stimuli are nonwords and when they are never repeated in the string of stimuli.
37	An increase in participants' public moral image will be related to an increased willingness to reconcile only for perpetrators, whereas an increase in participants' sense of power will be related to an increased willingness to reconcile only for victims.
38	Participants instructed to avoid race or use race in categorizing tools and guns exhibited less 1/f noise than participants in a control condition where no mention of race was made.
39	Participants with reduced self-regulation resources are expected to exhibit more pronounced confirmatory information processing than nondepleted and ego-threatened participants, whereas no significant differences regarding confirmatory information processing are expected between nondepleted and ego-threatened participants.
40	Participants will prefer descriptions of the city of Los Angeles that are more concrete/less abstract when they are exposed to the words "Los Angeles" during an earlier exercise. Participants who are not shown "Los Angeles" during this earlier exercise will prefer relatively less concrete/more abstract descriptions of the city of Los Angeles.
41	Word processing is slower for dense near semantic neighborhoods, i.e., words with many near neighbors are processed more slowly than words with few near neighbors.
42	Words denoting objects that typically occur high in the visual field hinder identification of targets appearing at the top of the display, whereas words denoting low objects hinder target identification at the bottom of the display.
43	Survival processing yields better memory retention than a control condition with a contextually rich (but non-survival-relevant) encoding scenario.
44	When there are no nonoccurrences of the outcome in the presence of just one cause (cause A), increasing the number of occurrences of the outcome in the presence of that cause alone does not alter the conditional contingency. Under the conditional contingency hypothesis, therefore, such manipulations should not have a significant effect on causal judgment. As opposed to this, the tested predictions are that (i) such occurrences raise judgments of A as cause for the outcome and (ii) lower judgments of an alternative cause B.
45	When participants read sequences of digits and a task requires the joint processing of nonadjacent pairs of digits, they learn exclusively the relation between these nonadjacent digits and not relations between adjacent digits, thus suggesting attention instead of spatial contiguity as the critical factor.
46	Drug use is positively correlated with learning from experience under "sunny" conditions (in which win-loss probabilities are known before making a series of choices) but not correlated under "cloudy" conditions (in which the win-loss probabilities are not known in advance and can only be learned through trial and error).
47	Drinking lemonade with sugar reduces the attraction effect (the reliance on intuitive, heuristic-based decision making) compared with drinking lemonade with sugar substitute among subjects with depleted mental resources.

Researcher degrees of freedom

(Simmons et al 2011: False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science)

Flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates

Several decisions to be made in research:

- Should more data be collected?
- Should some observations be excluded?
- Which variables should be reported?
- Which control variables should be considered?
- ...

Often impractical to make all such decisions beforehand, ...

... but if done during/after data collection, data can influence analytic techniques → probability of getting false positive results can go up dramatically (= "curve fitting")

- a further needed assumption: confirmation bias: decisions made so as to allow publishing positive findings

Often such "**p-hacking**" is not intentional, but a failure of imagination (failure to consider: how likely is it that I got this result; how else could the experiment have turned out and why?)



A garden of forking paths
(see Gelman & Loken

try it:

<https://fivethirtyeight.com/features/science-isnt-broken/>

see also

<https://shinyapps.org/apps/p-hacker/>

A hidden universe of uncertainty

(Breznau et al 2022: Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Idiosyncratic Uncertainty, PNAS)

PNAS

RESEARCH ARTICLE

SOCIAL SCIENCES



Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty

Edited by Douglas Massey, Princeton University, Princeton, NJ; received March 6, 2022; accepted August 22, 2022

This study explores how researchers' analytical choices affect the reliability of scientific findings. Most discussions of reliability problems in science focus on systematic biases. We broaden the lens to emphasize the idiosyncrasy of conscious and unconscious decisions that researchers make during data analysis. We coordinated 161 researchers in 73 research teams and observed their research decisions as they used the same data to independently test the same prominent social science hypothesis: that greater immigration reduces support for social policies among the public. In this typical case of social science research, research teams reported both widely diverging numerical findings and substantive conclusions despite identical start conditions. Researchers' expertise, prior beliefs, and expectations barely predict the wide variation in research outcomes. More than 95% of the total variance in numerical results remains unexplained even after qualitative coding of all identifiable decisions in each team's workflow. This reveals a universe of uncertainty that remains hidden when considering a single study in isolation. The idiosyncratic nature of how researchers' results and conclusions varied is a previously underappreciated explanation for why many scientific hypotheses remain contested. These results call for greater epistemic humility and clarity in reporting scientific findings.

Significance

Will different researchers converge on similar findings when analyzing the same data? Seventy-three independent research teams used identical cross-country survey data to test a prominent social science hypothesis: that more immigration will reduce public support for government provision of social policies. Instead of convergence, teams' results varied greatly, ranging from large negative to large positive effects of immigration on social policy support. The choices made by the research teams in designing their statistical tests explain very little of this variation; a hidden universe of uncertainty remains. Considering this variation, scientists, especially those working with the complexities of human societies and behavior, should exercise humility and strive to better account for the uncertainty in their work.

A fix: Registered reports

New procedures:

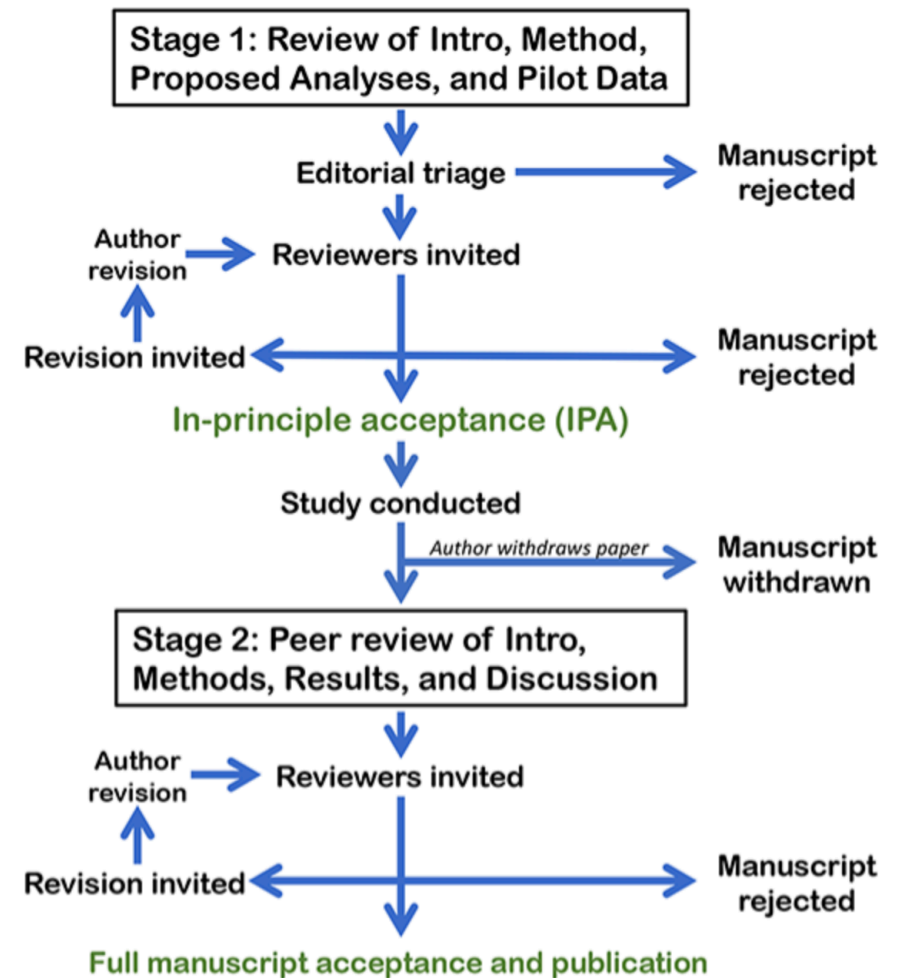
For author. Fixed protocol: Data collected only after design decisions are made

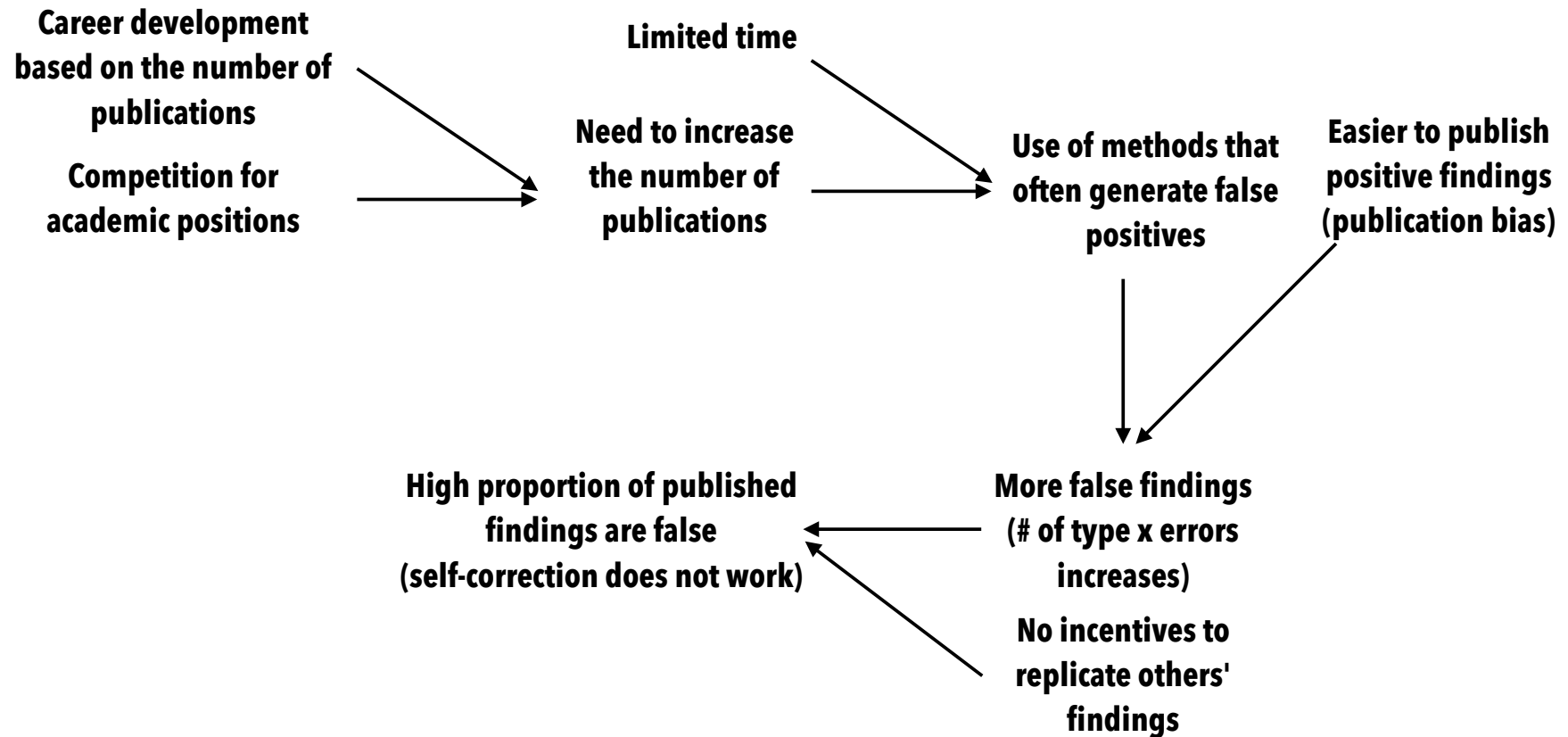
For journal. If study has been conducted according to protocol, journal publishes irrespective of whether results positive or negative

→ Helps with the researcher-degrees-of-freedom problem and publication bias

Not all studies should be preregistered, only **confirmatory research**

Fishing expeditions allowed in **exploratory research**





Often (esp. in exploratory research) there are good reasons to use methods that generate even a high proportion of false positives, e.g. searching "weak signals"

Note: self-correction in science breaks down only when both

1. lots of false positive findings in the literature
2. false positives not detected and removed

**why are there not enough
replication attempts?**

Why are there not enough replication attempts?



- often not enough information on the original publication to attempt replication
- expensive / no funding for replications
- controversies
- hard to get published (publication bias)
- under-rewarded (the priority rule)
- ...

Scientific Utopias I-III

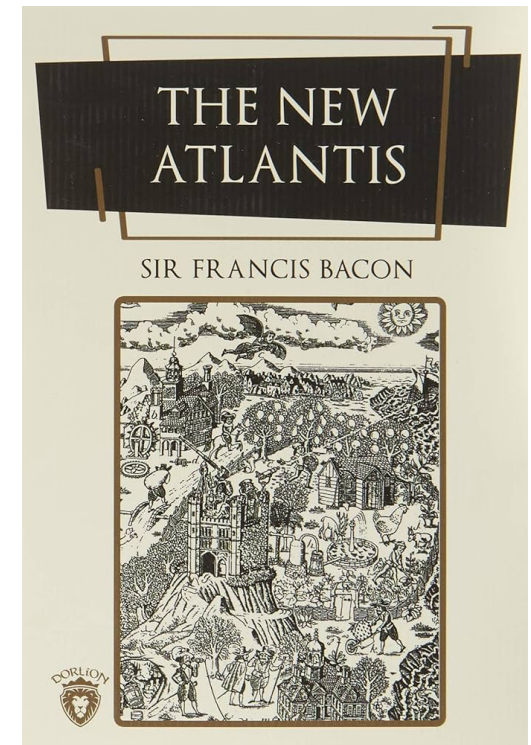
(e.g. Nosek et al. 2012. Scientific utopia II: Resutructuring incentives and practices to promote truth over publishability @Perspectives on Psychological Science)

Publishability \neq truth

- "the solution requires making incentives for getting it right competitive with the incentives for getting it published"

Accuracy motive vs. professional motives

- motivated reasoning: justifying research decisions in the name of accuracy when they actually serve career advancement
- motivated reasoning particularly influential when the situation is complex, the available information is ambiguous
- motivated reasoning not always intentional
 - details of design decisions hard to remember.
"Forgetting the details provides an opportunity for reimaging the study purpose and results to recall and understand them in their best (i.e., most publishable) light. "

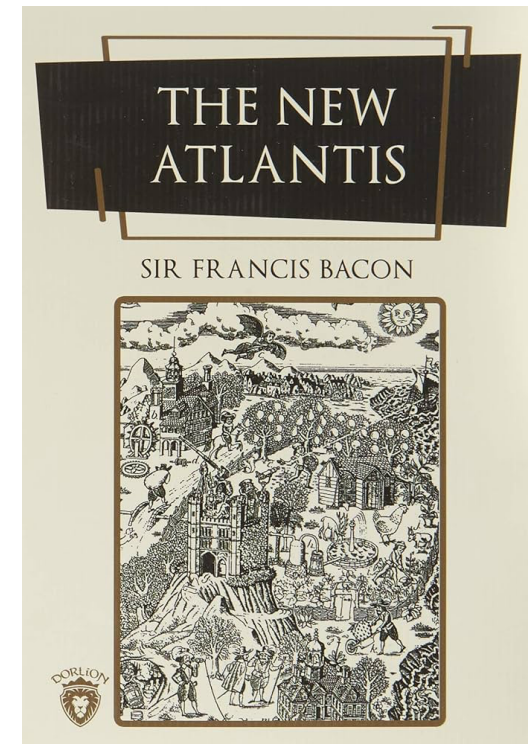


Scientific Utopias I-III

(e.g. Nosek et al. 2012. Scientific utopia II: Resutructuring incentives and practices to promote truth over publishability @Perspectives on Psychological Science)


Imagined solutions (utopias):

- strengthening long-term goals (getting it right) vs. short-term ones (getting it published)
 - promoting and rewarding paradigm-driven research
 - Author, reviewer, and editor checklists
 - Metrics for identifying important papers to replicate
 - Diversifying peer review practices
 - Lowering or removing the barrier for publication
 - Transparency: opening the data and the scientific workflow
-
- encouraging high-quality peer review
 - publishing reviews as scientific contributions
 - → a new role in scientific community, expert reviewer



Manifesto for reproducible science

(Munafo et al. 2017)



Improvements:

methods

- blinding (at many stages of research)
- methods training & support
- encouraging collaboration & team science (#diversity)

reporting

- pre-registration
- reporting guidelines (e.g. CONSORT)

reproducibility

- transparency & open science

evaluation

- diversifying peer review (e.g. pre- AND post-publication review)

incentives

- rewarding also for carefulness, replication, not only innovation

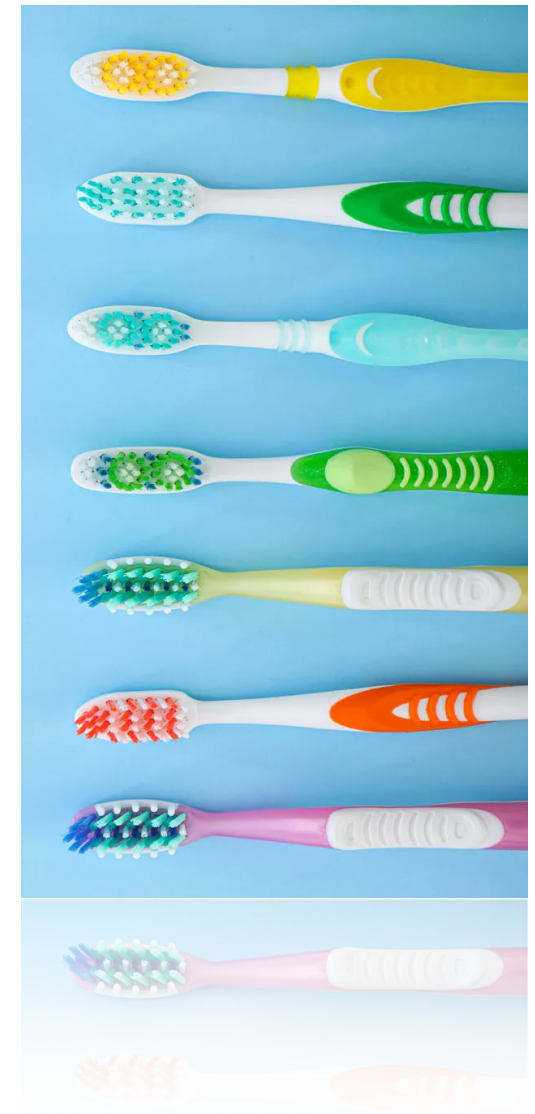
Theory Crisis

Gigerenzer 2010:

- present-day psychology such a patchwork of small territories, resembling, to use a political metaphor, Italy or Germany before unification around 1870
- Watkins (1984) wrote that a cognitive theory "is a bit like someone else's toothbrush—it is fine for that individual's use, but for the rest of us ... well, we would just rather not, thank you" (p. 86).'
- "Theory construction should be taught in graduate school"

Muthukrishna & Henrich 2019: Problem in theory. @ Nature Human Behavior

- methodological repairs (e.g., preregistration) are needed, but the problem runs deeper: no shared theory
- no general theory from which to derive testable hypotheses
- diagnosis: compare psych textbooks from those from other sciences (e.g. Econ: RCT!)



Playing 20 questions with nature



Newell 1973: You can't play 20 questions with nature and win

- Psychology in 1973:
 - small experiments
 - not derived from general theory
 - experiments do not contribute to general questions
 - no "coordination" between the small experiments

Almaatouq et al: **Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences** @ forthcoming in the *Behavioral and Brain Sciences*

the aftermath

Factors underlying false positives



A spectrum ...

- from outright fraud
- to questionable research practices (QRPs)
- unintentional methodological errors
- to pure bad luck (at $\alpha=.05$, 5% of times we get a positive when null is true)

Heterogeneity revolution

(Bryan, Tipton, Yeager 2021: Behavioural science is unlikely to change the world without a heterogeneity revolution. Nature Human Behavior)

"The recognition that most treatment effects are heterogeneous, so the variation in effect estimates across studies that defines the replication crisis is to be expected as long as heterogeneous effects are studied without a systematic approach to sampling and moderation"

When studied systematically, heterogeneity can be leveraged to build more complete theories of causal mechanism that could inform nuanced and dependable guidance to policymakers.

nature
human behaviour

PERSPECTIVE

<https://doi.org/10.1038/s41562-021-01143-3>

 Check for updates

Behavioural science is unlikely to change the world without a heterogeneity revolution

Christopher J. Bryan ¹✉, Elizabeth Tipton ²✉ and David S. Yeager ¹✉

In the past decade, behavioural science has gained influence in policymaking but suffered a crisis of confidence in the replicability of its findings. Here, we describe a nascent heterogeneity revolution that we believe these twin historical trends have triggered. This revolution will be defined by the recognition that most treatment effects are heterogeneous, so the variation in effect estimates across studies that defines the replication crisis is to be expected as long as heterogeneous effects are studied without a systematic approach to sampling and moderation. When studied systematically, heterogeneity can be leveraged to build more complete theories of causal mechanism that could inform nuanced and dependable guidance to policymakers. We recommend investment in shared research infrastructure to make it feasible to study behavioural interventions in heterogeneous and generalizable samples, and suggest low-cost steps researchers can take immediately to avoid being misled by heterogeneity and begin to learn from it instead.

Psychology's renaissance

(Nelson et al. 2018 @ Annual Review of Psychology)

Abstract

In 2010–2012, a few largely coincidental events led experimental psychologists to realize that their approach to collecting, analyzing, and reporting data made it too easy to publish false-positive findings. This sparked a period of methodological reflection that we review here and call Psychology's Renaissance. We begin by describing how psychologists' concerns with publication bias shifted from worrying about file-drawered studies to worrying about *p*-hacked analyses. We then review the methodological changes that psychologists have proposed and, in some cases, embraced. In describing how the renaissance has unfolded, we attempt to describe different points of view fairly but not neutrally, so as to identify the most promising paths forward. In so doing, we champion disclosure and preregistration, express skepticism about most statistical solutions to publication bias, take positions on the analysis and interpretation of replication failures, and contend that meta-analytical thinking *increases* the prevalence of false positives. Our general thesis is that the scientific practices of experimental psychologists have improved dramatically.

If a team of research psychologists were to emerge today from a 7-year hibernation, they would not recognize their field.

- Authors voluntarily posting their data.
- Top journals routinely publishing replication attempts, both failures and successes.
- Hundreds of researchers preregistering their studies.
- Crowded methods symposia at many conferences.
- Enormous increases in sample sizes.
- Some top journals requiring the full disclosure of measures, conditions, exclusions, and the rules for determining sample sizes.
- Several multilab replication efforts accepted for publication before any data were collected.

Overall, an unprecedented focus on replicability. What on earth just happened?