



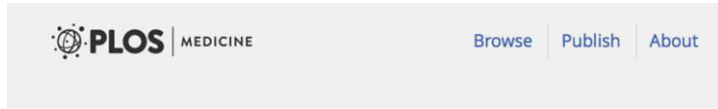
Metascience - how to improve science using science

Ljiljana B. Lazarević
Institute of psychology and LIRA lab, University of Belgrade, Serbia



"Science is the best thing that has happened to human beings ... but we can do it better."

John Ioannidis



OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • <http://dx.doi.org/10.1371/journal.pmed.0020124>

The **smaller the studies** conducted in a scientific field

The **smaller the effect sizes** in a scientific field

The **greater the number and the lesser the selection of tested relationships** in a scientific field

The **greater the flexibility in designs, definitions, outcomes, and analytical modes** in a scientific field

The **greater the financial and other interests and prejudices** in a scientific field

The **hotter a scientific field** (with more scientific teams involved)



the **less likely** the research findings are to be true.

What is Metascience?

Metascience, also known as **meta-research**, **research on research**, **the science of science** - is the use of scientific methodology to study science itself.

Metascience seeks to increase the quality of scientific research while reducing inefficiency.

Metascience uses research methods to study how research is done and what improvements can be made.

“Rhumba of R’s” - what constitutes scientific integrity (Steward, 2016)

Replication - testing the reliability of a prior finding with different data.

- Replication has long been established as a key practice in scientific research (Fidler & Wilcox, 2018; Schmidt, 2009).
- It plays a critical role in controlling the impact of sampling error, questionable research practices, publication bias and fraud

Robustness - testing the reliability of a prior finding using the same data and a different analysis strategy

Reproducibility - testing the reliability of a prior finding using the same data and the same analysis strategy


Rigor - following good scientific practices and avoiding bad

State of affairs in different fields

Findings are limited to those fields that investigated replicability

(e.g., psychology, economics, neurology, cancer research, pharmacology, philosophy, physics, political science).

State of affairs in different fields



Experimental Neurology
Volume 233, Issue 2, February 2012, Pages 597-605



Editorial


Replication and reproducibility in spinal cord injury research

Oswald Steward ^{a b c d}  , Phillip G. Popovich ^{a f}, W. Dalton Dietrich ^{g h}, Naomi Kleitman ⁱ

[Show more](#) 

Initiative to replicate spinal-cord-injury research in independent laboratories

2 successful replications out of 12 targeted studies

 OPEN ACCESS

Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in *The BMJ* and *PLOS Medicine*

Florian Naudet,¹ Charlotte Sakarovich,² Perrine Janiaud,¹ Ioana Cristea,^{1,3} Daniele Fanelli,^{1,4} David Moher,^{1,5} John P A Ioannidis^{1,6}

¹Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, USA

²Quantitative Sciences Unit, Division of Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, CA, USA

³Department of Clinical Psychology and Psychotherapy, Babes-Bolyai University, Romania

⁴Department of Methodology, London School of Economics and Political Science, UK

⁵Centre for Journalism, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada

⁶Departments of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics, Stanford University

ABSTRACT

OBJECTIVES
To explore the effectiveness of data sharing by randomized controlled trials (RCTs) in journals with a full data sharing policy and to describe potential difficulties encountered in the process of performing reanalyses of the primary outcomes.

DESIGN
Survey of published RCTs.

SETTING
PubMed/Medline.

ELIGIBILITY CRITERIA
RCTs that had been submitted and published by *The BMJ* and *PLOS Medicine* subsequent to the adoption of data sharing policies by these journals.

MAIN OUTCOME MEASURE
The primary outcome was data availability, defined as the eventual receipt of complete data with clear labelling. Primary outcomes were reanalyzed to assess

in contacting corresponding authors and lack of resources on their behalf in preparing the datasets. In addition, there was a range of different data sharing practices across study groups.

CONCLUSIONS
Data availability was not optimal in two journals with a strong policy for data sharing. When investigators shared data, most reanalyses largely reproduced the original results. Data sharing practices need to become more widespread and streamlined to allow meaningful reanalyses and reuse of data.

TRIAL REGISTRATION
Open Science Framework osf.io/c4zke.

Introduction
Patients, medical practitioners, and health policy analysts are more confident when the results and conclusions of scientific studies can be verified. For

Primary outcomes of 14 out of 17 (82%) randomized control trials (RCTs) published in *The BMJ* or *PLOS Medicine* successfully reproduced.

State of affairs in different fields

Is Economics Research Replicable?
Sixty Published Papers from Thirteen Journals Say
“Usually Not”

Andrew C. Chang* and Phillip Li†

September 4, 2015

Abstract

We attempt to replicate 67 papers published in 13 well-regarded economics journals using author-provided replication files that include both data and code. Some journals in our sample require data and code replication files, and other journals do not require

Key findings of 22 of 59 (approx. 37%) published economics papers (when authors had access to study data) were reproduced.

Current Issue First release papers Archive About  Submit manuscript

Science

HOME > SCIENCE > VOL. 351, NO. 6280 > EVALUATING REPLICABILITY OF LABORATORY EXPERIMENTS IN ECONOMICS

 **REPORT**      

Evaluating replicability of laboratory experiments in economics

COLIN F. CAMERER, ANNA DREBER, ESKIL FORSELL, TECK-HUA HO, JÜRGEN HUBER, MAGNUS JOHANNESSEN, MICHAEL KIRCHLER, JOHAN ALMENBERG, ADAM ALTMEJD, [...]

AND HANG WU +8 authors [Authors Info & Affiliations](#)

SCIENCE • 3 Mar 2016 • Vol 351, Issue 6280 • pp. 1433-1436 • DOI:10.1126/science.1230918

Experimental Economics Replication Project: initiative to replicate prominent findings in experimental economics in independent laboratories

61% of findings successfully replicated

State of affairs in different fields

JOURNAL ARTICLE

The Power of Bias in Economics Research

[Get access >](#)

John P. A. Ioannidis, T. D. Stanley, Hristos Doucouliagos

The Economic Journal, Volume 127, Issue 605, October 2017, Pages F236–F265,

<https://doi.org/10.1111/eoj.12461>

Published: 24 October 2017


“ Cite 🔑 Permissions ➦ Share ▼



Abstract

We investigate two critical dimensions of the credibility of empirical economics research: statistical power and bias. We survey 159 empirical economics literatures that draw upon 64,076 estimates of economic parameters reported in more than 6,700 empirical studies. Half of the research areas have nearly 90% of their results under-powered. The median statistical power is 18%, or less. A simple weighted average of those reported results that are adequately powered (power $\geq 80\%$) reveals that nearly 80% of the reported effects in these empirical economics literatures are exaggerated; typically, by a factor of two and with one-third inflated by a factor of four or more.

Majority of its studies have less than 50% probability of detecting the phenomenon under investigation.

State of affairs in different fields


elife

FEATURE ARTICLE



REPRODUCIBILITY IN CANCER BIOLOGY

Making sense of replications

REPRODUCIBILITY PROJECT
CANCER BIOLOGY

Abstract The first results from the Reproducibility Project: Cancer Biology suggest that there is scope for improving reproducibility in pre-clinical cancer research.
DOI: 10.7554/eLife.23383.001

BRIAN A NOSEK AND TIMOTHY M ERRINGTON*

What is replication? In one sense, the answer is easy. Replication is independently repeating the methodology of a previous study and obtaining the same results. In another sense, the answer is

reflects the current beliefs about what is needed to produce a finding. Conducting a direct replication tests those beliefs empirically. In a con-

Reproducibility Project: Cancer Biology: an initiative to replicate prominent findings in cancer biology

Of 12 replications:

- 4 reproduced important parts of the original article,
- 4 replicated some parts of the original article but not others,
- 2 were not interpretable, and
- 2 did not replicate the original findings


nature human behaviour

[Explore content](#)
[About the journal](#)
[Publish with us](#)
[Subscribe](#)

[nature](#) > [nature human behaviour](#) > [letters](#) > article

Letter | Published: 27 August 2018

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

[Colin F. Camerer](#), [Anna Dreber](#), [Felix Holzmeister](#), [Teck-Hua Ho](#), [Jürgen Huber](#), [Magnus Johannesson](#), [Michael Kirchler](#), [Gideon Nave](#), [Brian A. Nosek](#) , [Thomas Pfeiffer](#), [Adam Altmejd](#), [Nick Buttrick](#), [Taizan Chan](#), [Yiling Chen](#), [Eskil Forsell](#), [Anup Gampa](#), [Emma Heikensten](#), [Lily Hummer](#), [Taisuke Imai](#), [Siri Isaksson](#), [Dylan Manfredi](#), [Julia Rose](#), [Eric-Jan Wagenmakers](#) & [Hang Wu](#)

Social Sciences Replication Project: an initiative to replicate 21 social-science findings in *Science* and *Nature*

13 studies (62% of findings) successfully replicated

State of affairs in different fields

Amgen (Thousand Oaks, CA) - In an attempt to develop treatments for different types of tumors, 53 landmark studies published in biomedical journals were replicated in course of 10 years - **Only 6 (11%) were successfully replicated.**

Bayer HealthCare (Germany) in 2011 reported that **only about 25%** of published preclinical studies **could be validated** to the point at which projects could continue.

Source: Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483(7391), 531-533.

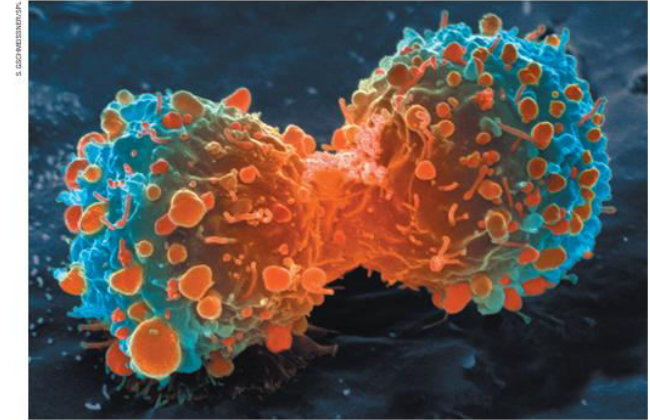
COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge **p.534**

EARTH SYSTEMS Past climates give valuable clues to future warming **p.537**

HISTORY OF SCIENCE Descartes' lost letter tracked using Google **p.540**

OBITUARY Wylie Vale and an elusive stress hormone **p.542**



Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low¹. Sadly, clinical

trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and

investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models² make it difficult for even ▶

State of affairs in different fields

Rev.Phil.Psych.
https://doi.org/10.1007/s13164-018-0400-9

 CrossMark

Estimating the Reproducibility of Experimental Philosophy

Florian Cova^{1,2}  • Brent Strickland^{3,4} • Angela Abatista⁵ • Aurélien Allard⁶ • James Andow⁷ • Mario Attie⁸ • James Beebe⁹ • Renatas Berniūnas¹⁰ • Jordane Boudesseul¹¹ • Matteo Colombo¹² • Fiery Cushman¹³ • Rodrigo Diaz¹⁴ • Noah N'Djaye Nikolai van Dongen¹⁵ • Vilius Dranseika¹⁶ • Brian Earp¹⁷ • Antonio Gaitán Torres¹⁸ • Ivar Hannikainen¹⁹ • José V. Hernández-Conde²⁰ • Wenjia Hu²¹ • François Jaquet¹ • Kareem Khalifa²² • Hanna Kim²³ • Markus Kneer²⁴ • Joshua Knobe²⁵ • Miklos Kurthy²⁶ • Anthony Lantian²⁷ • ...

Initiative to replicate prominent findings in experimental philosophy in independent laboratories

78% of findings successfully replicated

From the Sections

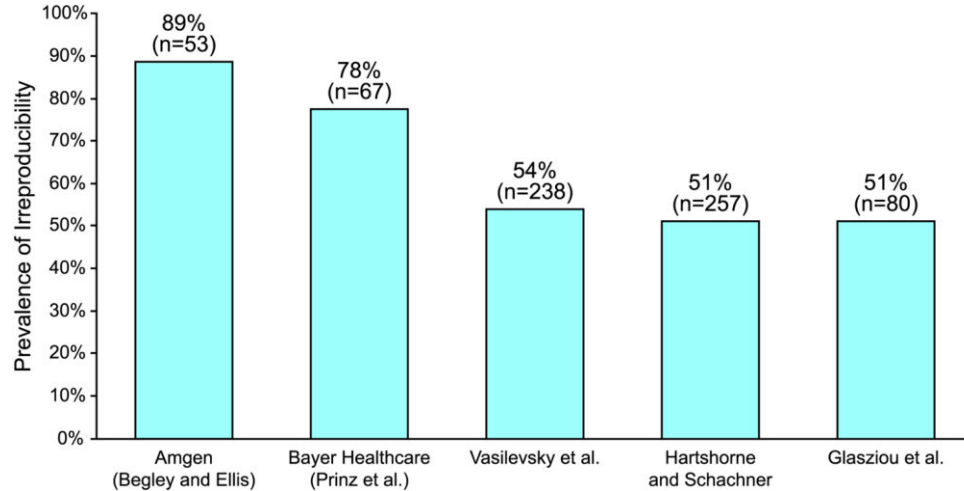
Lessons from a Decade of Replications at the *Quarterly Journal of Political Science*

Nicholas Eubank, *Stanford Graduate School of Business*

ABSTRACT To allow researchers to investigate not only whether a paper's methods are theoretically sound but also whether they have been properly implemented and are robust to alternative specifications, it is necessary that published papers be accompanied by their underlying data and code. This article describes experiences and lessons learned at the *Quarterly Journal of Political Science* since it began requiring authors to provide this type

Only 4 out of 24 articles (17%) were reproducible without author assistance.

Why these results matter?



27.6% of failures are due to flawed study design
 25.5% to data analysis and reporting,
 10.8% to poor laboratory protocols
 36.1% poor tools - subpar biological reagents and reference materials are used

Published estimates of irreproducibility in preclinical research range from 51% to 89%.

Data from 2012: an estimated US\$114.8B in the US is spent annually on life sciences research (pharmaceutical industry 61.8%)



Of this amount, an estimated US\$56.4B (49%) is spent on preclinical research.



Approximately US\$28B/year is spent on research that cannot be replicated (conservative cumulative irreproducibility rate).

Why these results matter?

Estimates are that the cost of irreproducibility has probably gone up since 2015.

The American pharmaceutical industry spent an estimated \$83 billion on research and development in 2019 (Congressional Budget Office, 2021).

If 50% is irreproducible \longrightarrow \$40 billion in US only and about \$90 billion globally each year.

Can we improve state of affairs and how?

Evidence says:
YES

Collaborative research

Collaboration Type	Relevance	Benefits	Limitation and Barriers to Entry	Resources and Examples
Participating teams run different studies	Which existing results are replicable and generalizable?	Assess credibility of previous studies	Efforts focused on previously established findings	Open Science Collaboration: https://osf.io/vmrgu/
Multiteam collaborations	What about when a specific question or intervention needs a definitive answer?	Informs about generalizability, heterogeneity of effect	Massive resources to create Recognition for effort may not match existing incentive structure	www.manyclassess.org https://osf.io/wx7ck/ Overview of Many Labs 2: https://cos.io/our-services/research/many-labs-2-project-overview/
Collaborative analysis	What about when there are many plausible and reasonable analytic choices to answer a research question?	Makes analytic flexibility transparent and develops consensus assessment	Devotes large amount of analyst time to a single question	Many analysts, one data set: https://psyarxiv.com/gkwst/
Preregistered adversarial collaboration	What if there is disagreement within the field on how to interpret existing results?	Reduces post hoc "Whataboutism" and provides clarity regarding a program's strengths and weakness	Requires buy-in and participation of particular researchers who disagree. Researchers must be willing/able to change their view based on results	Example of Adversarial Collaboration Agreement: https://osf.io/deany Matzke, D., van Rijn, H., Wagenmakers, E. J., Slagter, H., van der Molen, M., & Nieuwenhuis, S. (2014, June 3). <i>The effect of horizontal eye movements on free recall performance. A purely confirmatory replication study</i> . Retrieved from osf.io/pxt3m
Persistent collaboration	How could education researchers organize to conduct large-scale collaborations?	Facilitate large scale data accumulation	Massive resources to create Recognition for effort may not match current incentive structure	PsyAccelerator: https://psysciacc.org StudySwap: https://osf.io/view/StudySwap/

Source: Makel, M. C., Smith, K. N., McBee, M. T., Peters, S. J., & Miller, E. M. (2019). A Path to Greater Credibility: Large-Scale Collaborative Education Research. *AERA Open*, 5(4).
<https://doi.org/10.1177/2332858419891963>

Examples of good collaborative research

Selected for a Viewpoint in Physics
PRL 119, 161101 (2017) PHYSICAL REVIEW LETTERS week ending 20 OCTOBER 2017



GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral

B. P. Abbott *et al.**

(LIGO Scientific Collaboration and Virgo Collaboration)

(Received 26 September 2017; revised manuscript received 2 October 2017; published 16 October 2017)

On August 17, 2017 at 12:41:04 UTC the Advanced LIGO and Advanced Virgo gravitational-wave detectors made their first observation of a binary neutron star inspiral. The signal, GW170817, was detected with a combined signal-to-noise ratio of 32.4 and a false-alarm-rate estimate of less than one per 8.0×10^4 years. We infer the component masses of the binary to be between 0.86 and $2.26 M_{\odot}$, in agreement with masses of known neutron stars. Restricting the component spins to the range inferred in binary neutron stars, we find the component masses to be in the range 1.17 – $1.60 M_{\odot}$, with the total mass of the system $2.74^{+0.04}_{-0.04} M_{\odot}$. The source was localized within a sky region of 28 deg^2 (90% probability) and

Around 10,000 visiting scientists from over 113 countries, which represent half of the world's particle physicists, come to CERN for their research. They represent 580 universities and over 85 nationalities. The construction and operation budget contributions are proportional to the GDP of each of the member states.

Laser Interferometer Gravitational-Wave Observatory (LIGO) is a facility for gravitational-wave research. It consists of researchers from California Institute of Technology (Caltech) and the Massachusetts Institute of Technology (MIT), and collaborators from the over 80 scientific institutions world-wide that are members of the **LIGO Scientific Collaboration**

Physics Letters B 716 (2012) 1–29



Contents lists available at SciVerse ScienceDirect

Physics Letters B

www.elsevier.com/locate/physletb



Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC [☆]

ATLAS Collaboration*

This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.

ARTICLE INFO

ABSTRACT

Article history:
Received 31 July 2012

A search for the Standard Model Higgs boson in proton–proton collisions with the ATLAS detector at the LHC is presented. The dataset used corresponds to an integrated luminosity of approximately 36.1 fb^{-1}










Examples of good collaborative research

Science Current Issue First release papers Archive About

HOME > SCIENCE > VOL. 376, NO. 6588 > THE COMPLETE SEQUENCE OF A HUMAN GENOME

SPECIAL ISSUE RESEARCH ARTICLE HUMAN GENOMICS

The complete sequence of a human genome

SERGEY NURK  SERGEY KOREN  ARANG RHIE  MIKKO RAUTIAINEN  ANDREY V. BIKADZE  ALLA MIKHENKO MITCHELL R. VOLLGER 
 NICOLAS ALTIERO  LEV URALSKY  L.-J. AND ADAM M. PHILLIPPY  **+90 authors** [Authors Info & Affiliations](#)

SCIENCE • 31 Mar 2022 • Vol 376, Issue 6588 • pp. 44-59 • DOI:10.1126/science.abc6987

562,383 392


Abstract

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromo-

The efforts of several laboratories located in several countries to complete an initial sequencing of the human genome. Its goal is to determine the sequence of nucleotide base pairs that make up human DNA and to map all the genes of the human genome.

It remains the world's largest collaborative biological project. A large number of discoveries and publications have emerged from this project, due in part to the public availability of the data.

This group is a cooperation between the world's ice research centers on all matters concerning sea ice and icebergs. Presently, IICWG has member organizations from 11 countries and provides a forum for coordination of research activities on ice matters, including icebergs, and acts as an advisory body for the relevant international sea organizations and programs.

 **National Snow and Ice Data Center**
 a part of CIRES at the University of Colorado Boulder

[NEWS & ANALYSES](#) [DATA](#) [OUR RESEARCH](#) [LEARN](#)

Home > International Ice Charting Working Group

International Ice Charting Working Group

[OVERVIEW](#) [MEETINGS](#) [TASK TEAMS](#) [IICWG BUSINESS](#)

Psychology offered some new solutions for
improvement of science

2010's is a decade of crisis (Giner-Sorolla, 2019)

NEUROSCIENCE — DECEMBER 6, 2009

Are psi phenomena real? A study on precognition once exploded science

How a controversial study on psychic powers caused a revolution in psychology research.



OPEN ACCESS Freely available online

PLoS one

Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect

Stuart J. Ritchie^{1*}, Richard Wiseman², Christopher C. French³

¹ Psychology Department, The University of Edinburgh, Edinburgh, United Kingdom, ² School of Psychology, University of Hertfordshire, Hatfield, United Kingdom, ³ Anomalous Psychology Research Unit, Goldsmiths, University of London, London, United Kingdom

Abstract

Nine recently reported parapsychological experiments appear to support the existence of precognition. We describe three

Diederik Stapel now has 58 retractions

Social psychologist [Diederik Stapel](#) has notched his 58th retraction, after admitting he fabricated data in yet another article.

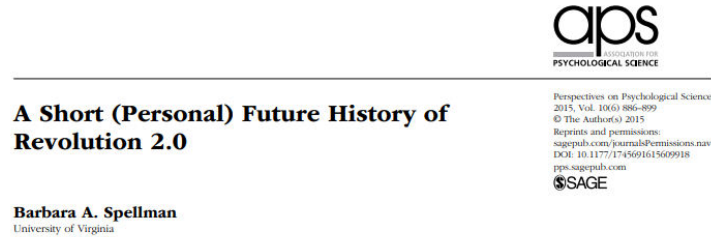
He's holding onto his 4th place spot on [our leaderboard](#).



Subsequent analyses showed that:

- selective reporting of analyses and studies was almost certainly going on (Schimmack, 2012),
- Even multi-study papers have a problem for inference (Simmons, Nelson & Simonsohn, 2011)
- Research are almost exclusively evaluated by p-value thresholds, ignoring statistical power and effect sizes (Cumming, 2014).

2010's is a decade of revolution (Spellman, 2015)



Abstract
Crisis of replicability is one term that psychological scientists use for the current introspective phase we are in—I argue instead that we are going through a revolution analogous to a political revolution. Revolution 2.0 is an uprising focused on how we should be doing science now (i.e., in a 2.0 world). The precipitating events of the revolution have already been well-documented: failures to replicate, questionable research practices, fraud, etc. And the fact that none of these events is new to our field has also been well-documented. I suggest four interconnected reasons as to why this time is different: changing technology, changing demographics of researchers, limited resources, and misaligned incentives. I



Abstract
 The credibility revolution (sometimes referred to as the “replicability crisis”) in psychology has brought about many changes in the standards by which psychological science is evaluated. These changes include (a) greater emphasis on transparency and openness, (b) a move toward preregistration of research, (c) more direct-replication studies, and (d) higher standards for the quality and quantity of evidence needed to make

Changes in the standards by which psychological science is evaluated:

- greater emphasis on transparency and openness,
- a move toward preregistration of research,
- more direct-replication studies, and
- higher standards for the quality and quantity of evidence needed to make strong scientific claims.

2010's is a decade of renaissance (Nelson et al., 2015)



Annual Review of Psychology

Psychology's Renaissance

Leif D. Nelson,¹ Joseph Simmons,²
and Uri Simonsohn²

¹Haas School of Business, University of California, Berkeley, California 94720;
email: Leif_Nelson@haas.berkeley.edu

²The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104;
email: jimmio@upenn.edu, urisoehn@gmail.com

Milestones:

Simmons et al (2011) published a paper “False-Positive Psychology” reporting the surprisingly severe consequences of selectively reporting data and analyses, i.e., *p-hacking*.

Daniel Kahneman circulated email calling for researchers to resolve the debate by conducting systematic replications - after Doyen et al., (2012) failed to replicate one of the most famous findings in social psychology, that priming people with elderly stereotypes made them walk more slowly (Bargh et al. 1996).

Brian Nosek organized several replication efforts in 2011 and together with Jeffrey Spies developed Open Science Framework platform in 2012, and Center for Open Science in 2013.

Reproducibility Project: Psychology (RP:P)

A large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

We replicated 100 experimental and correlational studies published in three flagship journals using high-powered designs and original materials when available.

- 97% of original studies had significant results ($P < .05$).

Criteria for evaluating reproducibility: p-value, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes.

RESEARCH

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration^{*,†}

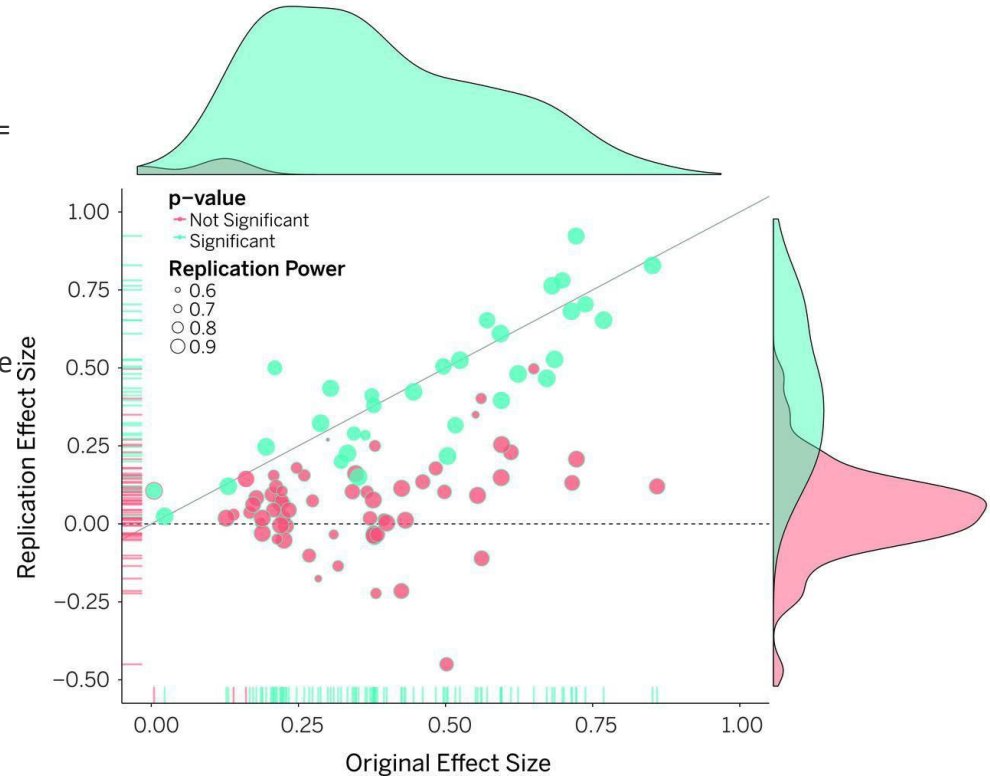
Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication 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.

RP:P

Mean effect size (r) of the replication effects ($M_r = 0.197$, $SD = 0.257$) - half the magnitude of the mean effect size of the original effects ($M_r = 0.403$, $SD = 0.188$), representing a substantial decline.

- 36% 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;
- if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects.

Replication success **was better predicted by the strength of original evidence than by characteristics of the original and replication teams.**



Many Labs 2 - Investigating Variation in Replicability Across Samples and Settings

We conducted pre-registered replications of 28 classic and contemporary published findings.

Protocols were peer reviewed in advance, to examine variation in effect magnitudes across samples and settings.

Each protocol was administered to approximately half of 125 samples that comprised $N = 15,305$ participants from 36 countries and territories.

Advances in Methods and Practices in Psychological Science
Volume 1, Issue 4, December 2018, Pages 443-490
© The Author(s) 2018, Article Reuse Guidelines
<https://doi.org/10.1177/2515245918810225>



Registered Replication Report and Commentaries

Many Labs 2: Investigating Variation in Replicability Across Samples and Settings

Richard A. Klein¹, Michelangelo Vianello², Fred Hasselman^{3,4}, Byron G. Adams^{5,6}, Reginald B. Adams, Jr.⁷, Sinan Alper⁸, Mark Aveyard⁹, Jordan R. Axt¹⁰, Mayowa T. Babalola¹¹, Štěpán Bahník¹², Rishree Batra¹³, Mihály Berkics¹⁴, Michael J. Bernstein¹⁵, Daniel R. Berry¹⁶, Olga Bialobrzeska¹⁷, Evans Dami Binan¹⁸, Konrad Bocian¹⁹, Mark J. Brandt⁵, Robert Busching²⁰, Anna Cabak Rédei²¹, Huajian Cai²², Fanny Cambier^{23,24}, Katarzyna Cantarero²⁵, Cheryl L. Carmichael²⁶, Francisco Ceric^{27,28}, Jesse Chandler^{29,30}, Jen-Ho Chang^{31,32}, Armand

Many Labs 2

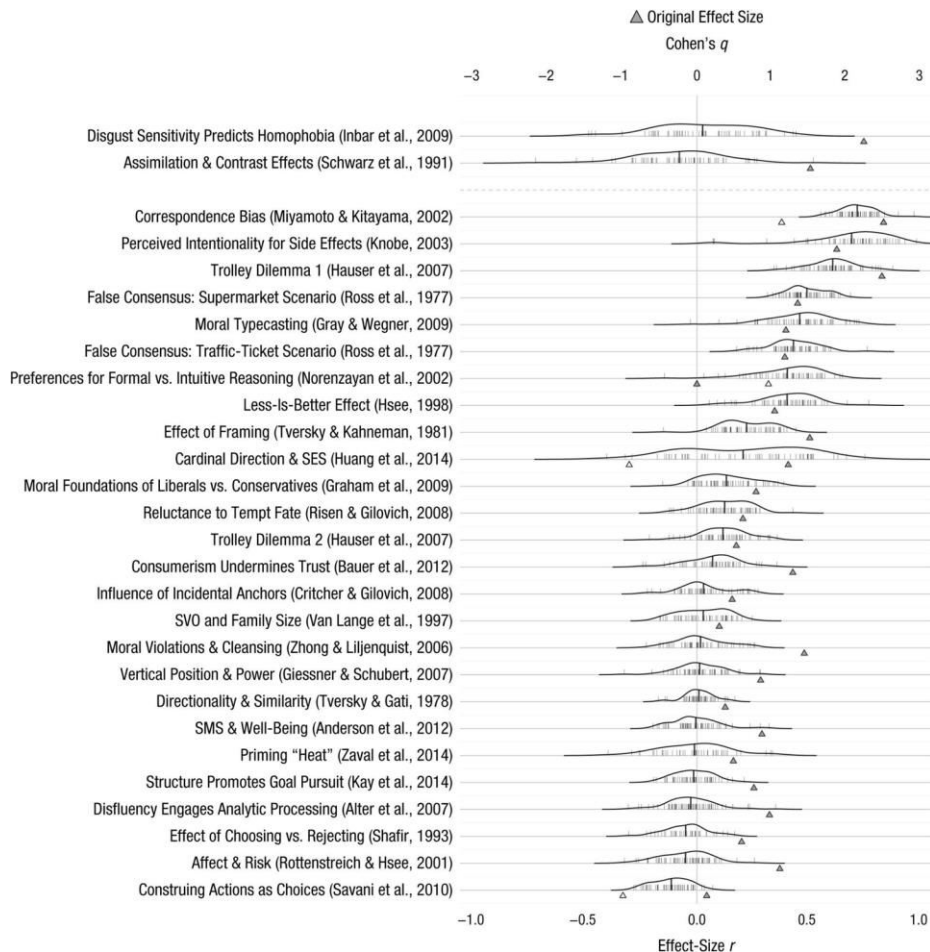
15 (54%) of the replications provided evidence of a statistically significant effect in the same direction as the original finding - using the conventional criterion of statistical significance ($p < .05$)

14 (50%) of the replications still provided such evidence, a reflection of the extremely high-powered design - with a strict significance criterion ($p < .0001$)

7 (25%) of the replications yielded effect sizes larger than the original ones, and 21 (75%) yielded effect sizes smaller than the original ones

The median comparable Cohen's d s were 0.60 for the original findings and 0.15 for the replications.

The effect sizes were small (< 0.20) in 16 of the replications (57%), and 9 effects (32%) were in the direction opposite the direction of the original effect.



Many Labs 3: Evaluating participant pool quality across the academic semester via replication

20 laboratories attempted to replicate 10 psychology findings at different times of the semester

This crowdsourced project examined time of semester variation in 10 known effects, 10 individual differences, and 3 data quality indicators over the course of the academic semester in 20 participant pools ($N = 2696$) and with an online sample ($N = 737$).

3 of 10 findings replicated; most unaffected by time of semester



Journal of Experimental Social Psychology
Volume 67, November 2016, Pages 68–82



Many Labs 3: Evaluating participant pool quality across the academic semester via replication ☆

Charles R. Ebersole ^a, Olivia E. Atherton ^b, Aimee L. Belanger ^c,
Hayley M. Skulborstad ^c, Jill M. Allen ^d, Jonathan B. Banks ^e, Erica Baranski ^f,
Michael J. Bernstein ^g, Diane B.V. Bonfiglio ^h, Leanne Boucher ^e, Elizabeth R. Brown ⁱ,
Nancy I. Budiman ^b, Athena H. Cairo ^j, Colin A. Capaldi ^k, Christopher R. Chartier ^h,
Joanne M. Chung ^b, David C. Cicero ^l, Jennifer A. Coleman ^j, John G. Conway ^m,
William E. Davis ⁿ, Thierry Devos ^o, Melody M. Fletcher ^p, Komi German ^b,
Jon E. Grahe ^q, Anthony D. Hermann ^r, Joshua A. Hicks ⁿ, Nathan Honeycutt ^o

Source: Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... & Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68–82. <https://doi.org/10.1016/j.jesp.2015.10.012>

Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability

We selected 10 replication studies from the RP:P (OSC, 2015) for which the original authors had expressed concerns about the replication designs before data collection.

Commenters suggested that lack of adherence to expert review and low-powered tests were the reasons that most of these RP:P studies failed to replicate the original effects.

We revised the replication protocols and received formal peer review prior to conducting new replication studies.

We administered the RP:P and revised protocols in multiple laboratories.

Special Section: Many Labs 5 Project
Empirical Article: Registered Report

aps
ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability



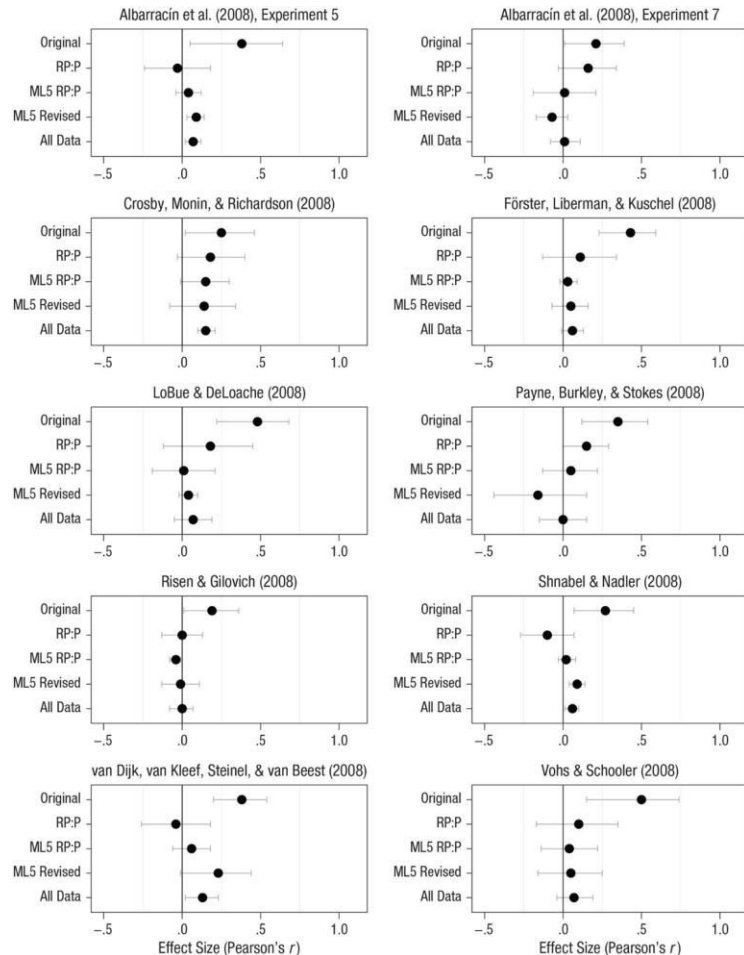
Charles R. Ebersole¹, Maya B. Mathur², Erica Baranski³,
Diane-Jo Bart-Plange⁴, Nicholas R. Buttrick⁵,
Christopher R. Chartier⁶, Katherine S. Corker⁵, Martin Corley⁶,
Joshua K. Hartshorne⁷, Hans IJzerman^{8,9}, Ljiljana B. Lazarevic^{10,11},
Hugh Rabagliati⁶, Ivan Ropovik^{12,13}, Balazs Aczel¹⁴, Lena F. Aeschbach¹⁵,
Luca Andrichetto¹⁶, Jack D. Arnal¹⁷, Holly Arrow¹⁸, Peter Babincak¹⁹,
Bence E. Bakos¹⁴, Gabriel Banik¹⁹, Ernest Baskin²⁰,
Dedovic Balenovic²¹, Michael H. Bernstein^{22,23}, Michael Biala²⁴

Advances in Methods and
Practices in Psychological Science
2020, Vol. 3(3), 309–331
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245920958687
www.psychologicalscience.org/AMPPS
SAGE

Many Labs 5

Revised protocols produced effect sizes similar to those of the RP:P protocols ($\Delta r = .002$ or $.014$, depending on analytic approach).

The median effect size for the revised protocols ($r = .05$) was similar to that of the RP:P protocols ($r = .04$) and the original RP:P replications ($r = .11$), and smaller than that of the original studies ($r = .37$).



Using prediction markets to improve science

Prediction markets

Idea proposed by Hanson (1995) to overcome major problems in academia (*publish or perish*; mainstream theories and ideas favorized; strong hierarchical organization where those on the top get all the credit; publications, grants and tenures are evaluated mostly for investigating what is accepted as mainstream, etc.)

Hanson proposes market-based alternative (“idea-futures”) where scientists more formally can stake their reputation if not doing good science, and where incentives are not offered for academics *mainly for telling a good story, rather than for being right*.

DARPA SCORE

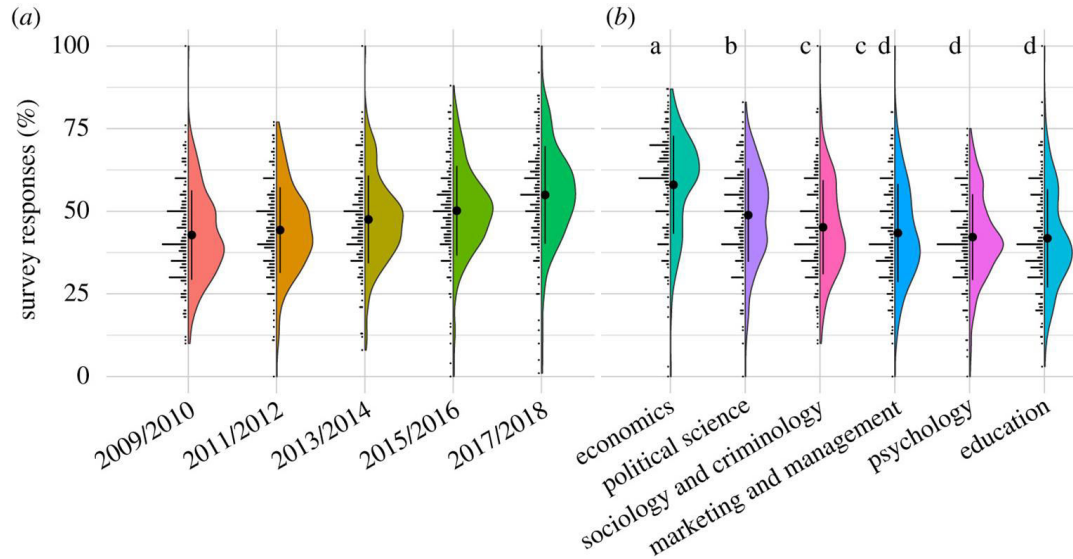
Information about replication outcomes can be elicited from the research community (Camerer et al., 2016, 2018; Dreber et al., 2015; Forsell et al., 2018) - forecasting the outcomes of hypothetical replications can help assessing replication probabilities without requiring the resources for actually conducting replications.

Dreber et al. (2015) - Prediction markets are able to quickly identify findings that are unlikely to replicate, they predict the outcomes of replications well and outperform a survey of individual forecasts.

The Defense Advanced Research Projects Agency (DARPA) programme 'Systematizing Confidence in Open Research and Evidence' (SCORE) aims to generate confidence scores for a large number of research claims from empirical studies in the social and behavioural sciences (Gordon et al., 2020).

- Quantitative assessment of how likely a claim will hold up in an independent replication
- A small subset of the claims (about 5%) is assessed through replication, and the replication outcomes are used to evaluate the accuracy of the confidence scores.
- The research claims are sampled from studies published during a 10 year period (2009–2018) across 60 journals from a number of academic disciplines.

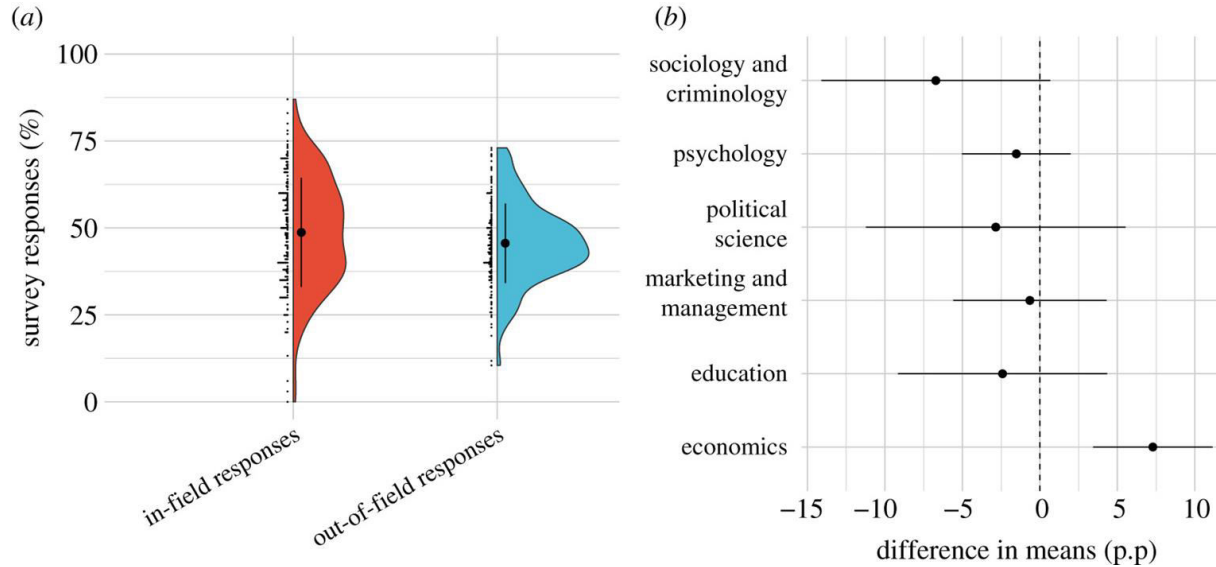
DARPA-SCORE



(a) Results show that participants expect replication rates to increase over time, from 43% in 2009/2010 to 55% in 2017/2018 due to recent methodological changes in the social and behavioural sciences having positive impact on replication rates.

(b) Expected replication rates differ between fields, with the highest replication rate in economics (average survey response 58%), and the lowest in psychology and in education (average survey response of 42% for both fields).

DARPA-SCORE



(a) In-field versus out-of-field responses. Participants predict a higher replication rate for their fields of interest, as compared to other fields.

(b) Difference of evaluation of a field by in-field and out-field participants (in per cent points). Participants with interest in economics predict a higher replication rate for this field compared to participants with no interest in economics. For other fields, such an effect is not observed. Points and error bars indicate the mean ± 1 s.d.

How can we motivate researchers to practice
open and credible science?

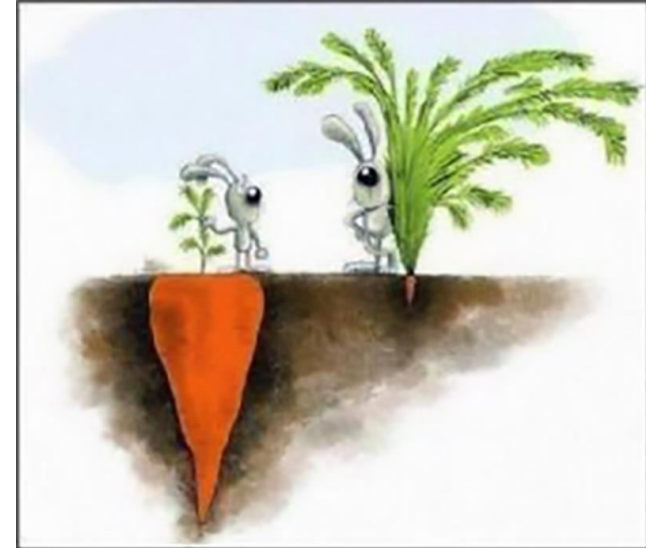
Open science and Incentives for researchers

Incentives for scientists – why should I spend my **valuable** time to share the data, code, materials, etc.?

Making data publicly available is time-consuming!

Open Practice Badges are incentives for researchers to share data, materials, or to pre-register their work.

They are designed to be displayed on published articles to show that authors have engaged in these open practices.





Received: 23 March 2023 | Accepted: 29 May 2023

DOI: 10.1111/bjc.12431

RESEARCH ARTICLE

The anatomy of COVID-19-related conspiracy beliefs: Exploring their nomological network on a nationally representative sample

Goran Knežević¹ | Ljiljana B. Lazarević¹ | Ljiljana Mihić² |
Milica Pejović Milovančević^{3,4} | Zorica Terzić^{3,5} | Oliver Tošković¹ |
Olivera Vuković^{3,4} | Jovana Todorović^{3,5} | Nada P. Marić^{3,4}



META-RESEARCH ARTICLE

Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency

Mallory C. Kidwell^{1*}, Ljiljana B. Lazarević², Erica Baranski³, Tom E. Hardwicke⁴, Sarah Piechowski⁵, Lina-Sophia Falkenberg⁶, Curtis Kennett⁶, Agnieszka Slowik⁷, Carina Sonleitner⁷, Chelsey Hess-Holden⁸, Timothy M. Errington¹, Susann Fiedler⁸, Brian A. Nosek^{1,8}

¹ Center for Open Science, Charlottesville, Virginia, United States of America, ² University of Belgrade, Belgrade, Serbia, ³ University of California, Riverside, Riverside, California, United States of America, ⁴ University College London, London, United Kingdom, ⁵ Max Planck Institute for Research on Collective Goods, Bonn, Germany, ⁶ Mississippi State University, Starkville, Mississippi, United States of America, ⁷ University of Vienna, Vienna, Austria, ⁸ University of Virginia, Charlottesville, Virginia, United States of America

* Mallory@cos.io; Mallory.Kidwell@utah.edu



OPEN ACCESS

Citation: Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, Falkenberg L-S, et al. (2016) Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLoS Biol* 14(5): e1002456. doi:10.1371/journal.pbio.1002456

Academic Editor: Malcolm R. Macleod, University of Edinburgh, UNITED KINGDOM

Published: May 12, 2016

Copyright: © 2016 Kidwell et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data and materials

Abstract

Beginning January 2014, *Psychological Science* gave authors the opportunity to signal open data and materials if they qualified for badges that accompanied published articles. Before badges, less than 3% of *Psychological Science* articles reported open data. After badges, 23% reported open data, with an accelerating trend; 39% reported open data in the first half of 2015, an increase of more than an order of magnitude from baseline. There was no change over time in the low rates of data sharing among comparison journals. Moreover, reporting openness does not guarantee openness. When badges were earned, reportedly available data were more likely to be actually available, correct, usable, and complete than when badges were not earned. Open materials also increased to a weaker degree, and there was more variability among comparison journals. Badges are simple, effective signals to promote open practices and improve preservation of data and materials by using independent repositories.

Badges to acknowledge open science practices

Beginning January 2014, *Psychological Science* gave authors the opportunity to signal open data and materials if they qualified for badges that accompanied published articles.

Analyzed journals:

- *Psychological Science* (PSCI; $N = 838$) - respected journal that publishes empirical research from any area of psychology
- *Journal of Personality and Social Psychology* (JPSP; $N = 419$),
- *Journal of Experimental Psychology: Learning, Memory, and Cognition* (JEPLMC; $N = 483$),
- *Developmental Psychology* (DP; $N = 634$), and
- *Clinical Psychological Science* (CPS; $N = 104$).

We examined the entire population of empirical articles from the target and comparison journals ($N = 2,478$).

A total of 220 additional articles published in these journals between 2012 and May 2015 are not part of this corpus because they were not reports of empirical research (i.e., editorials, theoretical reviews, commentaries).



Do open science badges really work?

“Badges are simple, effective signals to promote open practices and improve preservation of data and materials by using independent repositories.” (Kidwell, Lazarevic et al., 2016).

- However, badges are not panacea - although sharing rates increase dramatically, researchers still do not use this as much as they could be used and do not share materials and (especially) data.
- Hardwicke et al., (2021) - evaluated analytic reproducibility in 25 PSCI papers after badges were introduced
 - 36% reproducible without requesting input from original authors
 - 24% reproducible with original authors involvement
 - 12% not fully reproducible with no substantive author response
 - 12% not fully reproducible despite original authors involvement
 - Non-reproducibility was primarily caused by unclear reporting of analytic procedures.
 - Open data alone is not sufficient to ensure analytic reproducibility.



Do mandatory open-data policies work?

Cognition journal - mandatory open data policies introduced in March 1, 2015.

- Data availability was almost 100% after the evaluation period
- Reusability reached 75%
- Analytic reproducibility -Prior to requesting assistance from original authors, reproducibility success rate of just 31%, but with original author assistance it increases to 63% (37% could not be reproduced even with author assistance).

Mandatory open data policies can increase the frequency and quality of data sharing.

However, suboptimal data curation, unclear analysis specification and reporting errors can impede analytic reproducibility, undermining the utility of data sharing and the credibility of scientific findings.

Pre-registration

Pre-registration - what it is and why we should do it?

Pre-registration of an analysis plan - committing to analytic steps without advance knowledge of the research outcomes (Nosek et al., 2018).

The analysis plan is posted on an independent registry such as <https://clinicaltrials.gov/> or <https://osf.io/>.

The registry preserves the preregistration and makes it discoverable

Observed (collected) data do not influence selection of the analytical tests - analytical strategy is known in advance.

Pre-registration distinguishes analyses and outcomes that result from predictions from those that result from postdictions.

[Example 1](#)

[Example 2](#)

Pre-registration

Researchers claimed that pre-registering studies is a major step toward greater research credibility ([Munafò et al., 2017](#), [Nosek et al., 2018](#)).

Strømmland (2019) - Experimental economy

- In absence of preregistration effect sizes are inflated and replications overestimate statistical power.
- pre-registration improves estimation of effect sizes and therefore reproducibility

Van der Akker et al. (2023) - Psychology - Comparison of 193 studies that earned a Preregistration Challenge prize or pre-registration badge to 193 related studies that were not pre-registered.

- preregistration increases statistical power and impact, but robust evidence that preregistration prevents *p*-hacking and HARKing were not found

Pre-registration increases confidence that the reported confirmatory analyses were not cherry-picked from a broader set - decreases *p*-hacking

Pre-registration decreases misinterpreting exploratory results as confirmatory ones – i.e., Hypothesizing After Results are Known, or “HARKing” (Kerr, 1998).

Pre-registration does not guarantee that every published finding will be true, but without it you can safely bet that many more will be false (Simmons et al., 2021)

Open science and teaching



Open science and teaching

Learning students about good scientific practices has impact on the scientific field!

The **Collaborative Replications and Education Project** (CREP; <http://osf.io/wfc6u>) is a framework for undergraduate students to participate in the production of high-quality direct replications

CREP's primary purpose is educational: to teach students good scientific practices by performing direct replications of highly cited works in the field using open science methods.

Publishing Research With Undergraduate Students via Replication Work: The Collaborative Replications and Education Project

Jordan R. Wagge^{1*}, Mark J. Brandt², Ljiljana B. Lazarevic³, Nicole Legate⁴, Cody Christopherson⁵, Brady Wiggins⁶ and Jon E. Grahe⁷

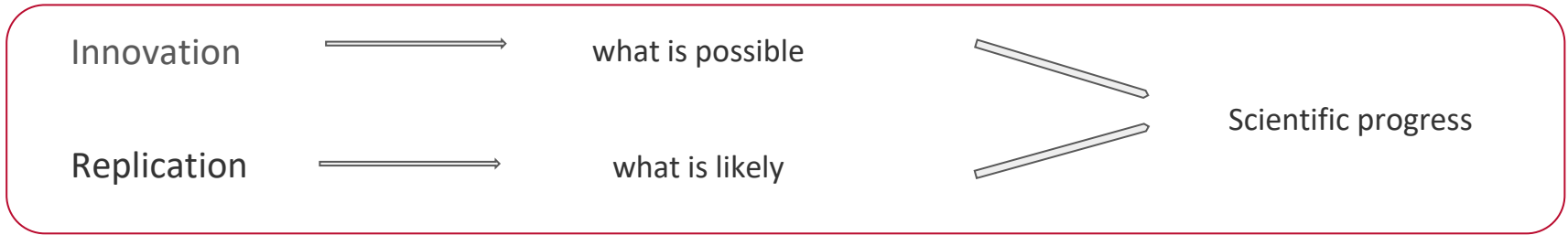
¹ School of Psychology, Avila University, Kansas City, MO, United States; ² Department of Social Psychology, Tilburg University, Tilburg, Netherlands; ³ Institute of Psychology, University of Belgrade, Belgrade, Serbia; ⁴ Department of Psychology, Winstate Institute of Technology, Chicago, IL, United States; ⁵ Department of Psychology, Southern Oregon University, Ashland, OR, United States; ⁶ Department of Psychology, Brigham Young University-Idaho, Rexburg, ID, United States; ⁷ Department of Psychology, Pacific Lutheran University, Tacoma, WA, United States

Keywords: replication, pedagogy, psychology, publishing, undergraduates, teaching, projects, open science



Take home message

Innovative ideas are pivotal for scientific progress, but they can become old news fast



Refusal to publish new tests of published ideas as unoriginal influences scientific progress

Incentives for individual scientists prioritize novelty over replication - thus, influencing the credibility of scientific findings.

Credible science is possible only when we have evidence-based assessment of the quality of research and critically evaluate them.



Thank you

ljiljana.lazarevic@f.bg.ac.rs

